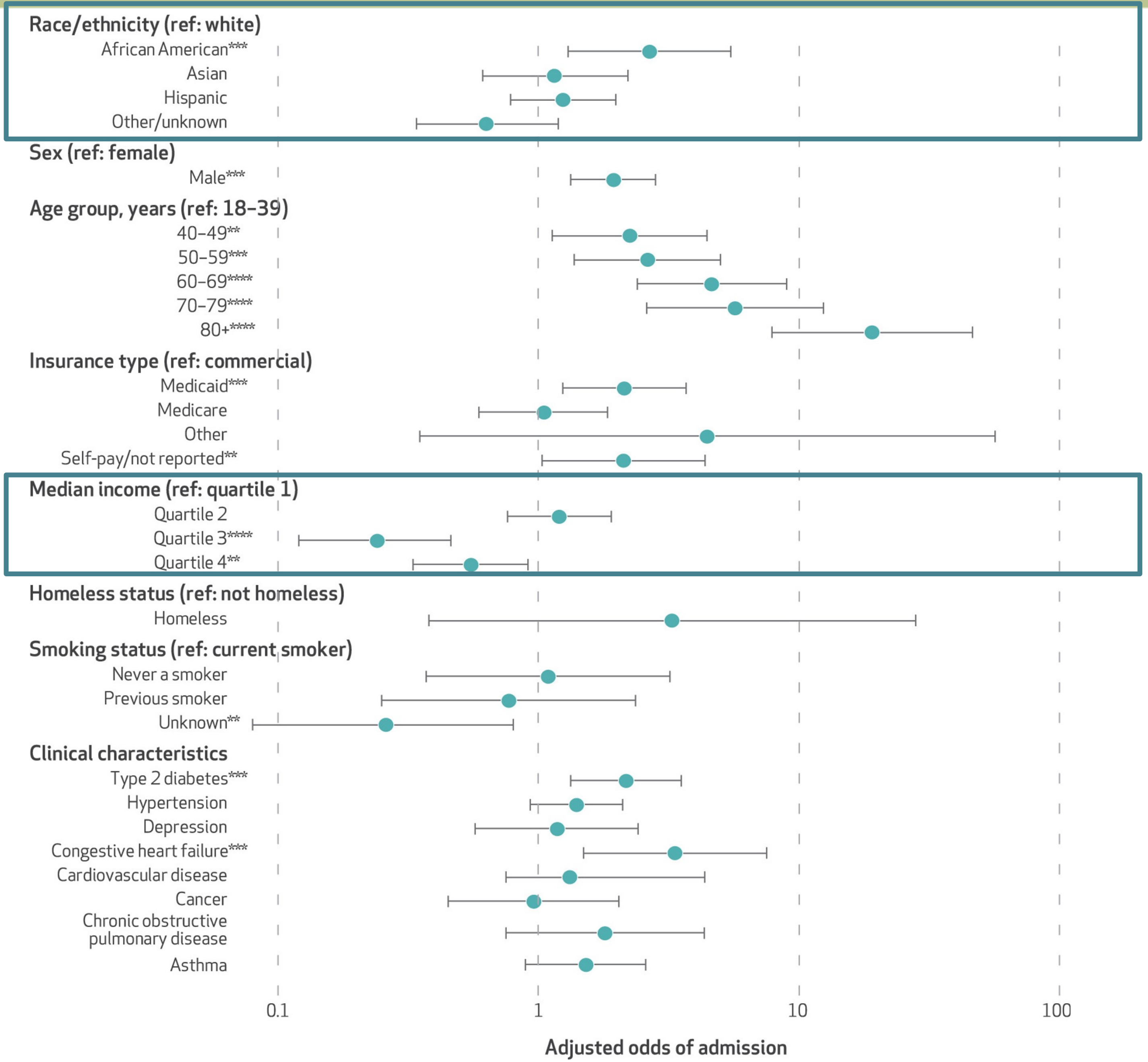
INSTITUTE OF HEALTH INFORMATICS, SCHOOL OF LIFE AND MEDICAL SCIENCES

Quantifying Health Inequalities Induced by Data and AI Models

Honghan Wu, Aneeta Sylolypavan, Minhong Wang & Sarah Wild
Associate Professor in Health Informatics | Rutherford Fellow | Turing Fellow

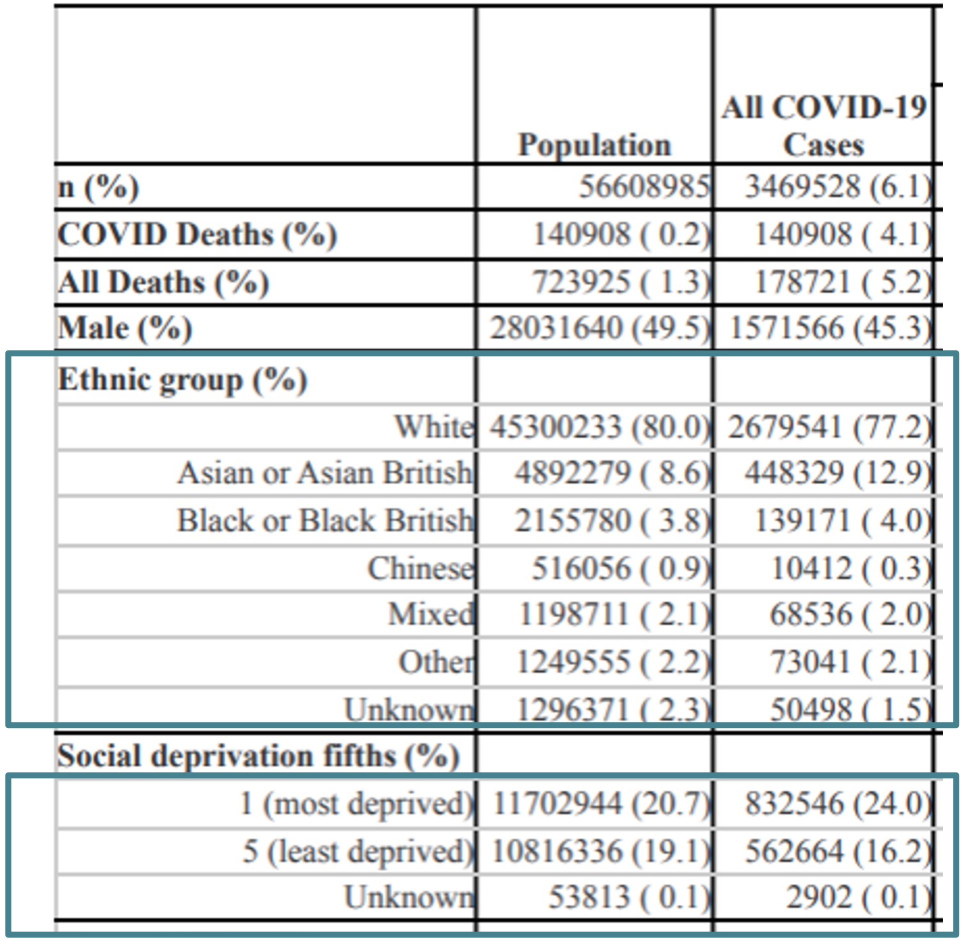
31 March 2022

Background - disparities in healthcare



Azar, Kristen MJ, et al. "Disparities In Outcomes Among COVID-19 Patients In A Large Health Care System In California: Study estimates the COVID-19 infection fatality rate at the US county level." Health Affairs 39.7 (2020): 1253-1262.

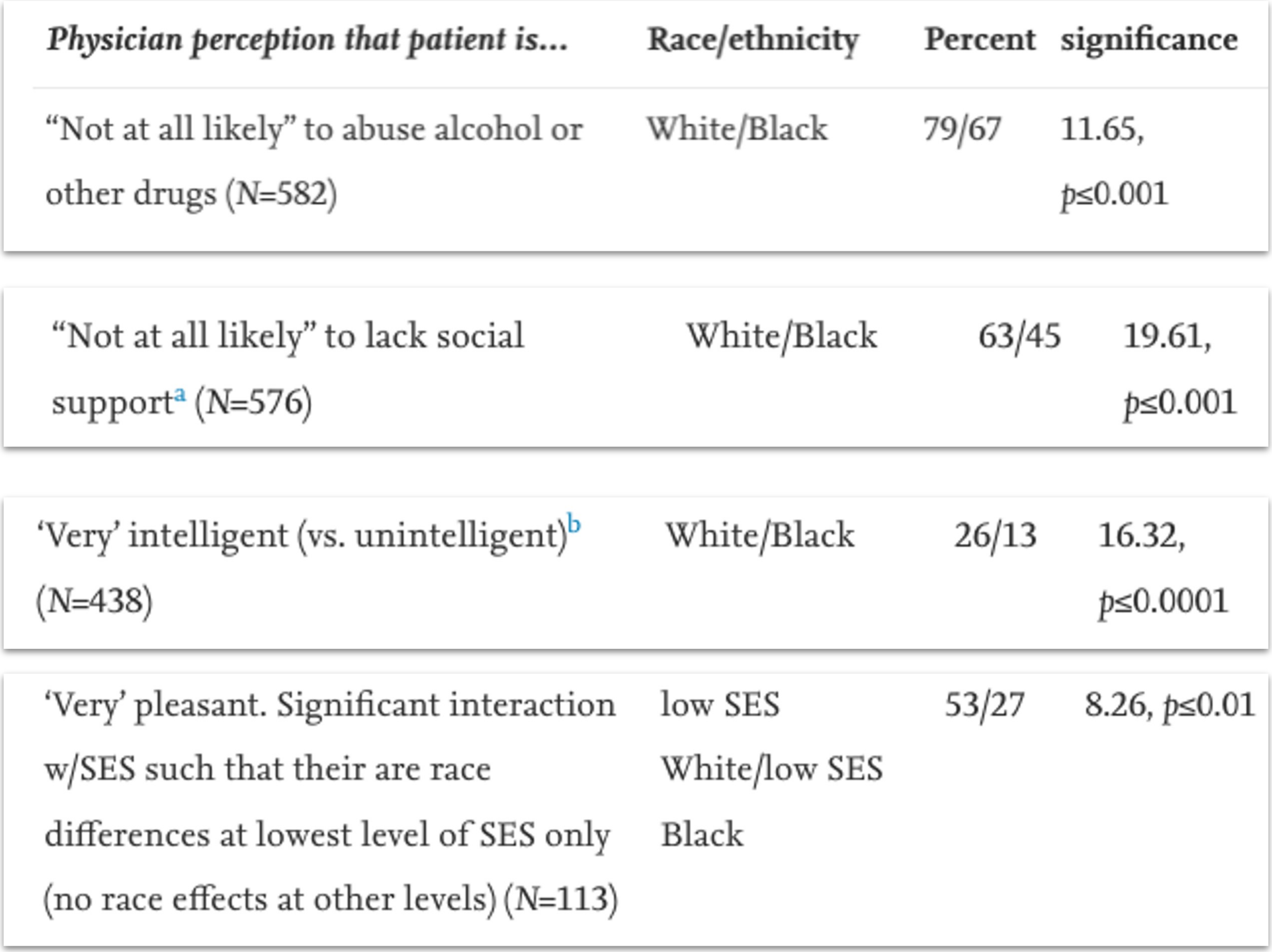
Background - disparities in healthcare



	Population	All COVID-19 Cases
n (%)	56608985	3469528 (6.1)
COVID Deaths (%)	140908 (0.2)	140908 (4.1)
All Deaths (%)	723925 (1.3)	178721 (5.2)
Male (%)	28031640 (49.5)	1571566 (45.3)
Ethnic group (%)		
White	45300233 (80.0)	2679541 (77.2)
Asian or Asian British	4892279 (8.6)	448329 (12.9)
Black or Black British	2155780 (3.8)	139171 (4.0)
Chinese	516056 (0.9)	10412 (0.3)
Mixed	1198711 (2.1)	68536 (2.0)
Other	1249555 (2.2)	73041 (2.1)
Unknown	1296371 (2.3)	50498 (1.5)
Social deprivation fifths (%)		
1 (most deprived)	11702944 (20.7)	832546 (24.0)
5 (least deprived)	10816336 (19.1)	562664 (16.2)
Unknown	53813 (0.1)	2902 (0.1)

Thygesen, Johan H., et al. "Understanding COVID-19 trajectories from a nationwide linked electronic health record cohort of 56 million people: phenotypes, severity, waves & vaccination." medRxiv (2021).

