Quantifying Health Inequalities Induced by Data and AI Models

Honghan Wu[1*] , **Aneeta Sylolypavan**[1] , **Minhong Wang**[1] , **Sarah Wild**[2]
[1]Institute of Health Informatics, University College London, London, United Kingdom
[2]Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
{honghan.wu, aneeta.sylolypavan, minhong.wang}@ucl.ac.uk, sarah.wild@ed.ac.uk

Abstract

AI technologies are being increasingly tested and applied in critical environments including healthcare. Without an effective way to detect and mitigate AI induced inequalities, AI might do more harm than good, potentially leading to the widening of underlying inequalities. This paper proposes a generic allocation-deterioration framework for detecting and quantifying AI induced inequality. Specifically, AI induced inequalities are quantified as the area between two allocation-deterioration curves. To assess the framework's performance, experiments were conducted on ten synthetic datasets (N>33,000) generated from HiRID - a real-world Intensive Care Unit (ICU) dataset, showing its ability to accurately detect and quantify inequality proportionally to controlled inequalities. Extensive analyses were carried out to quantify health inequalities (a) embedded in two real-world ICU datasets; (b) induced by AI models trained for two resource allocation scenarios. Results showed that compared to men, women had up to 33% poorer deterioration in markers of prognosis when admitted to HiRID ICUs. All four AI models assessed were shown to induce significant inequalities (2.45% to 43.2%) for non-White compared to White patients. The models exacerbated data embedded inequalities significantly in 3 out of 8 assessments, one of which was >9 times worse.

1 Introduction

Artificial intelligence (AI) technologies in medicine hold great potential to facilitate better decision-making and efficient service delivery in health and social care, hence, widely and highly expected to improve clinical outcomes in the near future [Topol, 2019]. However, a critical and alarming caveat is that AI driven decision making systems, particularly those using data-driven technologies, are subject to, or themselves cause, bias and discrimination that may exacerbate existing health inequity among racial and ethnicity groups [Leslie *et al.*, 2021].

"Bias in, bias out" is the catchphrase used to highlight concerns about the fact that data driven AI models make inferences by finding 'patterns' from the data they analyse. As racial and ethnic disparities have long existed in health and care [Nelson, 2002; van Ryn and Burke, 2000], inferences from such biased data would inevitably channel embedded inequality into decisions or suggestions they derive. So, effective mitigation is required [Bailey *et al.*, 2017].

Training data might induce bias even when there is no embedded inequalities from service deliveries. Underrepresentation of minority groups in datasets creates a real technical challenge for data driven approaches to draw sensible conclusions for such groups, creating another probable cause of inequality exacerbated by AI. Small samples of a minority group will cause computational models to draw inaccurate predictions for them [Rajkomar *et al.*, 2018].

In addition to biases rooted in the data, further bias could arise in the whole pipeline of AI development and deployment. In particular, health inequalities might be introduced through model selection, the feature engineering process (the choice of the input variables) and label determinations (the choice of target variables) [Passi and Barocas, 2019].

To mitigate these inequalities, conceptual frameworks have been proposed [Rajkomar *et al.*, 2018], and qualitative analysis and checklist based guidance have been suggested [Vyas *et al.*, 2020]. While these tools are useful for understanding the possible types of biases and where they might arise, these solutions are not able to *quantify health inequalities* so that AI practitioners can debug, evaluate and audit potential biases in data, model developments and deployments.

All in all, we currently lack effective technical tools for AI practitioners to deal with the critical issue of detecting and mitigating health inequalities embedded in the training data and/or induced by the AI models' developments and deployments. In this paper,

- we propose a generic quantification framework called *allocation-deterioration indices* that can quantify health inequalities from both datasets and AI models. It is available at https://github.com/knowlab/DAindex-Framework.
- we propose novel and pragmatic solutions to address several technical issues associated with the deterioration index computation including boundary bias of kernel

[*]Contact Author

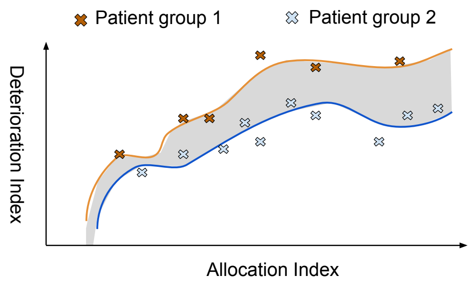

Figure 1: Area between Allocation-Deterioration curves: a generic inequality quantification metric for AI models and derivation data

density estimation and challenges associated with discrete random variables.

- ten synthetic datasets (N>33,000) and two real-world health datasets (total N>70,000) are used and three clinical decision-making scenarios are tested in an extensive set of experiments.

2 Method

Inspired by [Obermeyer *et al.*, 2019] and generalising from it, we define and quantify health inequalities in a generic resource allocation scenario using a so-called allocation-deterioration framework, as conceptualised in Figure 1.

