# Fairness in AI enabled clinical decision making: why isn't it being evaluated?

Honghan Wu
Associate Professor | Turing Fellow
Institute of Health informatics, UCL
April 2023

# Inequity in health is long-standing & v damaging for AI

BMJ Journals

Log In Basket search

**BMJ Health & Care Informatics**

Latest content Archive Authors About

Home / Archive / Volume 29, Issue 1

Article Text

Article info

Original research

Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction

PDF

PDF + Supplementary Material

Isabel Straw and Honghan Wu

Correspondence to Dr Isabel Straw: isabelstraw@doctors.org.uk

false negative predictions for liver diseases are twice in women as in men

## AI is the future of the NHS. It's also disadvantaging women and ethnic minorities

EXCLUSIVE

Experts warn that new research into AI in healthcare shows a failure to consider the full range of potential bias against particular groups of people will have life or death consequences

Decisions made by AI models to determine who should have an operation can be biased against women and non-white people, studies have shown (Peter Byrne/PA)

By Tom Bawden

Science & Environment Correspondent

https://inews.co.uk/news/science/why-ai-could-lead-to-a-poorer-performing-nhs-for-women-and-ethnic-minorities-1715312

# Fairness notions and metrics

arXiv > cs > arXiv:1909.11869

Search...

Help | Advanc

Computer Science > Computers and Society

[Submitted on 26 Sep 2019]

## This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology

Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, Richmond Y. Wong

The explosion in the use of software in important sociotechnical systems has renewed focus on the study of the way technical constructs reflect policies, norms, and human values. This effort requires the engagement of scholars and practitioners from many disciplines. And yet, these disciplines often conceptualize the operative values very differently while referring to them using the same vocabulary. The resulting conflation of ideas confuses discussions about values in technology at disciplinary boundaries. In the service of improving this situation, this paper examines the value of shared vocabularies, analytics, and other tools that facilitate conversations about values in light of these disciplinary specific conceptualizations, the role such tools play in furthering research and practice, outlines different conceptions of "fairness" deployed in discussions about computer systems, and provides an analytic tool for interdisciplinary discussions and collaborations around the concept of fairness. We use a case study of risk assessments in criminal justice applications to both motivate our effort--describing how conflation of different concepts under the banner of "fairness" led to unproductive confusion--and illustrate the value of the fairness analytic by demonstrating how the rigorous analysis it enables can assist in identifying key areas of theoretical, political, and practical misunderstanding or disagreement, and where desired support alignment or collaboration in the absence of consensus.

Comments: 36 pages
Subjects: **Computers and Society (cs.CY)**; Human-Computer Interaction (cs.HC)

*“The concept of fairness is **vast and ambiguous**, and differently used across disciplines.”*

# Fairness notions and metrics

© 2019 The Author(s)
JoGH © 2019 ISoGH

VIEWPOINTS

## Artificial intelligence and algorithmic bias: implications for health systems

Trishan Panch[1,2], Heather Mattie[3], Rifat Atun[4]

[1] Department of Health Policy and Management, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA
[2] Wellframe, Boston, Massachusetts, USA
[3] Department of Biostatistics and Executive Director, Health Data Science Masters Program, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA
[4] Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA

*“there is **no broadly recognized quantitative summary metric** for fairness and hence evaluation is ultimately qualitative, and subject to implicit biases of the evaluators.”*

# Fairness notions and metrics

2018 ACM/IEEE International Workshop on Software Fairness

**Fairness Definitions Explained**

Sahil Verma
Indian Institute of Technology Kanpur, India
vsahil@iitk.ac.in

Julia Rubin
University of British Columbia, Canada
mjulia@ece.ubc.ca

*“reviewed publications in major conferences and journals on ML and fairness, such as NIPS, Big Data, AAAI, FATML, ICML, and KDD, in the last six years.”*