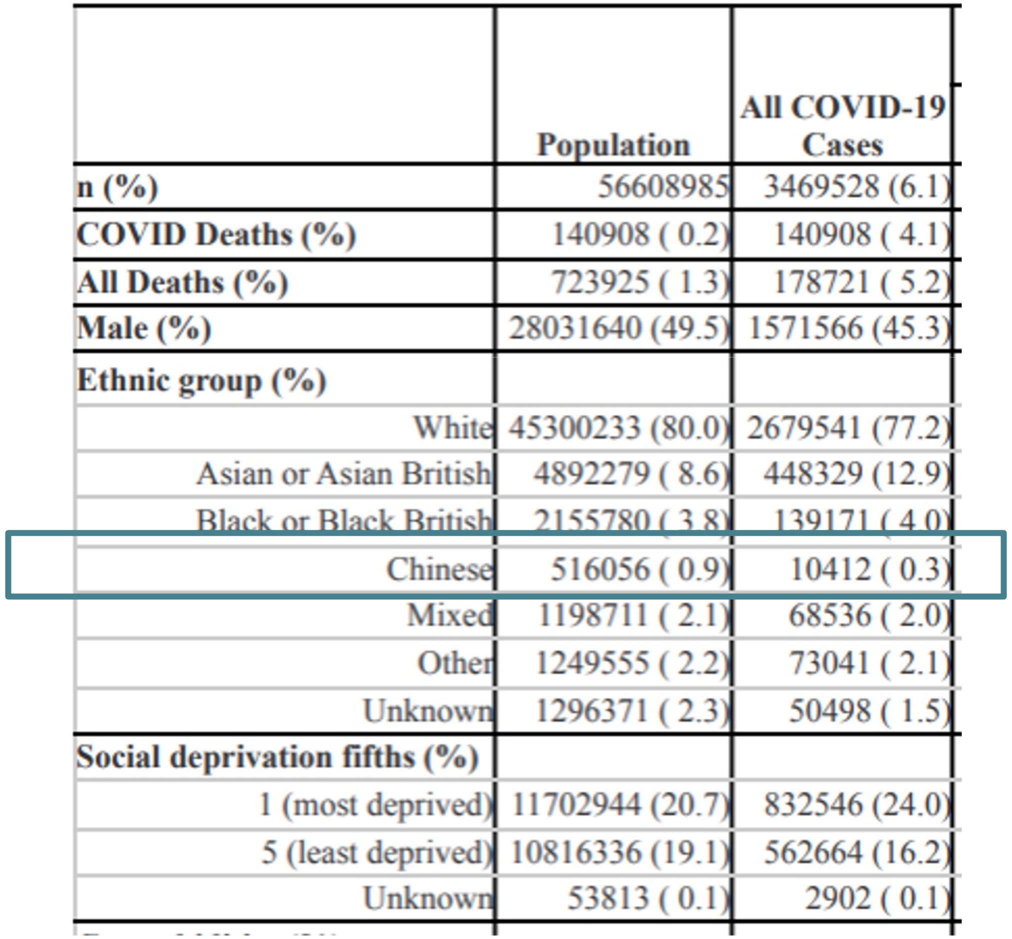
Background - inequality because of disecrimination



Physician perception that patient is...	Race/ethnicity	Percent	significance
"Not at all likely" to abuse alcohol or other drugs (N=582)	White/Black	79/67	11.65, *p*≤0.001
"Not at all likely" to lack social support[a] (N=576)	White/Black	63/45	19.61, *p*≤0.001
'Very' intelligent (vs. unintelligent)[b] (N=438)	White/Black	26/13	16.32, *p*≤0.0001
'Very' pleasant. Significant interaction w/SES such that their are race differences at lowest level of SES only (no race effects at other levels) (N=113)	low SES White/low SES Black	53/27	8.26, *p*≤0.01

Van Ryn, M., & Burke, J. (2000). The effect of patient race and socio-economic status on physicians' perceptions of patients. Social science & medicine, 50(6), 813-828.

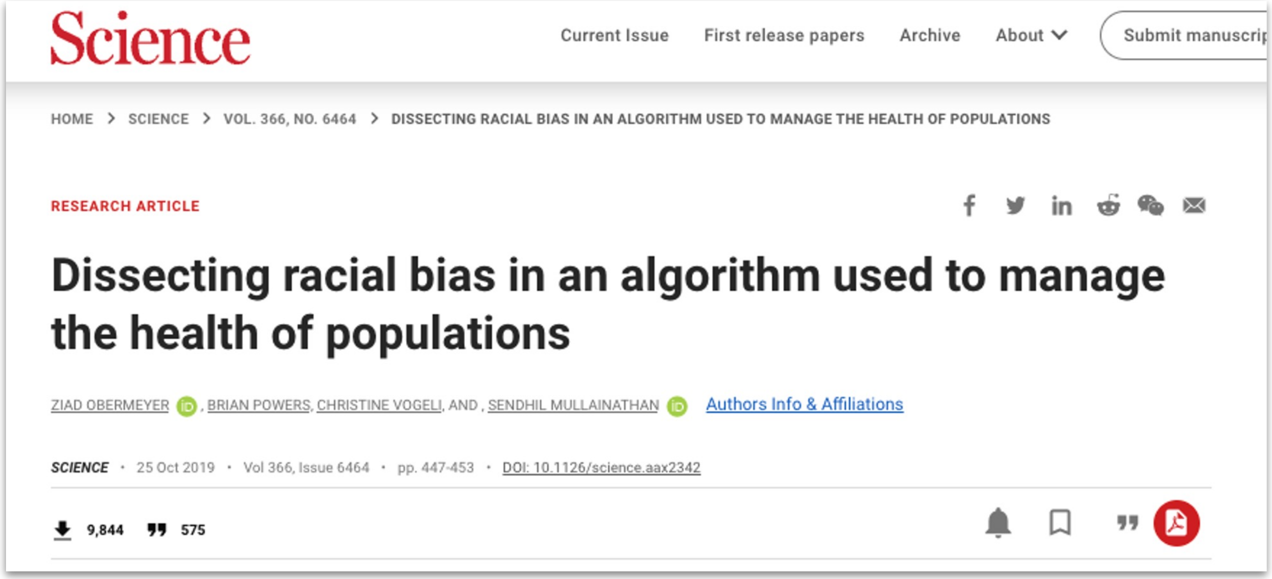
Background - inequality because of underrepresentation



	Population	All COVID-19 Cases
n (%)	56608985	3469528 (6.1)
COVID Deaths (%)	140908 (0.2)	140908 (4.1)
All Deaths (%)	723925 (1.3)	178721 (5.2)
Male (%)	28031640 (49.5)	1571566 (45.3)
Ethnic group (%)		
White	45300233 (80.0)	2679541 (77.2)
Asian or Asian British	4892279 (8.6)	448329 (12.9)
Black or Black British	2155780 (3.8)	139171 (4.0)
Chinese	516056 (0.9)	10412 (0.3)
Mixed	1198711 (2.1)	68536 (2.0)
Other	1249555 (2.2)	73041 (2.1)
Unknown	1296371 (2.3)	50498 (1.5)
Social deprivation fifths (%)		
1 (most deprived)	11702944 (20.7)	832546 (24.0)
5 (least deprived)	10816336 (19.1)	562664 (16.2)
Unknown	53813 (0.1)	2902 (0.1)

Thygesen, Johan H., et al. "Understanding COVID-19 trajectories from a nationwide linked electronic health record cohort of 56 million people: phenotypes, severity, waves & vaccination." medRxiv (2021).

Background - inequality because of model developments



HOME > SCIENCE > VOL. 366, NO. 6464 > DISSECTING RACIAL BIAS IN AN ALGORITHM USED TO MANAGE THE HEALTH OF POPULATIONS

RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

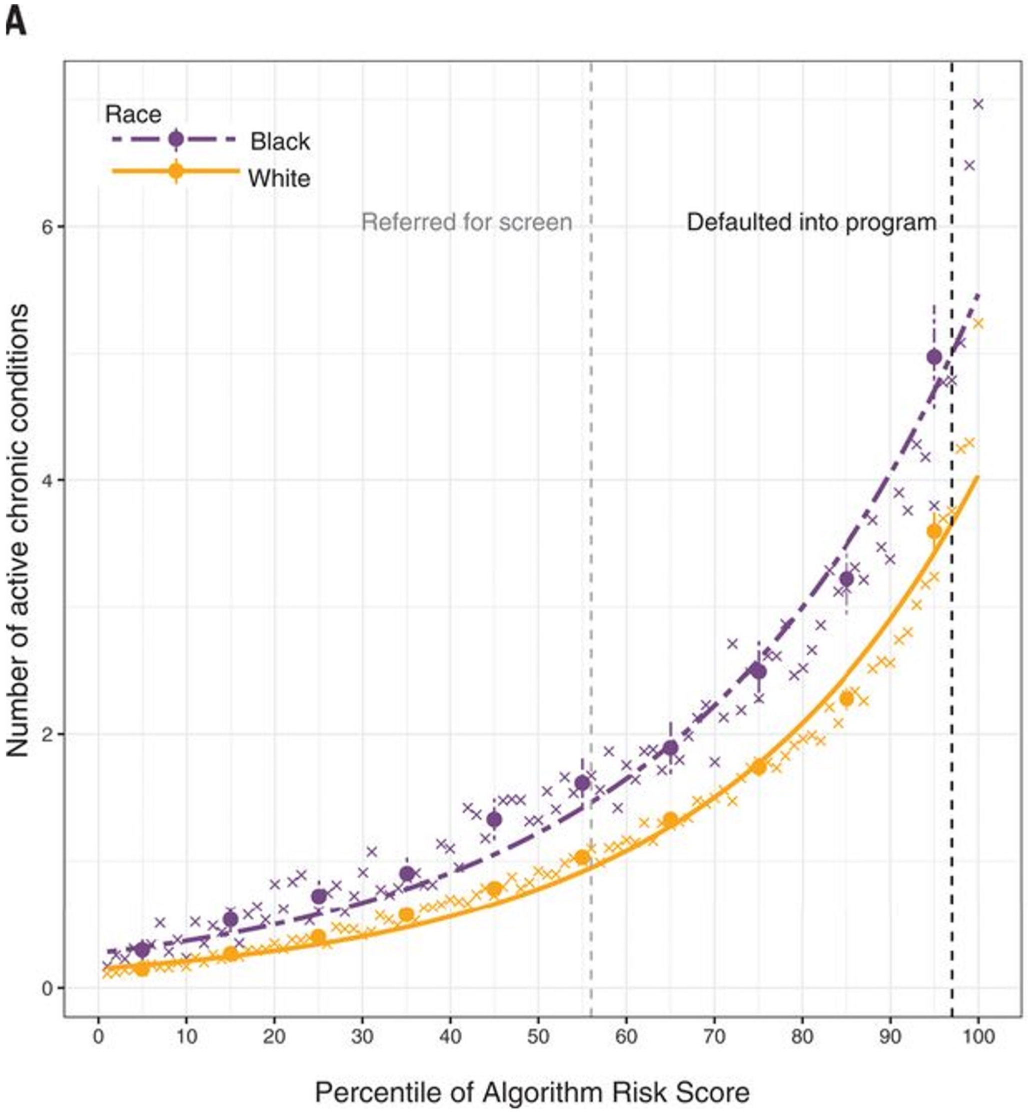
ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND, SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE • 25 Oct 2019 • Vol 366, Issue 6464 • pp. 447-453 • DOI: 10.1126/science.aax2342

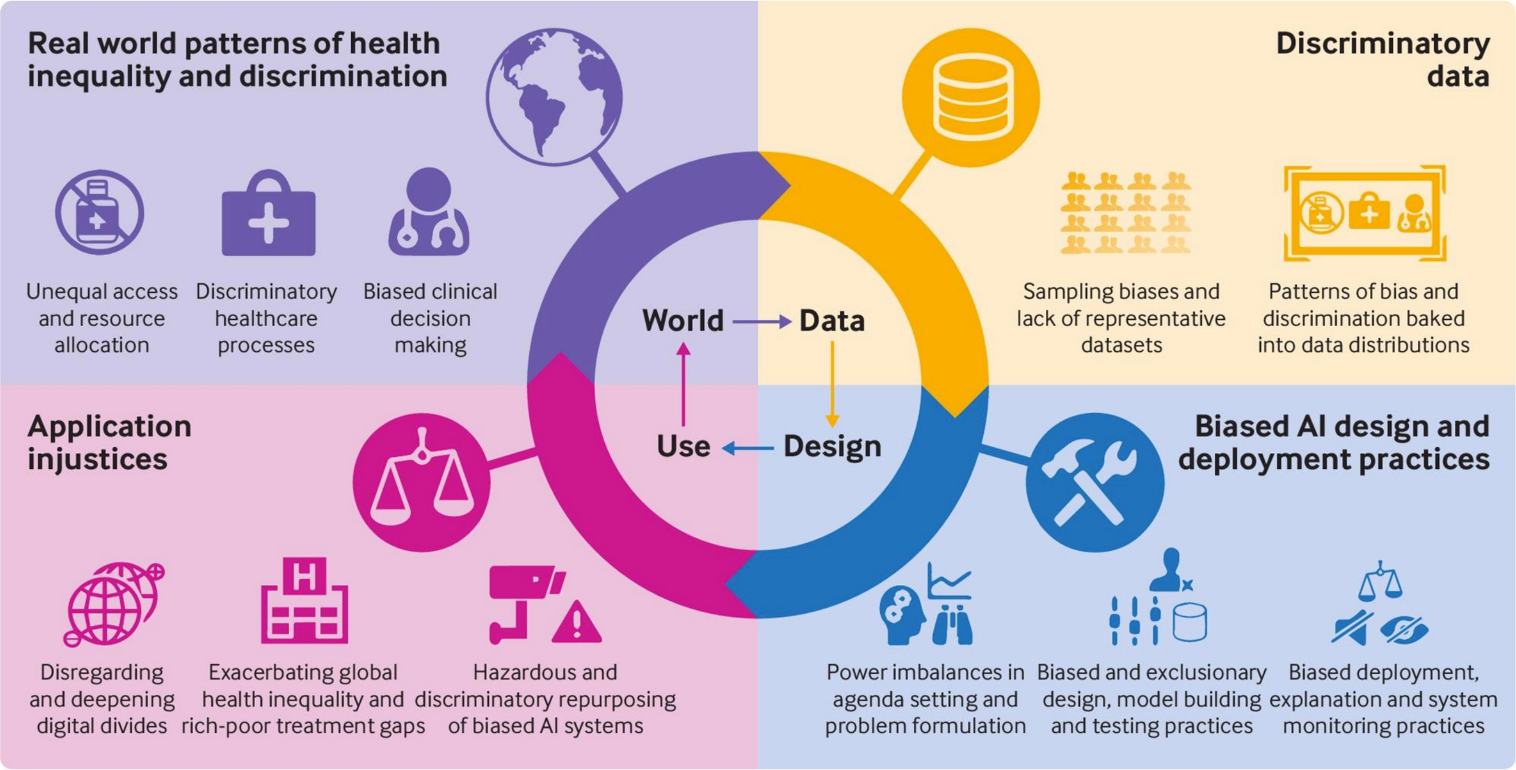
9,844 575

"

The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients.



Background - inequalities joined together


Real world patterns of health inequality and discrimination
Unequal access and resource allocation
Discriminatory healthcare processes
Biased clinical decision making
Discriminatory data
Sampling biases and lack of representative datasets
Patterns of bias and discrimination baked into data distributions
World
Data
Use
Design
Application injustices
Disregarding and deepening digital divides
Exacerbating global health inequality and rich-poor treatment gaps
Hazardous and discriminatory repurposing of biased AI systems
Biased AI design and deployment practices
Power imbalances in agenda setting and problem formulation
Biased and exclusionary design, model building and testing practices
Biased deployment, explanation and system monitoring practices
H

Leslie, David, et al. "Does "AI" stand for augmenting inequality in the era of covid-19 healthcare?." bmj 372 (2021).

Background - a snapshot of current recommendations

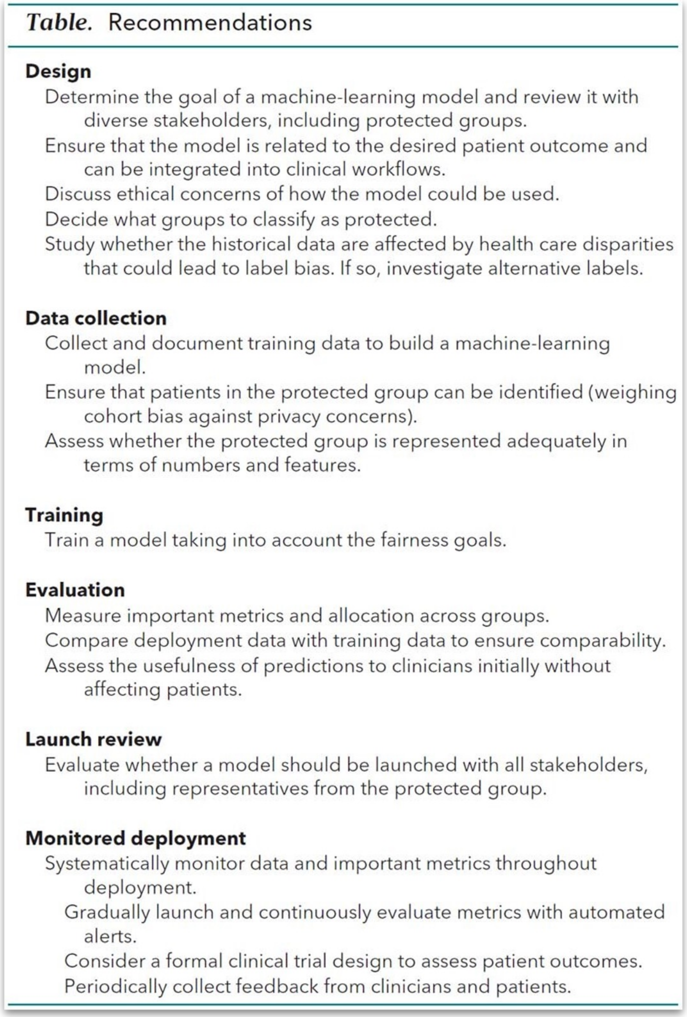


Table. Recommendations

Design

Determine the goal of a machine-learning model and review it with diverse stakeholders, including protected groups.

Ensure that the model is related to the desired patient outcome and can be integrated into clinical workflows.

Discuss ethical concerns of how the model could be used.

Decide what groups to classify as protected.

Study whether the historical data are affected by health care disparities that could lead to label bias. If so, investigate alternative labels.

Data collection

Collect and document training data to build a machine-learning model.

Ensure that patients in the protected group can be identified (weighing cohort bias against privacy concerns).

Assess whether the protected group is represented adequately in terms of numbers and features.

Training

Train a model taking into account the fairness goals.

Evaluation

Measure important metrics and allocation across groups.

Compare deployment data with training data to ensure comparability.

Assess the usefulness of predictions to clinicians initially without affecting patients.

Launch review

Evaluate whether a model should be launched with all stakeholders, including representatives from the protected group.

Monitored deployment

Systematically monitor data and important metrics throughout deployment.

Gradually launch and continuously evaluate metrics with automated alerts.

Consider a formal clinical trial design to assess patient outcomes.

Periodically collect feedback from clinicians and patients.