| | Definition | Paper | Citation # | Result |
|---|---|---|---|---|
| 3.1.1 | Group fairness or statistical parity | [12] | 208 | × |
| 3.1.2 | Conditional statistical parity | [11] | 29 | ✓ |
| 3.2.1 | Predictive parity | [10] | 57 | ✓ |
| 3.2.2 | False positive error rate balance | [10] | 57 | × |
| 3.2.3 | False negative error rate balance | [10] | 57 | ✓ |
| 3.2.4 | Equalised odds | [14] | 106 | × |
| 3.2.5 | Conditional use accuracy equality | [8] | 18 | × |
| 3.2.6 | Overall accuracy equality | [8] | 18 | ✓ |
| 3.2.7 | Treatment equality | [8] | 18 | × |
| 3.3.1 | Test-fairness or calibration | [10] | 57 | ✓✗ |
| 3.3.2 | Well calibration | [16] | 81 | ✓✗ |
| 3.3.3 | Balance for positive class | [16] | 81 | ✓ |
| 3.3.4 | Balance for negative class | [16] | 81 | × |
| 4.1 | Causal discrimination | [13] | 1 | × |
| 4.2 | Fairness through unawareness | [17] | 14 | ✓ |
| 4.3 | Fairness through awareness | [12] | 208 | × |
| 5.1 | Counterfactual fairness | [17] | 14 | – |
| 5.2 | No unresolved discrimination | [15] | 14 | – |
| 5.3 | No proxy discrimination | [15] | 14 | – |
| 5.4 | Fair inference | [19] | 6 | – |

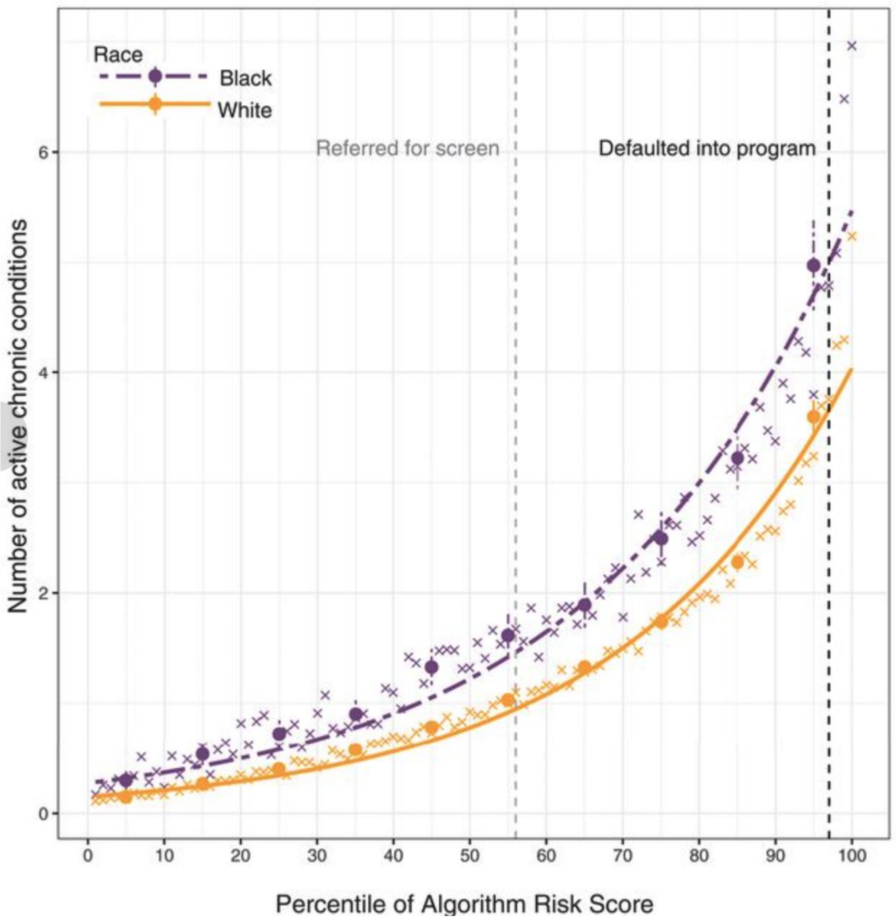
**Table 1: Considered Definitions of Fairness**