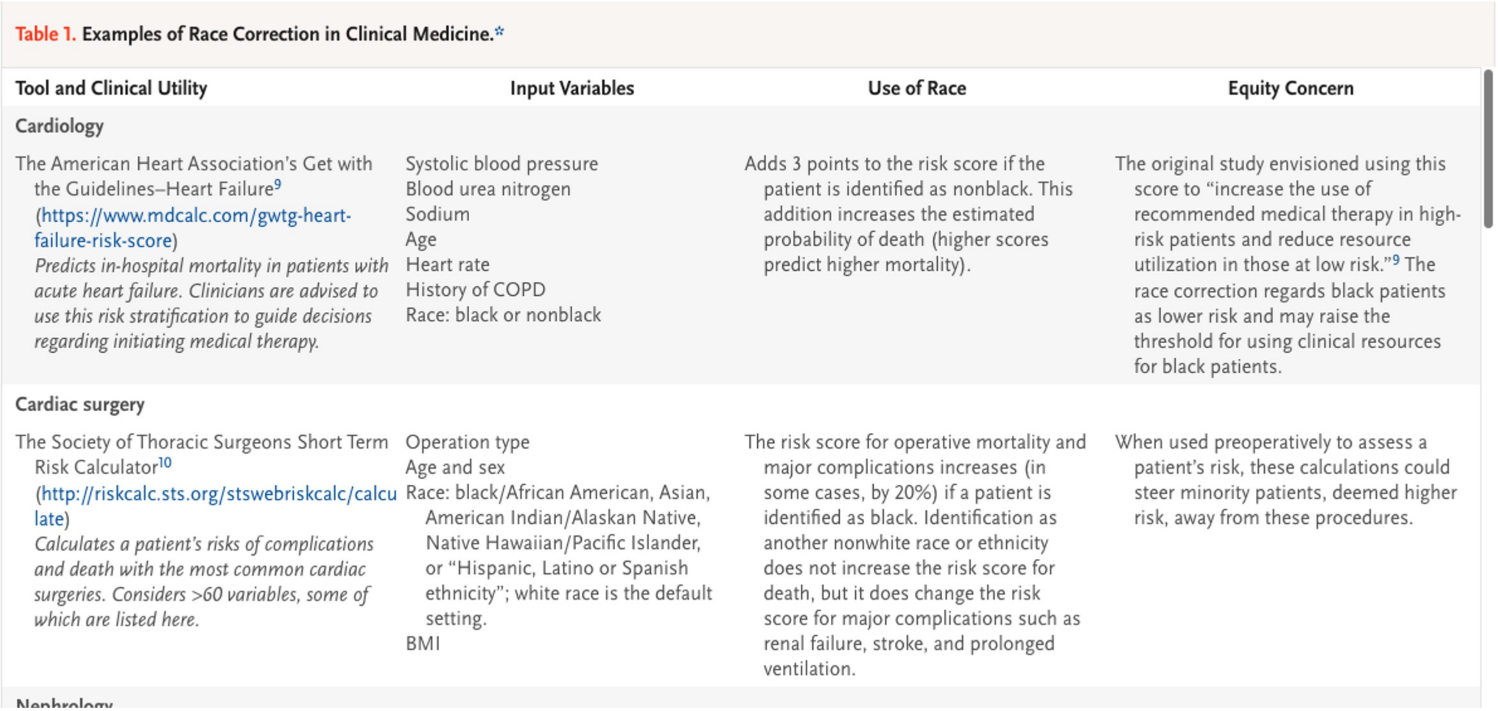


Table 1. Examples of Race Correction in Clinical Medicine.*

Tool and Clinical Utility	Input Variables	Use of Race	Equity Concern
Cardiology			
The American Heart Association's Get with the Guidelines–Heart Failure[9] (https://www.mdcalc.com/gwtg-heart-failure-risk-score) *Predicts in-hospital mortality in patients with acute heart failure. Clinicians are advised to use this risk stratification to guide decisions regarding initiating medical therapy.*	Systolic blood pressure Blood urea nitrogen Sodium Age Heart rate History of COPD Race: black or nonblack	Adds 3 points to the risk score if the patient is identified as nonblack. This addition increases the estimated probability of death (higher scores predict higher mortality).	The original study envisioned using this score to "increase the use of recommended medical therapy in high-risk patients and reduce resource utilization in those at low risk."[9] The race correction regards black patients as lower risk and may raise the threshold for using clinical resources for black patients.
Cardiac surgery			
The Society of Thoracic Surgeons Short Term Risk Calculator[10] (http://riskcalc.sts.org/stswebriskcalc/calculate) *Calculates a patient's risks of complications and death with the most common cardiac surgeries. Considers >60 variables, some of which are listed here.*	Operation type Age and sex Race: black/African American, Asian, American Indian/Alaskan Native, Native Hawaiian/Pacific Islander, or "Hispanic, Latino or Spanish ethnicity"; white race is the default setting. BMI	The risk score for operative mortality and major complications increases (in some cases, by 20%) if a patient is identified as black. Identification as another nonwhite race or ethnicity does not increase the risk score for death, but it does change the risk score for major complications such as renal failure, stroke, and prolonged ventilation.	When used preoperatively to assess a patient's risk, these calculations could steer minority patients, deemed higher risk, away from these procedures.
Nephrology			

> "*Many of these race-adjusted algorithms guide decisions in ways that may direct more attention or resources to white patients than to members of racial and ethnic minorities.*

Rajkomar, Alvin, et al. "Ensuring fairness in machine learning to advance health equity." Annals of internal medicine 169.12 (2018): 866-872.

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. New England Journal of Medicine, 383(9), 874-882.

Motivation of this work

Gap: there is no way to **quantify** health inequalities

Why quantification?

like <u>Precision/Recall/F1</u> for model accuracy

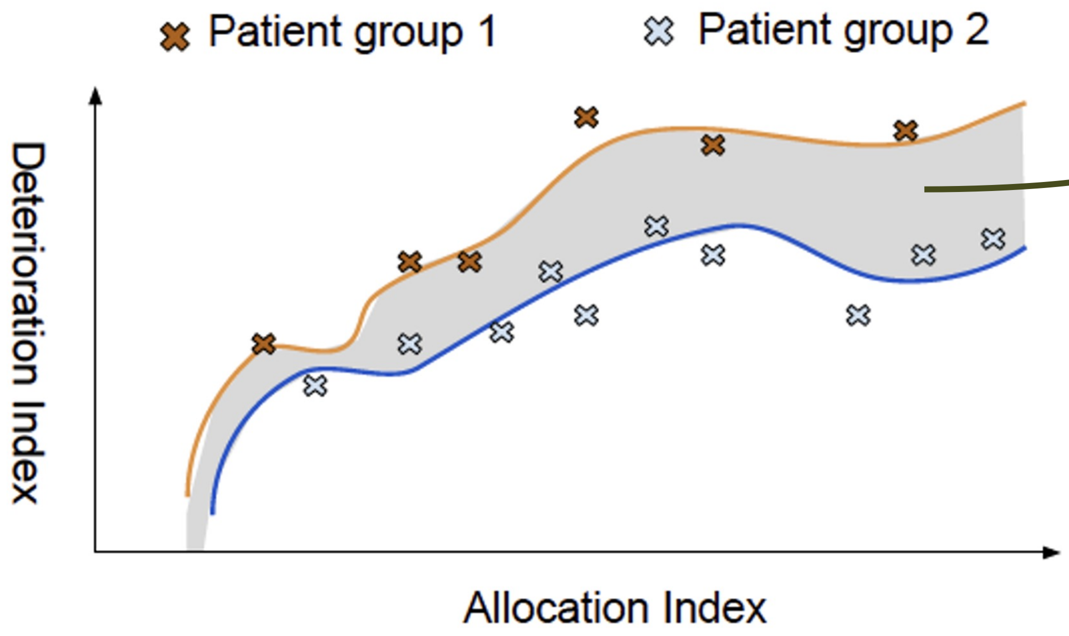
the quantification would enable debugging, evaluating and auditing potential biases in data, model developments and deployments

Method

The allocation-deterioration index

AI models =*abstracted*=> Resource Allocators

Deterioration index measures the deterioration status of patients (marker of prognosis)



Area under allocation-deterioration curves

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

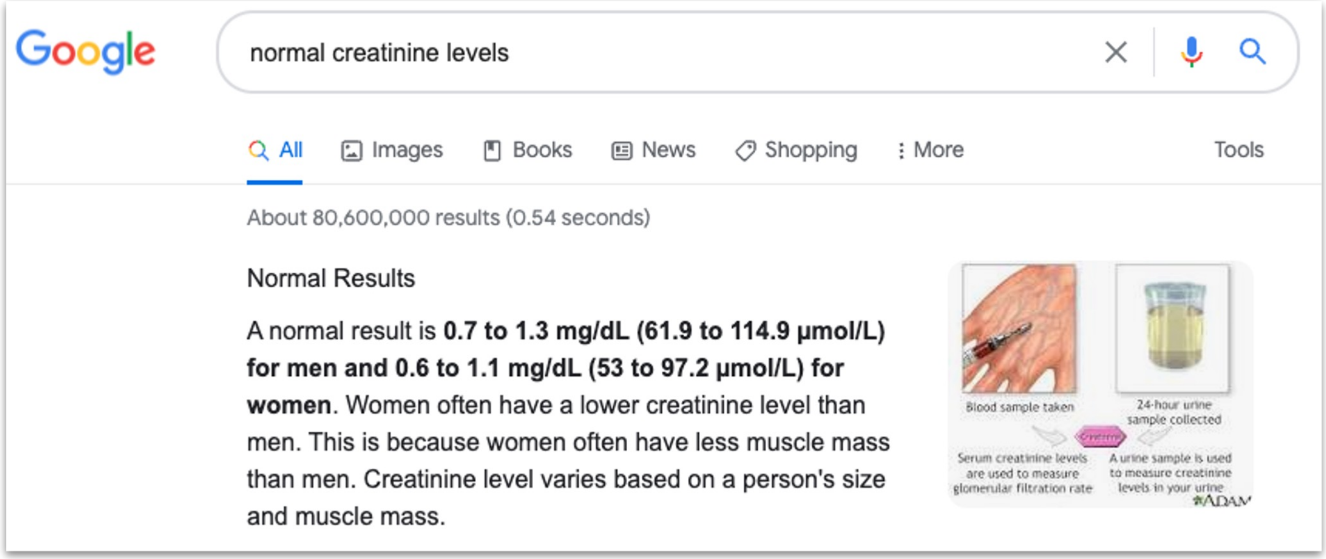
The deterioration index - formalisation

For a group of patients $P = \{p_1, p_2, ..., p_n\}$

with a numeric measurement function $m\colon P \to \mathbb{R}$

The deterioration index is $d\colon \mathbb{P}(P) \xrightarrow{m} [0, 1]$

The deterioration status is usually quantified as the degree to which the measured value is in excess of what is normal.



Google

normal creatinine levels

All · Images · Books · News · Shopping · More · Tools

About 80,600,000 results (0.54 seconds)

Normal Results

A normal result is **0.7 to 1.3 mg/dL (61.9 to 114.9 µmol/L) for men and 0.6 to 1.1 mg/dL (53 to 97.2 µmol/L) for women**. Women often have a lower creatinine level than men. This is because women often have less muscle mass than men. Creatinine level varies based on a person's size and muscle mass.

Blood sample taken
24-hour urine sample collected
Creatinine
Serum creatinine levels are used to measure glomerular filtration rate
A urine sample is used to measure creatinine levels in your urine
ADAM

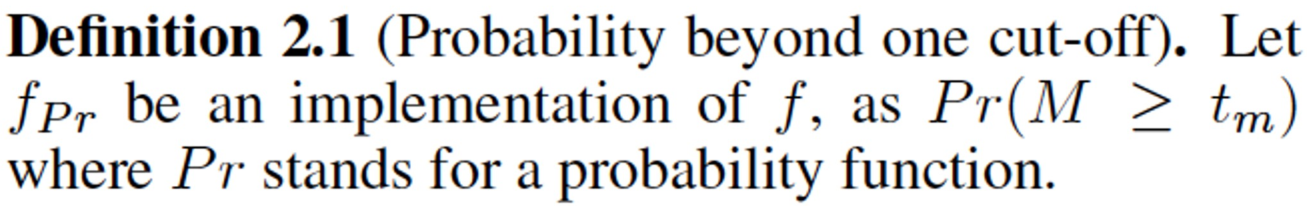
$$\{m(p) | p \in P\}$$

$$d(P; m) = f(\{M_1, M_2, ..., M_n\}; t_m)$$

A threshold

The deterioration index - definition 1

$$d(P;m) = f(\{M_1, M_2, ..., M_n\}; t_m)$$



Definition 2.1 (Probability beyond one cut-off)**.** Let f_{Pr} be an implementation of f, as $Pr(M \geq t_m)$ where Pr stands for a probability function.

Use Creatinine as m, two groups of patients: P1 and P2
P1: fpr=0.6
P2: fpr=0.3

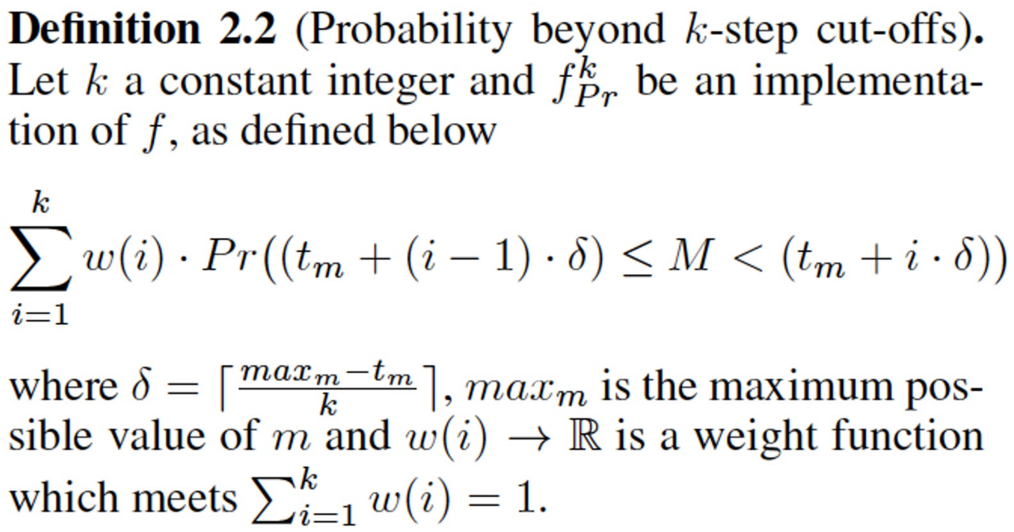
P1 is more deteriorated than P2 in terms of their kidney functions.

The deterioration index - definition 2

Implementation 1 does not quantify the level of exceeding the limit

Use Creatinine as m, two groups of patients: P1 and P2
M of P1: {0.8, 0.78, 10}
M of P2: {0.8, 0.78, 1.36}

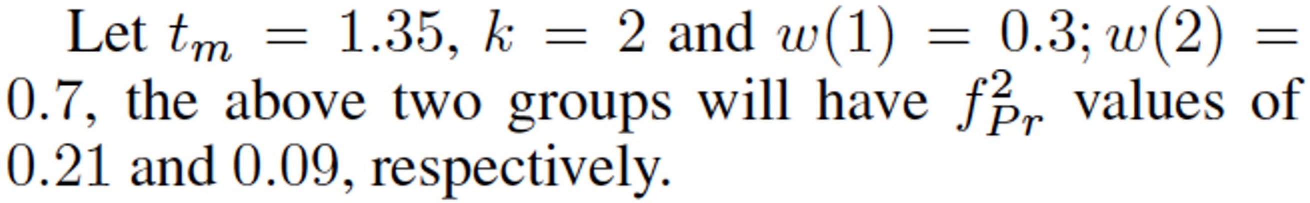
For fpr(M;1.35), then
P1: 0.3
P2: 0.3
However, P1 is clearly more deteriorated.



Definition 2.2 (Probability beyond k-step cut-offs)**.** Let k a constant integer and f^k_{Pr} be an implementation of f, as defined below

$$\sum_{i=1}^{k} w(i) \cdot Pr((t_m + (i-1)\cdot\delta) \leq M < (t_m + i\cdot\delta))$$

where $\delta = \lceil \frac{max_m - t_m}{k} \rceil$, max_m is the maximum possible value of m and $w(i) \rightarrow \mathbb{R}$ is a weight function which meets $\sum_{i=1}^{k} w(i) = 1$.



Let $t_m = 1.35$, $k = 2$ and $w(1) = 0.3; w(2) = 0.7$, the above two groups will have f^2_{Pr} values of 0.21 and 0.09, respectively.

The deterioration index - implementation

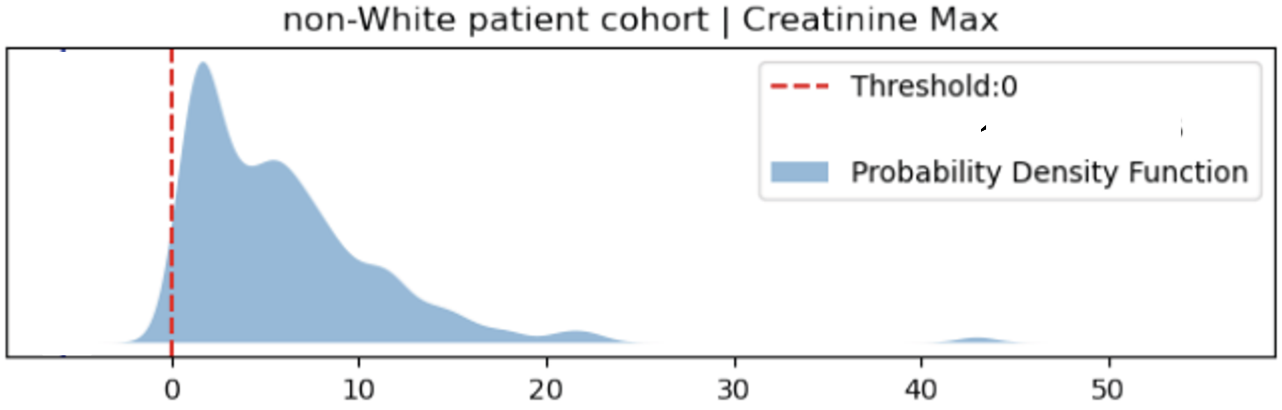
Use kernel density estimation to estimate the probability density function (PDF) of *Pr*

$$\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{v-M_i}{h}\right)$$

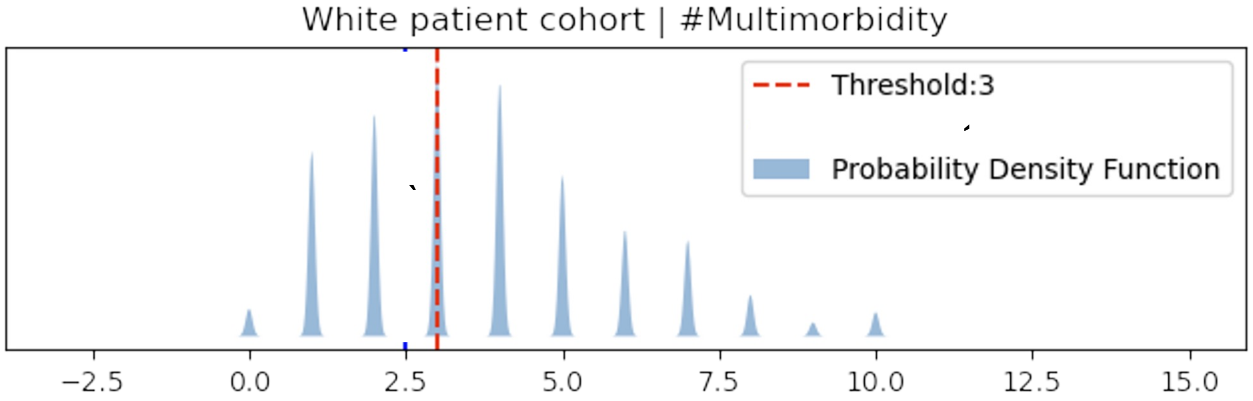
Gaussian kernel

$$exp(-v^2/2)/\sqrt{2\pi}$$

The deterioration index - boundary bias



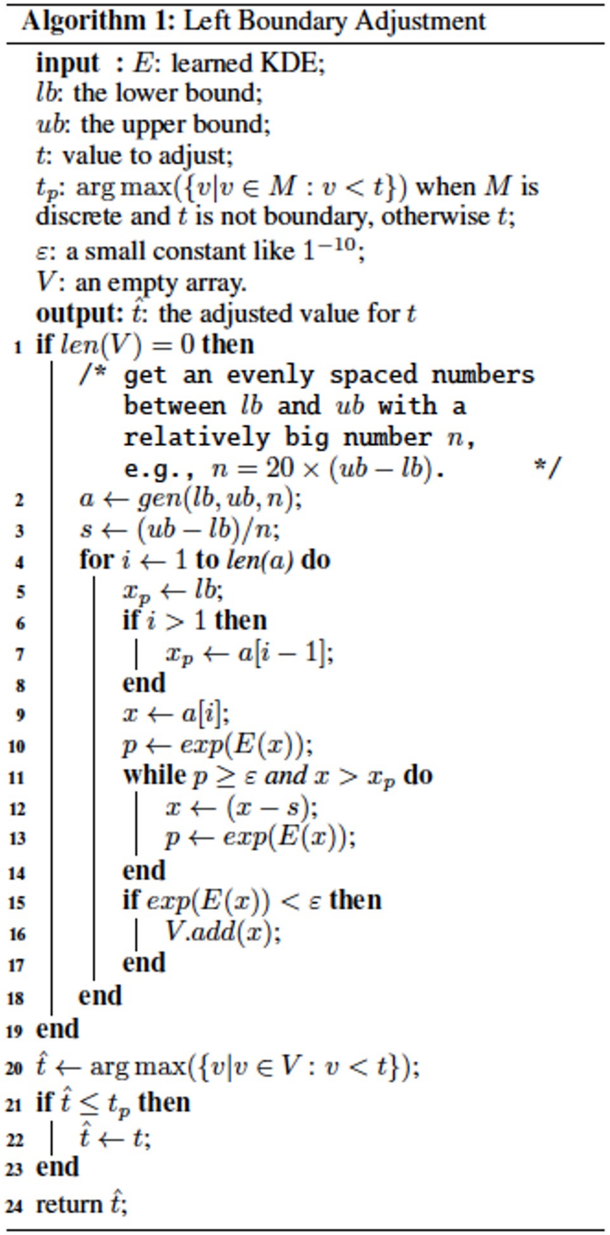
a PDF estimated for maximum Creatinine readings (ranged from 0 to 50) of a patient cohort from the MIMIC-III dataset



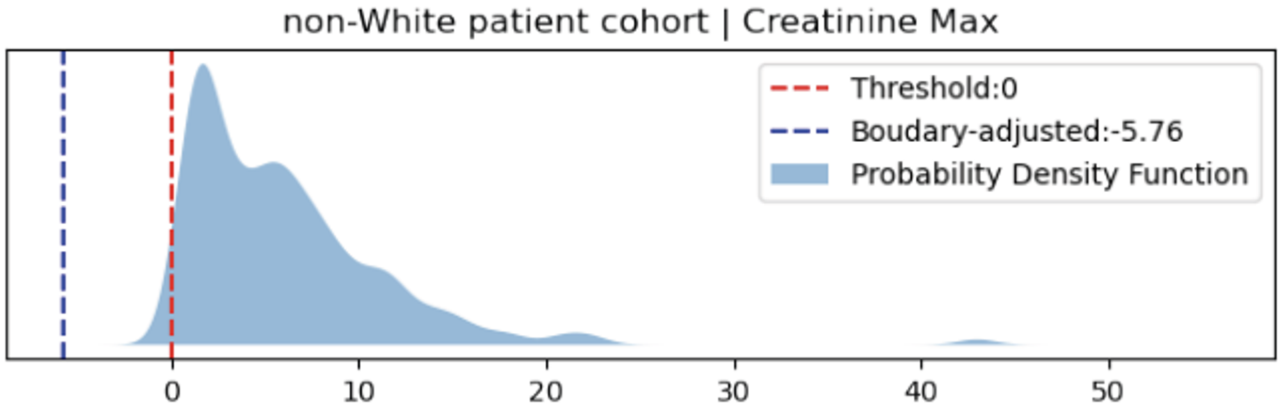
pulse-like PDFs for discrete random variables: # multimorbidities of a cohort of MIMIC-III patients

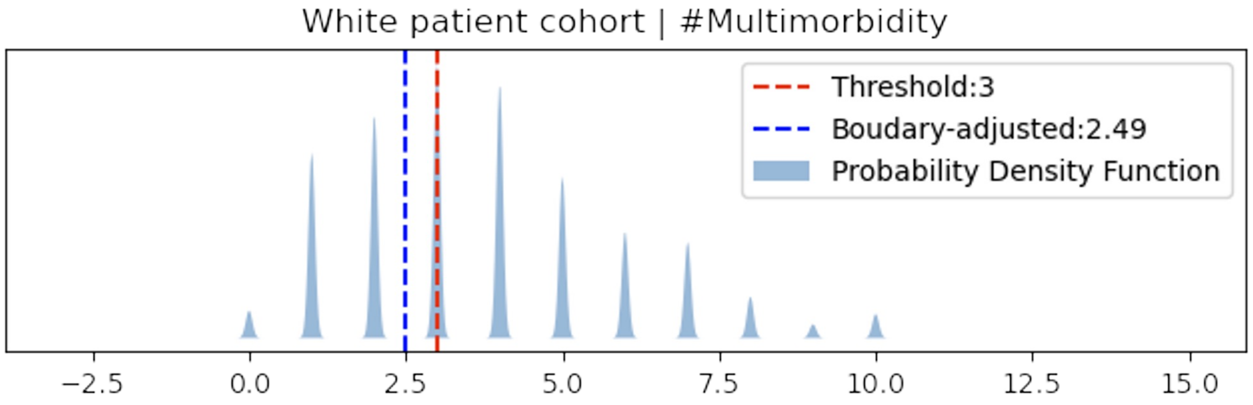
Gery Geenens. Probit transformation for kernel density estimation on the unit interval. Journal of the American Statistical Association, 109(505):346–358, 2014.