The reality, however, is existing metrics have not been well adopted in AI in Medicine. **Why?**

# Our hypothesis on the difficulty is three-fold

1. **Confusion on which one to choose**
   - *which one fits the context the best*
   - *“the same case can be considered fair according to some definitions and unfair according to others.”*
   - *too ‘complex’ / ‘generic’*

# Our hypothesis on the difficulty is four-fold




https://www.science.org/doi/10.1126/science.aax2342

## 2. ‘unfair’ target variable - y

- *due to the data embedded inequity*
- *pervasive in health data*

**Original Investigation**

January 9, 2020

**Association of Insurance Status and Racial Disparities With the Detection of Early-Stage Breast Cancer**

Naomi Y. Ko, MD, MPH, AM[1,2]; Susan Hong, MD, MPH[3]; Robert A. Winn, MD[4]; et al

» Author Affiliations | Article Information

*JAMA Oncol.* 2020;6(3):385-392. doi:10.1001/jamaoncol.2019.5672

*non-Hispanic black (OR, **1.46** [95% CI, 1.40-1.53]), American Indian or Alaskan Native (OR, **1.31** [95% CI, 1.07-1.61]) and Hispanic (OR, **1.35** [95% CI, 1.30-1.42]) women had higher odds of receiving a diagnosis of locally advanced disease (stage III) compared with non-Hispanic white women*

# Our hypothesis on the difficulty is three-fold

3. clinicians’ opinions on “individuals’ actual health needs” not easily integrable in the fairness frameworks