The deterioration index - boundary adjustment algorithm



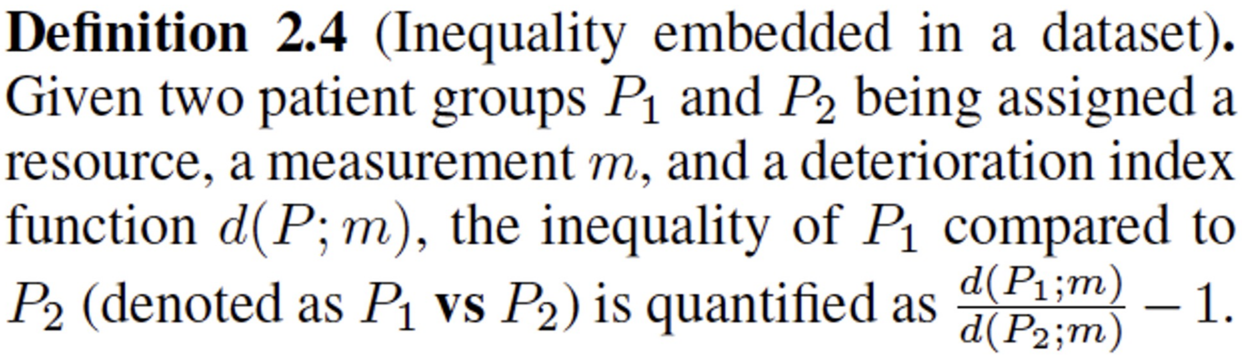
```
Algorithm 1: Left Boundary Adjustment
input : $E$: learned KDE;
    $lb$: the lower bound;
    $ub$: the upper bound;
    $t$: value to adjust;
    $t_p$: $\arg\max(\{v|v \in M : v < t\})$ when $M$ is
    discrete and $t$ is not boundary, otherwise $t$;
    $\varepsilon$: a small constant like $1^{-10}$;
    $V$: an empty array.
output: $\hat{t}$: the adjusted value for $t$
if $len(V) = 0$ then
    /* get an evenly spaced numbers
       between $lb$ and $ub$ with a
       relatively big number $n$,
       e.g., $n = 20 \times (ub - lb)$.        */
    $a \leftarrow gen(lb, ub, n)$;
    $s \leftarrow (ub - lb)/n$;
    for $i \leftarrow 1$ to $len(a)$ do
        $x_p \leftarrow lb$;
        if $i > 1$ then
            $x_p \leftarrow a[i-1]$;
        end
        $x \leftarrow a[i]$;
        $p \leftarrow exp(E(x))$;
        while $p \geq \varepsilon$ and $x > x_p$ do
            $x \leftarrow (x - s)$;
            $p \leftarrow exp(E(x))$;
        end
        if $exp(E(x)) < \varepsilon$ then
            $V.add(x)$;
        end
    end
end
$\hat{t} \leftarrow \arg\max(\{v|v \in V : v < t\})$;
if $\hat{t} \leq t_p$ then
    $\hat{t} \leftarrow t$;
end
return $\hat{t}$;
```





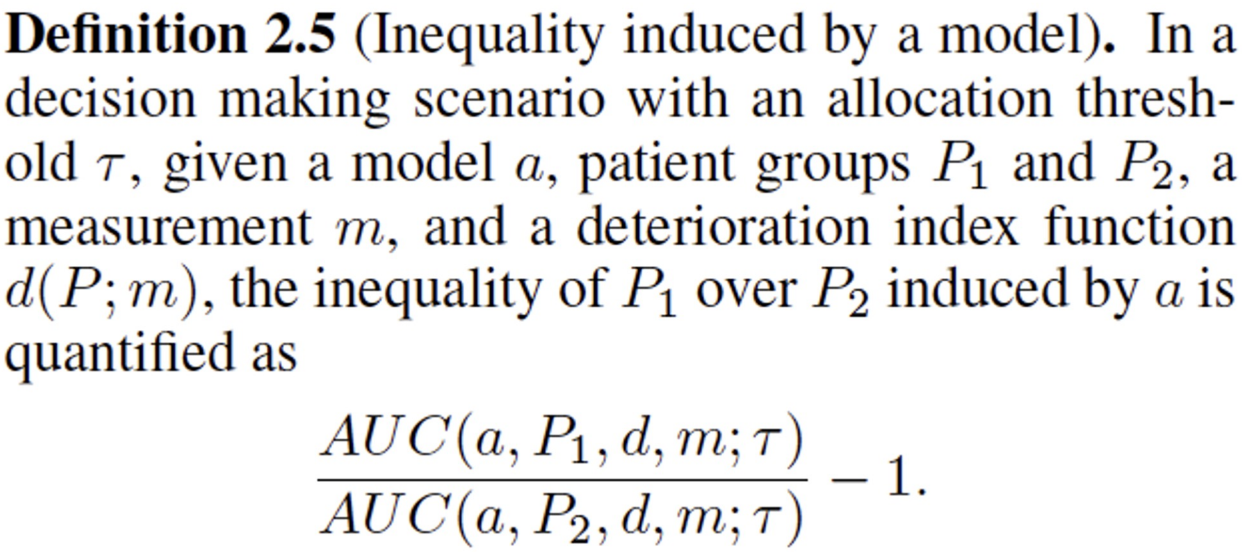
Inequality quantification definitions

Dataset



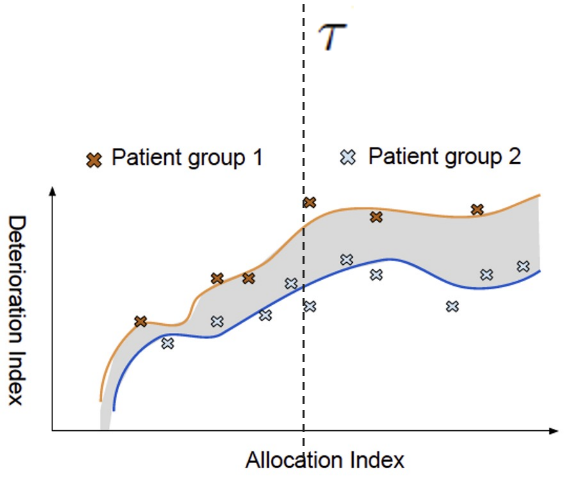
Definition 2.4 (Inequality embedded in a dataset)**.** Given two patient groups P_1 and P_2 being assigned a resource, a measurement m, and a deterioration index function $d(P; m)$, the inequality of P_1 compared to P_2 (denoted as P_1 **vs** P_2) is quantified as $\frac{d(P_1;m)}{d(P_2;m)} - 1$.

AI model



Definition 2.5 (Inequality induced by a model)**.** In a decision making scenario with an allocation threshold τ, given a model a, patient groups P_1 and P_2, a measurement m, and a deterioration index function $d(P; m)$, the inequality of P_1 over P_2 induced by a is quantified as

$$\frac{AUC(a, P_1, d, m; \tau)}{AUC(a, P_2, d, m; \tau)} - 1.$$



Results

Datasets and cohorts

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.

MIMIC-III:
a freely available database containing de-identified data for >40,000 ICU patients of the Beth Israel Deaconess Medical Centre, Boston, United States, between 2001-2012

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. Scientific data, 3(1):1–9, 2016.

Two case-control cohorts from MIMIC-III for two resource allocation scenarios for operations

(1) Renal Autotransplantation:
146 patients were identified using the ICD-9-CM Procedure Code 55.69. A control cohort (N=438) was then matched up using 1:3 ratio based on ethnicity, gender and age (+/- 3 years). The total cohort size is 584;

(2) Operations on Kidney:
584 patients were identified using the ICD-9-CM Procedure Code 55.xx, where ‘x’ means wildcard. A similar control matching method was used and identified 1,752 control patients. The total cohort size is 2,336.

Measurements & Deterioration Index Definition

Creatinine max value

Creatinine min value

Readings with the first 24 hours of admission. Creatinine measures kidney functions and normal ranges chosen were:

- 65.4 to 119.3 micromoles/L for women
- 52.2 to 91.9 micromoles/L for men.

ALT min value

ALT measures liver functions and normal ranges chosen were:

- 30 U/L for men
- 19 U/L for women

Normalised number of multimorbidities

$$\#MM \times \frac{65}{age}$$

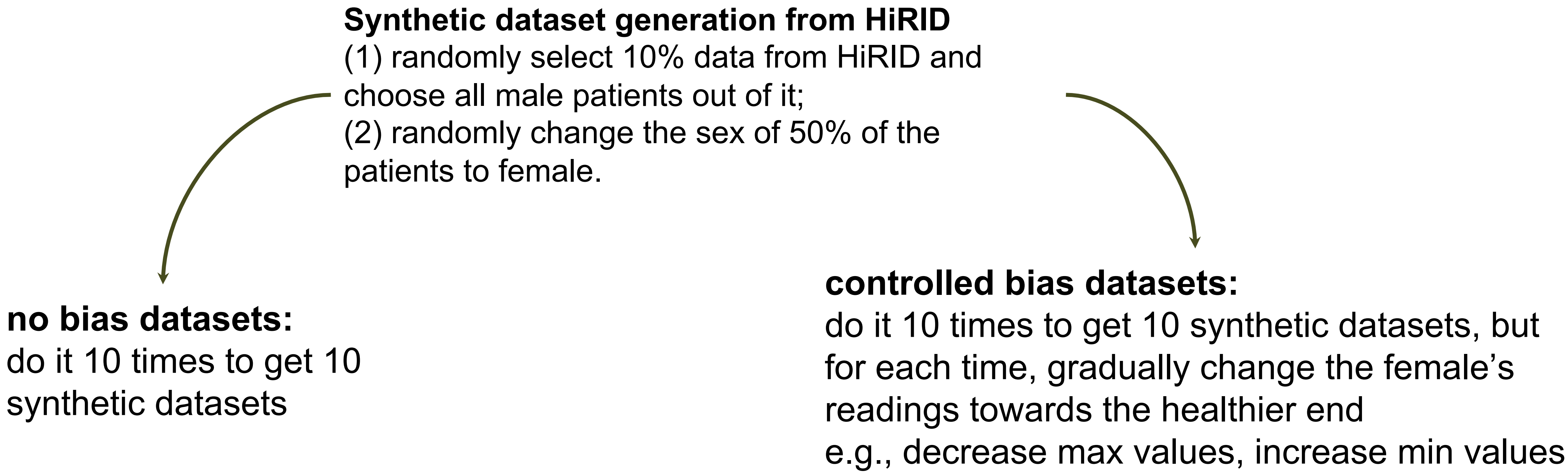
The deterioration index used a probability on **20-step** cut-offs.

Inequality quantification evaluation

does the deterioration index work?

For ICU admission scenario:

- can it detect when there is no bias?
- does it quantify the inequality accurately?



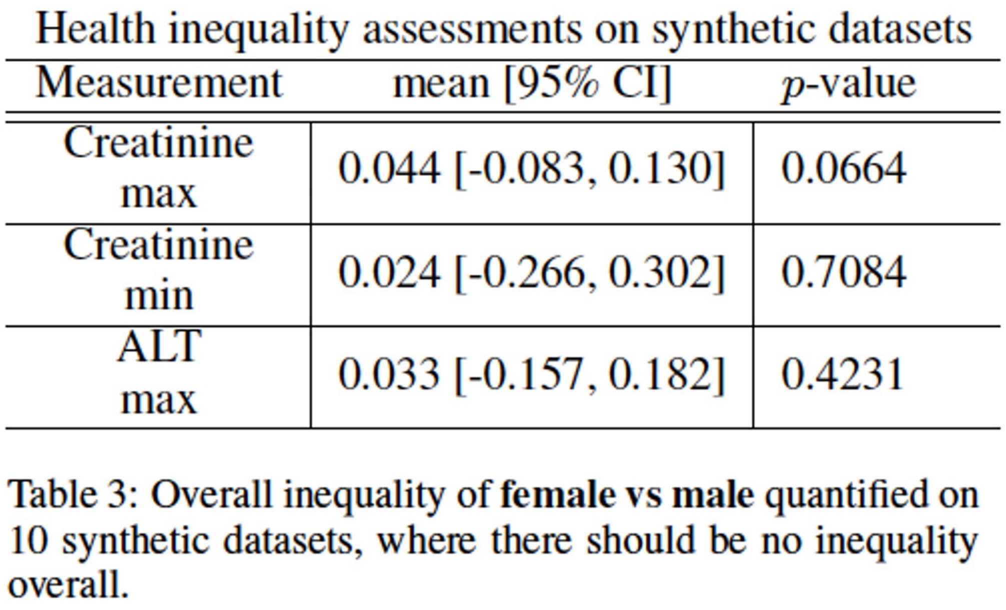
Synthetic dataset generation from HiRID
(1) randomly select 10% data from HiRID and choose all male patients out of it;
(2) randomly change the sex of 50% of the patients to female.

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

- **can it detect 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.

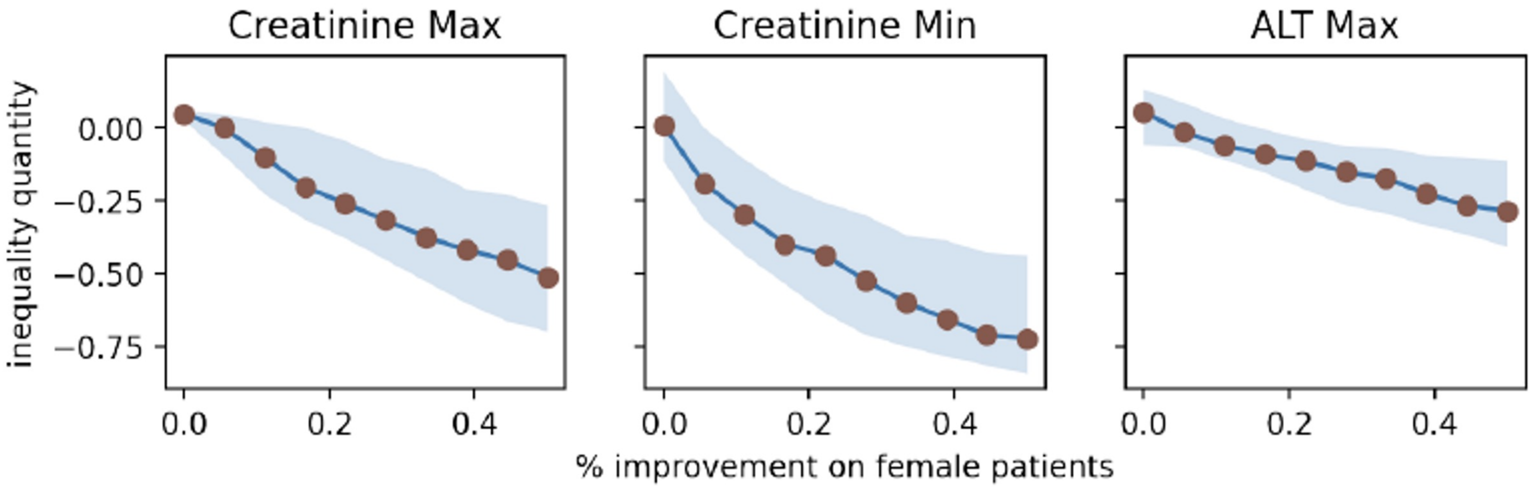
The p-value was generated for a T-test for the null hypothesis that the mean value was equal to 0, meaning NO inequality.

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

Inequality quantification evaluation

- **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.



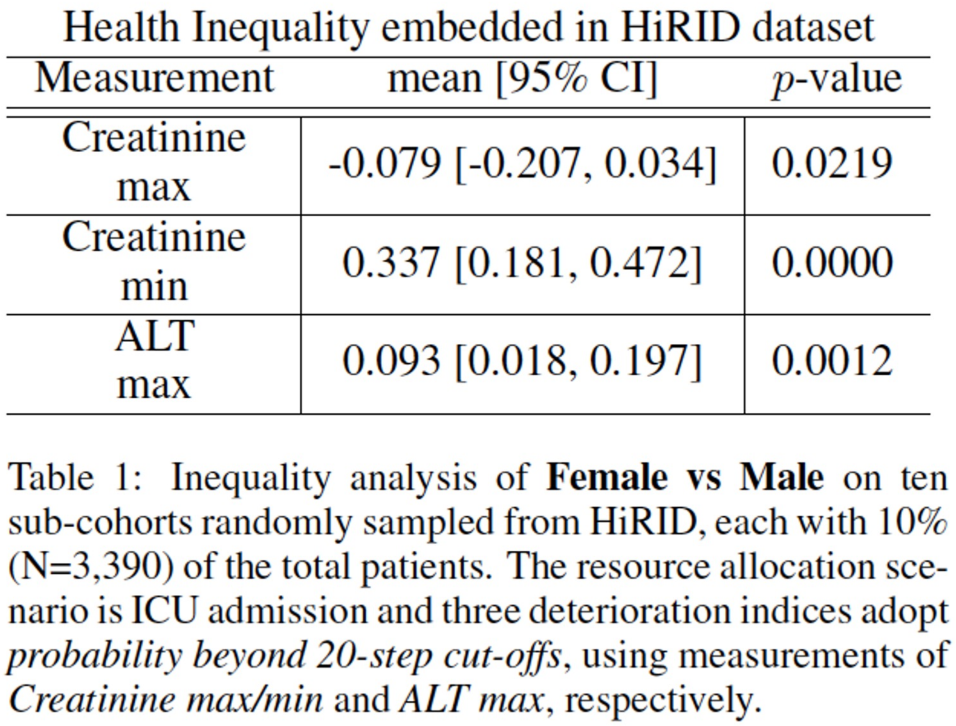
the Spearman rank-order correlation coefficients between the inequality quantities and the percentages of improvements are **-0.989, -0.974 and -0.993** for Creatinine Max/Min and ALT Max respectively.

Dataset embedded inequalities

- **ICU admission to HiRID: female vs male**

experiment datasets:

- randomly select 10% of HiRID patients (n=3,390)
- do it 10 times => 10 datasets



Health Inequality embedded in HiRID dataset

Measurement	mean [95% CI]	*p*-value
Creatinine max	-0.079 [-0.207, 0.034]	0.0219
Creatinine min	0.337 [0.181, 0.472]	0.0000
ALT max	0.093 [0.018, 0.197]	0.0012

Table 1: Inequality analysis of **Female vs Male** on ten sub-cohorts randomly sampled from HiRID, each with 10% (N=3,390) of the total patients. The resource allocation scenario is ICU admission and three deterioration indices adopt *probability beyond 20-step cut-offs*, using measurements of *Creatinine max/min* and *ALT max*, respectively.

Dataset embedded inequalities

- **Operations on Kidney: non-White patients vs White patients**

experiment dataset:

- Operations on Kidney - a cohort with 2,336 patients

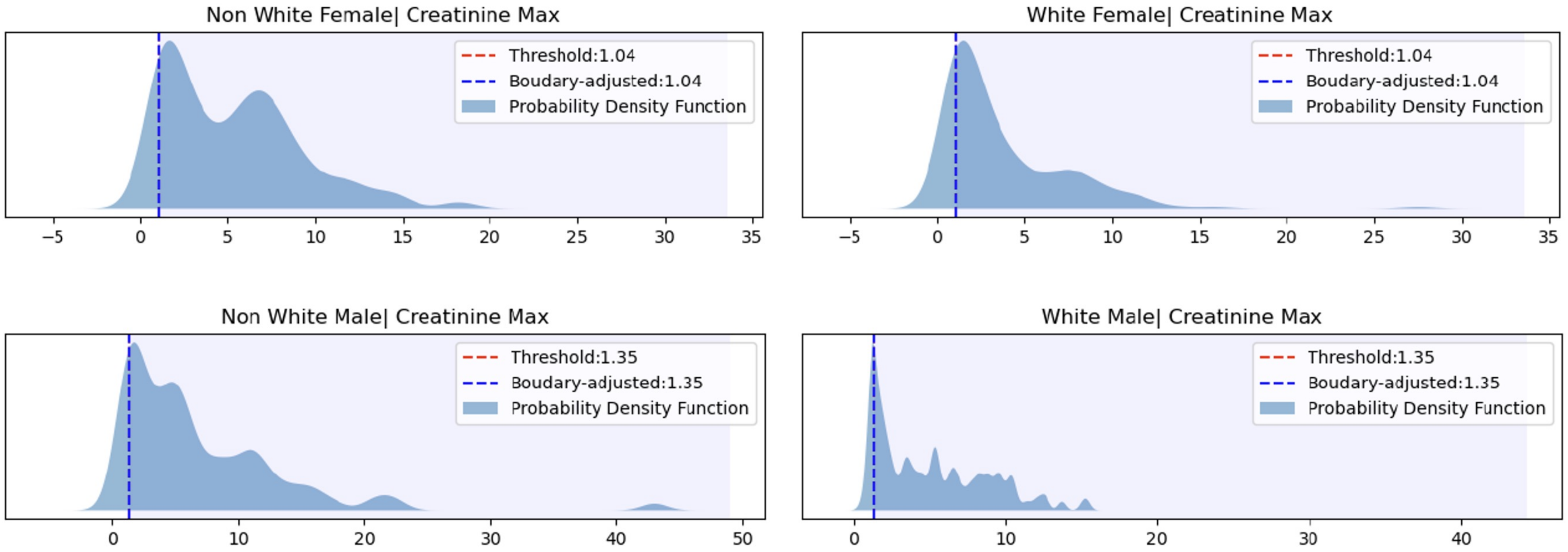


Figure 5: Probability density functions for quantifying inequalities of **non-White vs White** in the scenario of kidney operations in MIMIC-III dataset. Dashed lines denote thresholds (i.e., boundary values of abnormal readings) for computing deterioration index. Shaded area are regions where the probability integral happens for getting the deterioration index. The above two figures are females, which illustrate an inequality of 35.06%. The bottom two are males, where there is an inequality of 19.94%.