# Deterioration Allocation Framework

## Quantifying Health Inequalities Induced by Data and AI Models

Honghan Wu, Aneeta Sylolypavan, Minhong Wang, Sarah Wild

Watch video

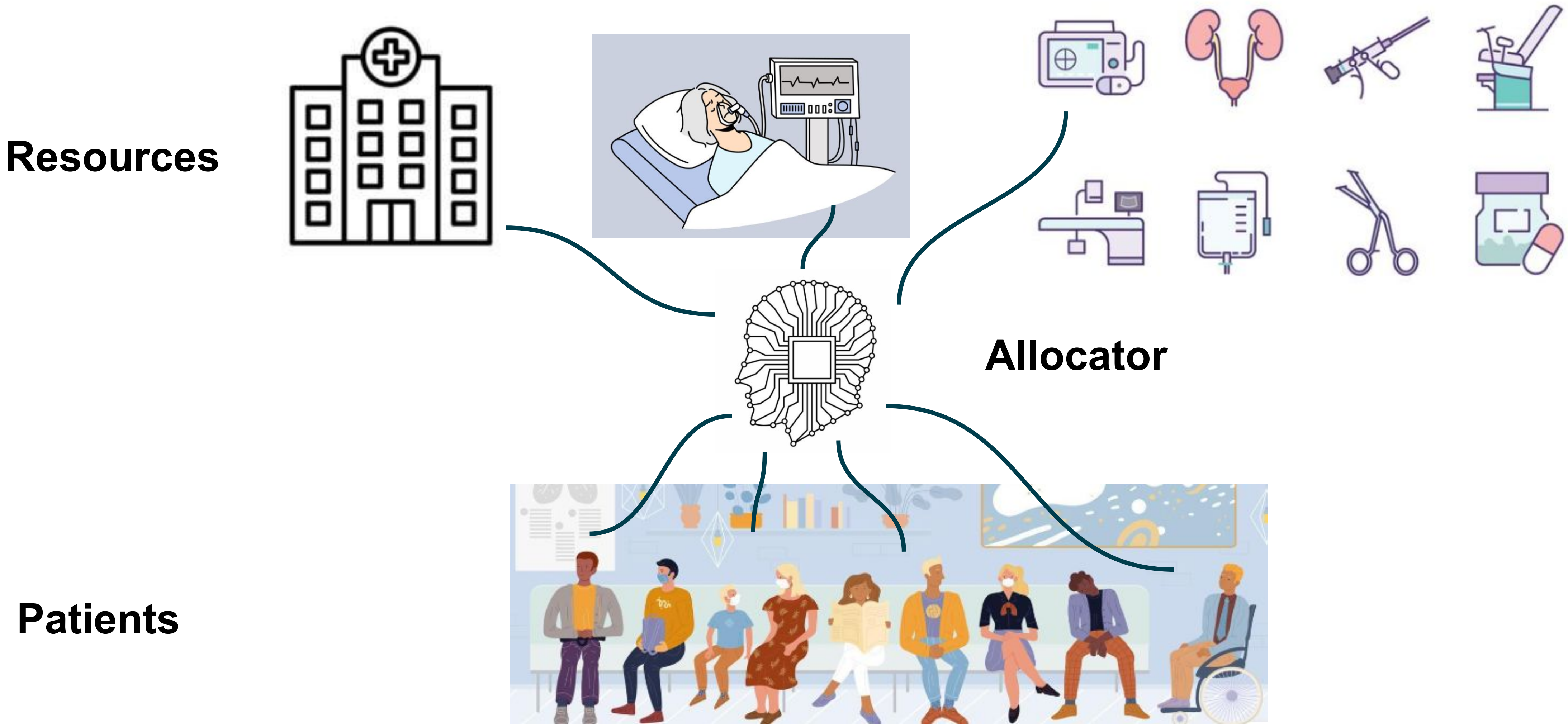
IJCAI ECAI
VIENNA 22

Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence
AI for Good. Pages 5192-5198. https://doi.org/10.24963/ijcai.2022/721

PDF BibTeX

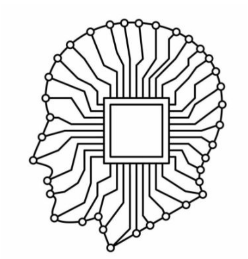
https://www.ijcai.org/proceedings/2022/721

# Abstracting clinical decision making: resource allocation




*Acknowledgement: images from https://www.vecteezy.com/*

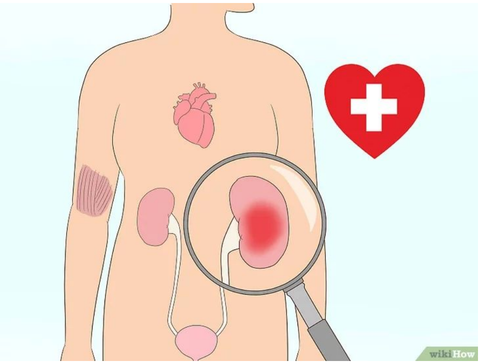
# Fairness in resource allocation scenario

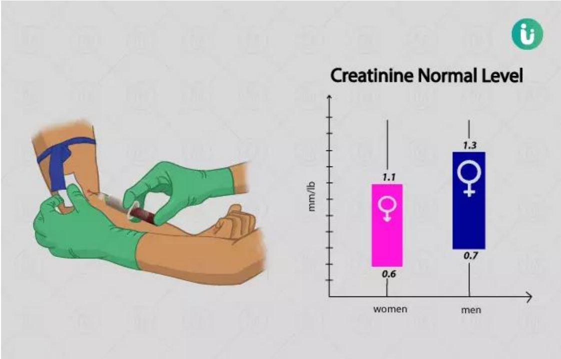


The same level of “**health needs**”