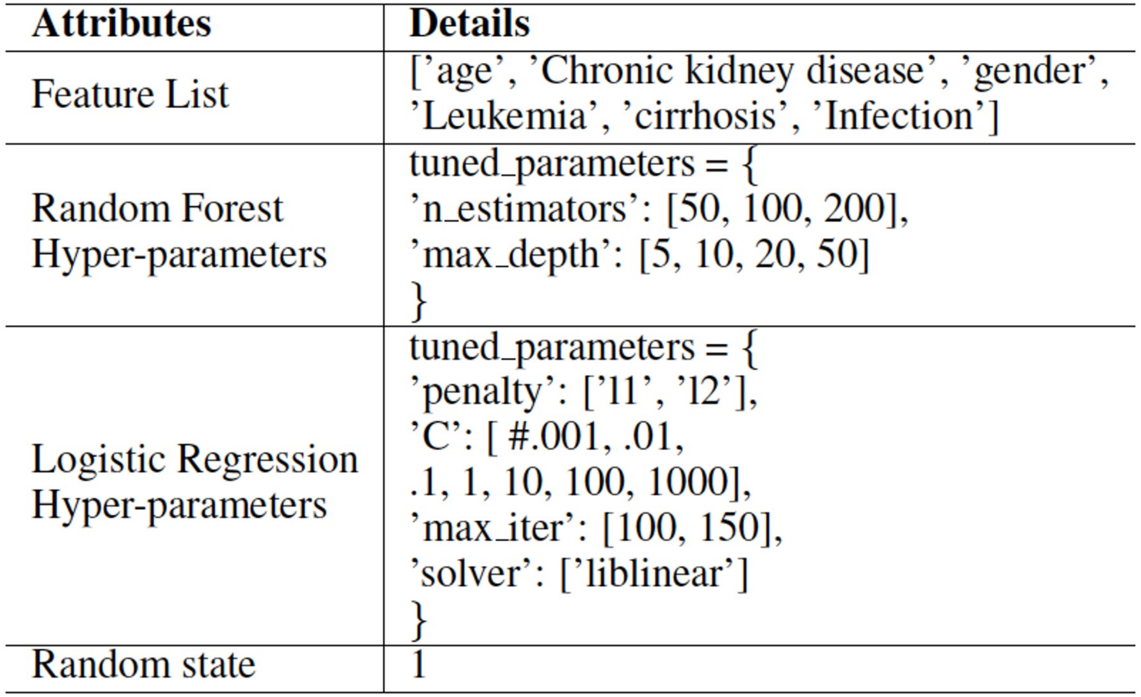
Model induced inequalities

- **Two kidney operations: non-White patients vs White patients**

experiment datasets:

- Operations on Kidney - a cohort with 2,336 patients
- Renal Autotransplantation - a cohort of 584 patients

AI models



Attributes	Details
Feature List	[’age’, ’Chronic kidney disease’, ’gender’, ’Leukemia’, ’cirrhosis’, ’Infection’]

Random Forest Hyper-parameters	tuned_parameters = {
	’n_estimators’: [50, 100, 200],
	’max_depth’: [5, 10, 20, 50]
	}

Logistic Regression Hyper-parameters	tuned_parameters = {
	’penalty’: [’l1’, ’l2’],
	’C’: [#.001, .01,
	.1, 1, 10, 100, 1000],
	’max_iter’: [100, 150],
	’solver’: [’liblinear’]
	}

Random state	1

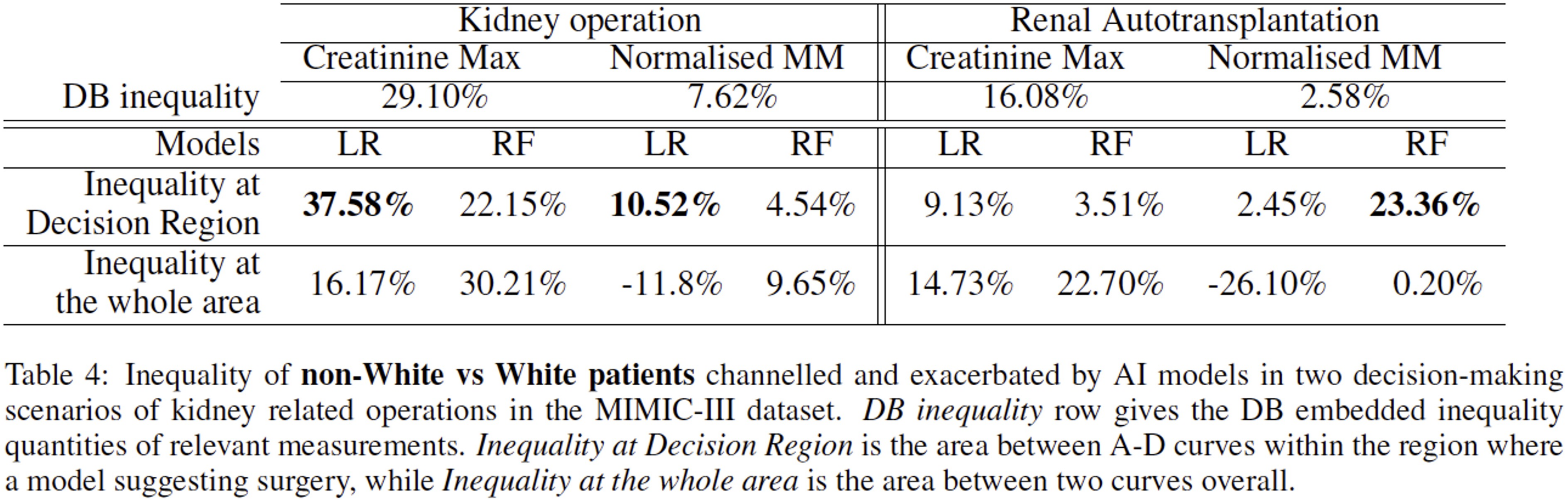
Table 5: AI Model’s hyperparameters and other reproducible setups

Performances (ROCAUC)

LR:
0.795 (IQR:0.784-0.805) and
0.867 (IQR:0.843-0.891) for Operations on Kidney and Renal Autotransplantation, respectively

RF:
0.830 (0.816-0.844) and
0.878 (0.853-0.904), respectively.

Model induced inequalities



	Kidney operation				Renal Autotransplantation			
	Creatinine Max		Normalised MM		Creatinine Max		Normalised MM	
DB inequality	29.10%		7.62%		16.08%		2.58%	
Models	LR	RF	LR	RF	LR	RF	LR	RF
Inequality at Decision Region	**37.58%**	22.15%	**10.52%**	4.54%	9.13%	3.51%	2.45%	**23.36%**
Inequality at the whole area	16.17%	30.21%	-11.8%	9.65%	14.73%	22.70%	-26.10%	0.20%

Table 4: Inequality of **non-White vs White patients** channelled and exacerbated by AI models in two decision-making scenarios of kidney related operations in the MIMIC-III dataset. *DB inequality* row gives the DB embedded inequality quantities of relevant measurements. *Inequality at Decision Region* is the area between A-D curves within the region where a model suggesting surgery, while *Inequality at the whole area* is the area between two curves overall.

Model induced inequalities

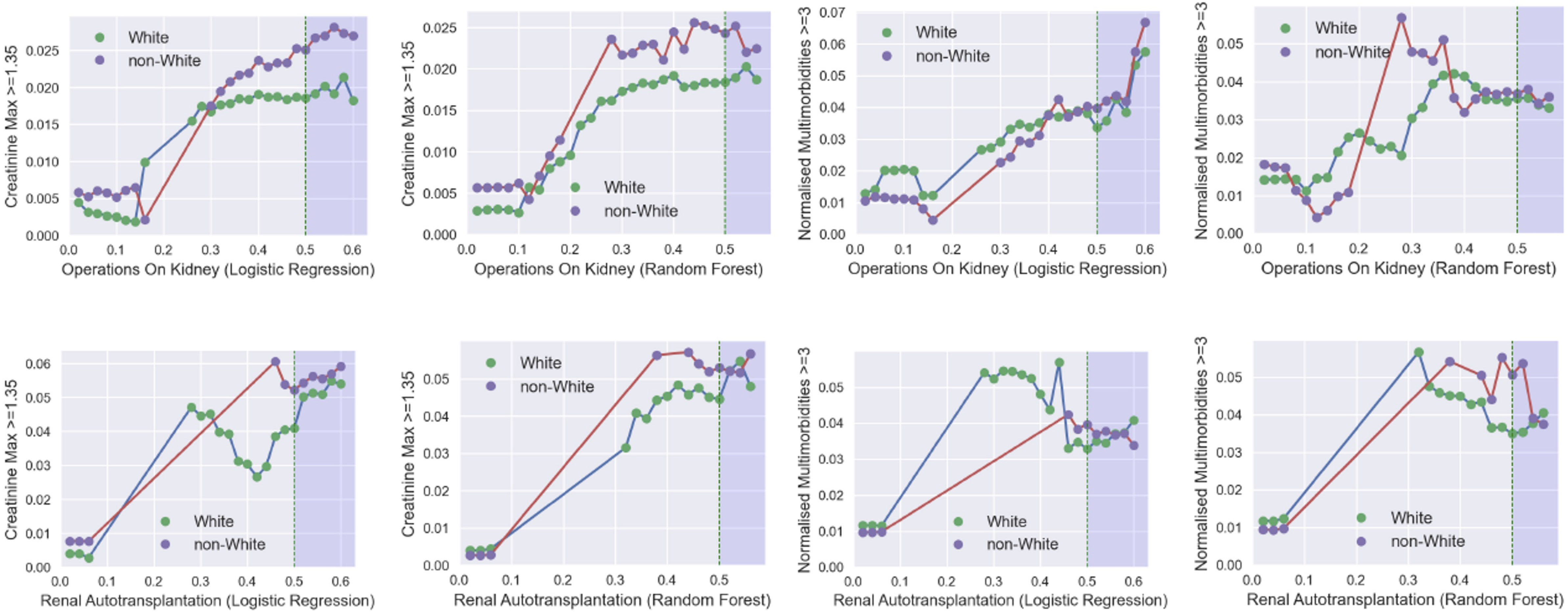
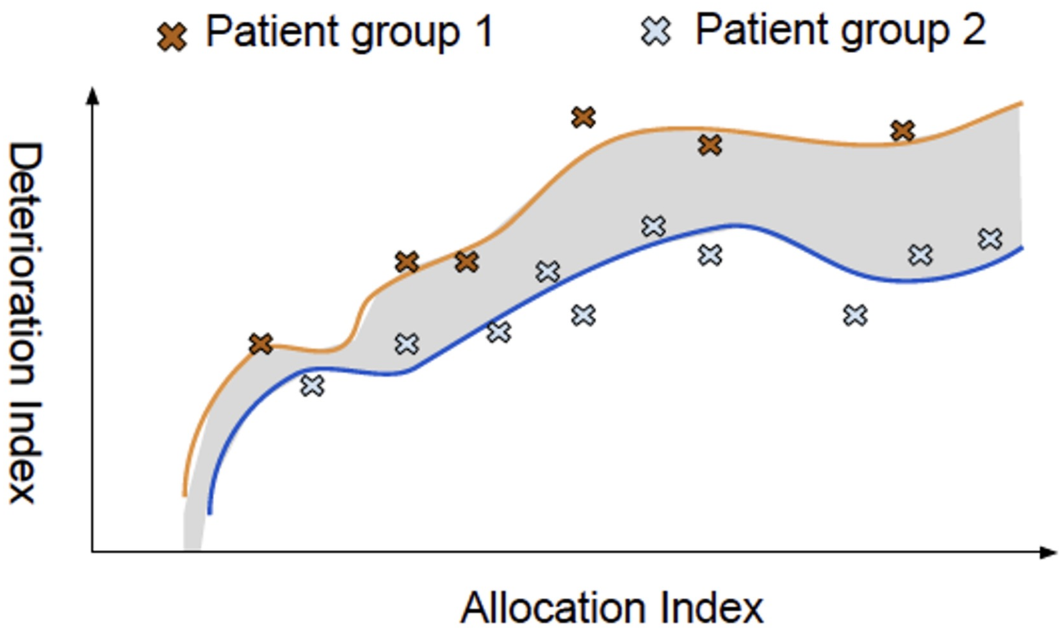


Figure 6: Allocation-Deterioration Indices of four models trained for predicting the needs of kidney related surgeries. The top row is for a generic *Operations on Kidney* and the bottom is for a particular *Renal Autotransplantation*. The left two columns are those using *deterioration index* defined on renal functions, while the right two are those using multimorbidities. In all cases, non-White patients are consistently more severe within the decision region (shaded area, allocation index > 0.5).

Summary



- we proposed a novel allocation-deterioration index framework for quantifying health inequalities
- it quantifies for both data embedded and AI induced inequalities
- experiments showed
    - it works (quantify zero or controlled inequalities correctly)
    - health inequalities exist in both ICU datasets: female vs male; non-white vs white
    - AI models induced inequalities, in most cases making them worse