gets

equal access to resources



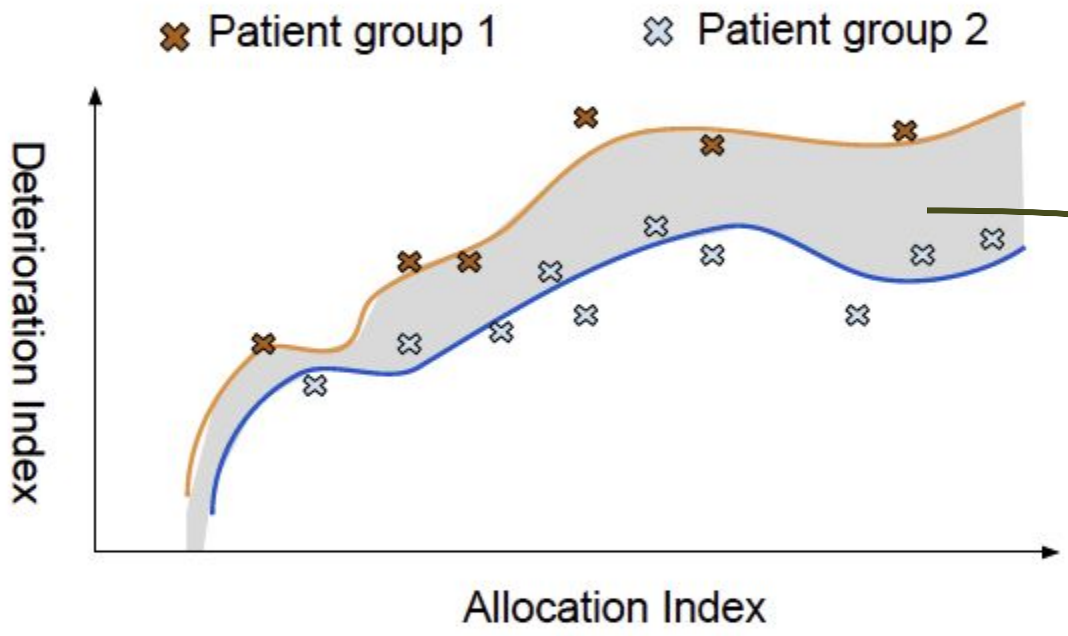

Allow the use of “objective” measurements for assessing health needs

*“25 560 diagnostic biomarkers,* ***102*** *prognostic biomarkers, 265 exposure biomarkers and 6746 predictive biomarkers”*

*Wishart, David S., et al. "MarkerDB: an online database of molecular biomarkers." Nucleic Acids Research 49.D1 (2021): D1259-D1267.*

# Fairness in resource allocation scenario

Deterioration index measures the deterioration status of patients (marker of prognosis) - ***health needs***




Area under allocation-deterioration curves

DA-AUC

Allocation index is the score derived from “a resource allocator”

# Relations of DA-AUC to existing fairness notions

| | Notion | | Use of $Y$ | Applicability in clinical decision making | Relation with DA-AUC |
|---|---|---|---|---|---|
| Group fairness | Demographic Parity | | – | Not directly applicable | - |
| | Conditional Demographic Parity | | – | Not directly applicable | - |
| | Error parity | Equal Accuracy | ✓ | Yes, but not fair when Y is biased | Equivalent to the DA-AUC when defining deterioration index as $DI = \begin{cases} 0, & if\ pred == y \\ 1, & otherwise \end{cases}$ |
| | | Equality of Odds | ✓ | Same as above | Equivalent to the DA-AUC when defining deterioration index as $DI = \begin{cases} 1, & if\ pred = 1\ and\ y = 0 \\ 0, & otherwise \end{cases}$ |
| | | Predictive Parity | ✓ | Same as above | Equivalent to the DA-AUC when defining deterioration index as $DI = \begin{cases} 1, & if\ pred = 0\ and\ y = 1 \\ 0, & otherwise \end{cases}$ |
| Individual fairness | FTU/Blindness | | – | Yes, but requires the crystallisation of 'similarity' definition at individual level | DA-AUC uses 'deterioration index' (e.g., creatinine for kidney function status) as the concrete 'similarity' definition for individuals |
| | Fairness Through Awareness | | – | Same as above | Same as above |
| Causality-based fairness | Counterfactual Fairness | | – | Yes, but not necessarily fair even when decisions are the same for different sensitive attribute values | DA-AUC can be evaluated in counterfactual setups (see our experiments in section 3.2 of DOI:10.24963/ijcai.2022/721) |
| | Path-specific Counterfactual Fairness | | – | Same as above | Same as above |

Fairness notions as defined by (*Castelnovo, 2022; DOI: 10.1038/s41598-022-07939-1*), their applicability in clinical decision making and DA-AUC's relation with these notions

# Fairness in resource allocation scenario

1. Confusion on which one to choose
2. ‘unfair’ target variable - y
3. Clinicians’ opinions

# Result

# Dataset and Resource allocation scenario

**HiRID:**
a freely accessible critical care dataset containing de-identified data for >33,000 ICU admissions to the Bern University Hospital, Switzerland, between 2008-2016

*M Faltys, M Zimmermann, X Lyu, M H  user, S Hyland, G R  atsch, and TM Merz. Hirid, a high time-resolution icu dataset (version 1.1.1), 2021.*

**ICU admission on HiRID**
female vs male when admitted to ICUs

# Deterioration indices for measuring health needs

| | |
|---|---|
| **Creatinine max value**<br><br>**Creatinine min value** | Readings with the first 24 hours of admission. Creatinine measures kidney functions and normal ranges chosen were:<br>- 65.4 to 119.3 micromoles/L for women<br>- 52.2 to 91.9 micromoles/L for men. |
| **ALT min value** | ALT measures liver functions and normal ranges chosen were:<br>- 30 U/L for men<br>- 19 U/L for women |
| **Normalised number of multimorbidities** | $\#MM \times \frac{65}{age}$ |

The deterioration index used a probability on **20-step** cut-offs.

## Exp1: does the deterioration index work?

For ICU admission scenario using **controlled experiments**
- can it detect when there is **no bias**?
- does it quantify the inequality **accurately**?

**Synthetic dataset generation from HiRID**
(1) randomly select 10% data from HiRID and choose all male patients out of it;
(2) randomly change the sex of 50% of the patients to female.

**no bias datasets:**
do it 10 times to get 10 synthetic datasets

**controlled bias datasets:**
do it 10 times to get 10 synthetic datasets, but for each time, gradually change the female's readings towards the healthier end
e.g., decrease max values, increase min values

# Inequality quantification evaluation

## Exp1.1: when there is no bias?

Health inequality assessments on synthetic datasets

| Measurement | mean [95% CI] | *p*-value |
|---|---|---|
| Creatinine max | 0.044 [-0.083, 0.130] | 0.0664 |
| Creatinine min | 0.024 [-0.266, 0.302] | 0.7084 |
| ALT max | 0.033 [-0.157, 0.182] | 0.4231 |

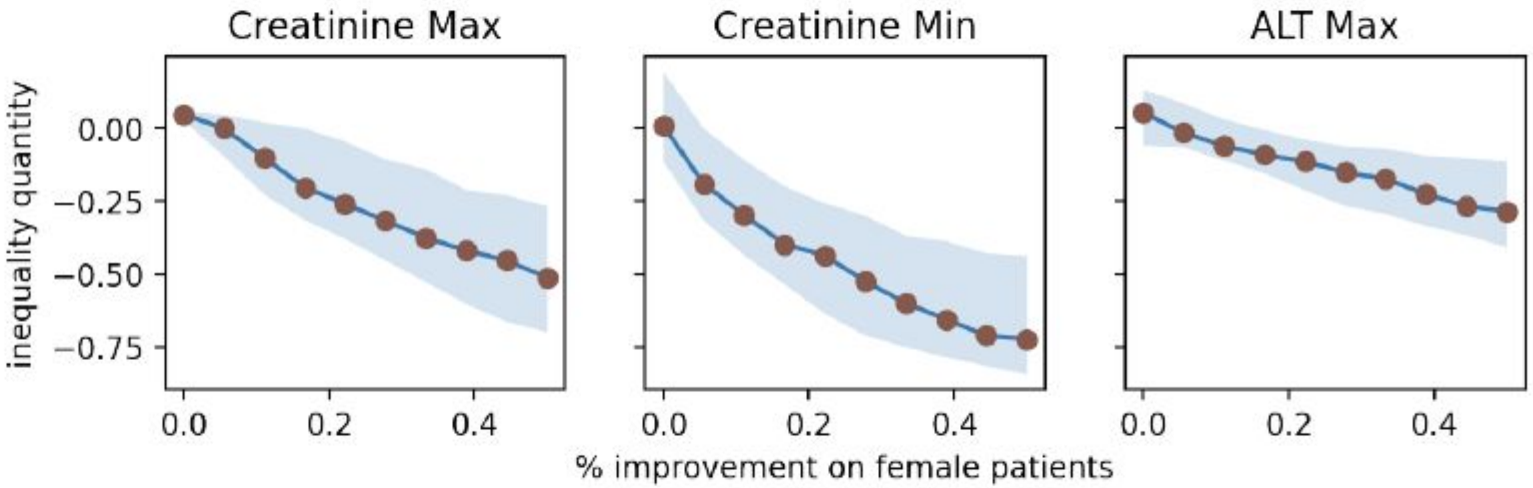
Table 3: Overall inequality of **female vs male** quantified on 10 synthetic datasets, where there should be no inequality overall.

The p-value was generated for a T-test for the null hypothesis that the mean value was equal to 0, meaning NO inequality.

p-values are not significant in all cases: could not reject the null hypothesis - meaning the mean values are 0s in all cases.

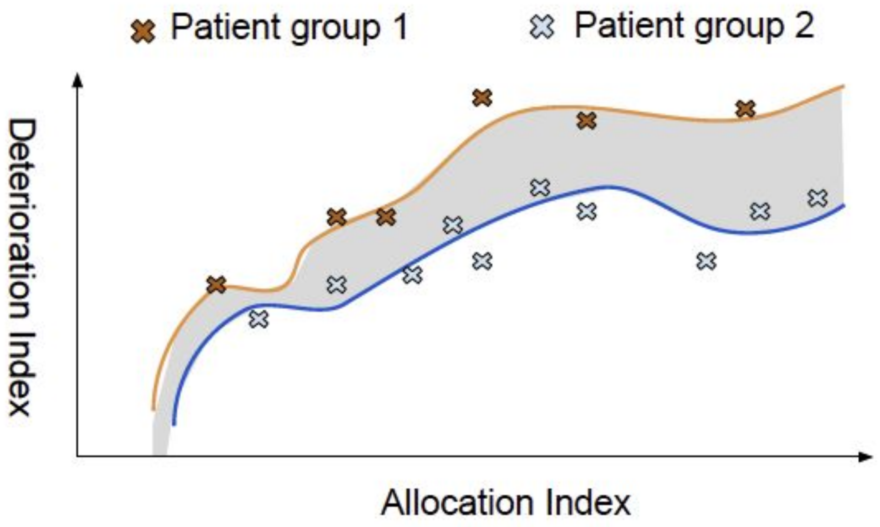
## Exp1.2: does it quantify the inequality accurately?

Figure 4: Inequality Quantification Evaluation on synthetic data: y-axis is the inequality quantity of female vs male. x-axis is the percentage of controlled improvements on readings of the female subcohort. Y-value of each point is the mean value of 10 runs on the same x-value, i.e., % of improvement. Shaded areas denote 25-75% quantile regions.

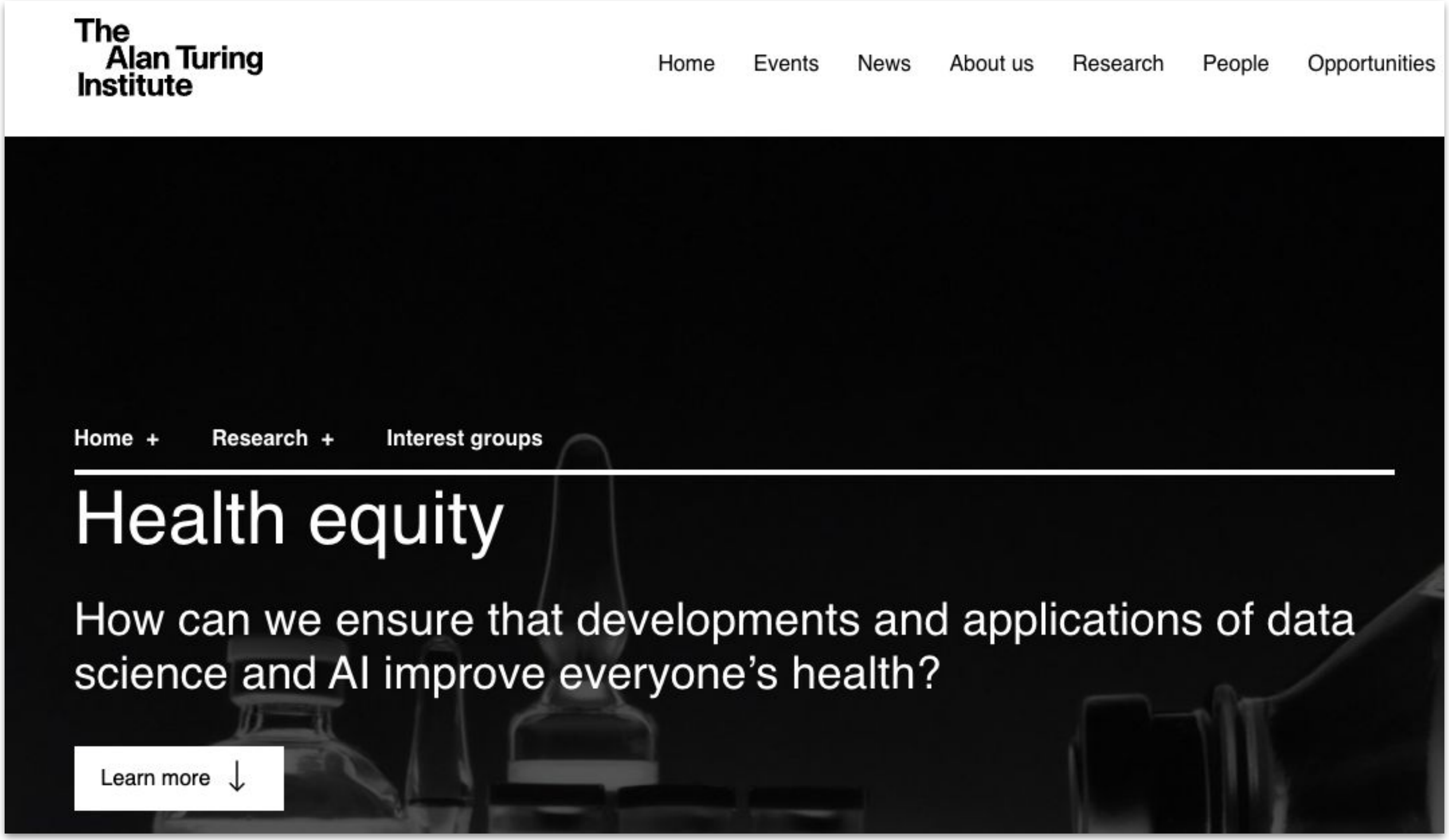



the Spearman rank-order correlation coefficients between the inequality quantities and the percentages of improvements are **-0.989, -0.974 and -0.993** for Creatinine Max/Min and ALT Max respectively.

# Summary




- There are many fairness definitions and metrics
- AI and data induced inequity are not assessed widely in the AI in medicine community
- Deterioration Allocation Framework is one effort to crystallise fairness notions for clinical decision making




https://www.turing.ac.uk/research/interest-groups/health-equity