The basic idea is to define two indices: allocation index and deterioration index. The allocation index is (to be derived) from the AI model of interest. Conceptually, AI models are abstracted as "resource allocators", such as predicting the probability of Intensive Care Unit admission. Note that the models themselves do not need to be particularly designed to allocate resources, for example, it could be risk prediction of cardiovascular disease (CVD) among people with diabetes [Dinh *et al.*, 2019]. Essentially, a resource allocator is a computational model that takes patient data as input and outputs a (normalised) score between 0 and 1. We call this score the allocation index. Deterioration index is a score between 0 and 1 to measure the deterioration status of patients. It can be derived from an objective measurement for disease prognosis (i.e., **a marker of prognosis** in epidemiology terminology), such as extensively used comorbidity scores [de Groot *et al.*, 2003; Obermeyer *et al.*, 2019] or biomarker measurements like those for CVDs [Vasan, 2006].

When we have the two indices, each patient can then be represented as a point in a two-dimensional space of (*allocation index*, *deterioration index*), as illustrated in Figure 1. A group of patients are then translated into a set of points in the space, for which a regression model could be fitted to approximate as a curve in the space. The same could be done for another group. *The area between the two curves is then the deterioration difference between their corresponding patient groups, quantifying the inequalities induced by the "allocator", i.e., the AI model that produces the allocation index.* The curve with the larger area under it represents the patient group which would be unfairly treated if the allocation index was to be used in allocating resources or services: a patient from this group would be deemed healthier than a patient from another group who is equally ill. The rest of this section gives technical details of realising key components of this conceptual framework.

2.1 Deterioration Index Definition

For a group of patients $P = \{p_1, p_2, ..., p_n\}$, a deterioration index is a function $d\colon \mathbb{P}(P) \xrightarrow{m} [0, 1]$, where the value 1 denotes the most deteriorated. m stands for a numeric measurement function $m\colon P \rightarrow \mathbb{R}$, where $p \in P$. For example, it can be counting the number of multimorbidities of a patient or the heart beat rates. Generally, let $\{m(p)|p \in P\}$ be $M = \{M_1, M_2, ..., M_n\}$, an independent and identically distributed sample.

The deterioration status is usually quantified as the degree to which the measured value is in excess of what is normal. For example, the normal range of Creatinine (a measure for kidney functions) for adult men is 0.74 to 1.35 mg/dL [1]. A reading of 5 mg/dL is apparently off-the-scale, however, it is probably quantified as less deteriorated than a reading of 10 mg/dL. Following this idea, without loss of generality, we quantify the deterioration index as

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

where f quantifies the degree to which these measures are in excess of a threshold t_m.

A simple implementation, as defined in Definition 2.1, is to use the probability of M having a value greater than a given cut-off t_m. This is intuitive as it quantifies the likelihood of having abnormal measurements within a patient group. For example, when using Creatinine as the measurement, t_m can be set as 1.35, the upper bound of normal readings for men. A group of patients with $f_{Pr} = 0.6$ is more deteriorated than another group with $f_{Pr} = 0.3$ in terms of their kidney functions.

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.

However, f_{Pr} is not able to discriminate a group with a distribution more skewed to the far end of the spectrum from another with a distribution closer to the cut-off when both having the same f_{Pr}. For example, for two groups with Creatinine measures as $\{0.8, 0.78, 10\}$ and $\{0.8, 0.78, 1.36\}$ respectively, both will have the same $f_{Pr}(M; 1.35) = 0.3$, while the former is clearly more deteriorated as it has a much higher abnormal reading.

To address this issue, we propose a 'probability beyond k-step cut-offs' as defined in Definition 2.2. It splits the relevant value range into k steps and allows the quantification to put more weights on more deteriorated values using a weight function.

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

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

[1] https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646

where $\delta = \lceil \frac{max_m - t_m}{k} \rceil$, max_m is the maximum possible value of m and $w(i) \to \mathbb{R}$ 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_{Pr}^2 values of 0.21 and 0.09, respectively.

2.2 Deterioration Index Implementation

To obtain the probabilities in the above definitions, we adopt a kernel density estimation approach, which is a standard non-parametric method for estimating a probability density. Let g_M be the probability density function (PDF) of a measurement random sample M. A kernel density estimator (KDE) is $\hat{g}_M(v; h) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{v - M_i}{h})$, where K is the kernel function and h is the bandwidth parameter used for smoothing the estimate. We use a Gaussian kernel $K(v) = exp(-v^2/2)/\sqrt{2\pi}$ in our implementation, as Gaussianity has been assessed over diverse settings and has shown good performances in our experiments.

However, there is a well-known boundary bias issue [Geenens, 2014] with the kernel density estimation. That is when the random variable is bounded to a closed interval, the KDE will exhibit significant bias at at the end-points of the interval because $\hat{g}_M(v; h)$ will have nonzero probability mass outside the interval. Unfortunately, in the clinical domain, almost all measurements are bounded to a closed interval. The figure at the left of Figure 2 illustrates one example of such a case. It is the PDF plot for the *Creatinine Max* distribution of a non-White adult patient cohort from the MIMIC-III dataset [Johnson *et al.*, 2016]. The PDF is estimated using a Gaussian KDE with a bandwidth of 1.0, which was chosen from a hyper-parameter tuning with grid search. The range of the values is $[0, 50]$. Clearly, there is a nonzero region to the left of the minimal possible value 0 (the red dashed vertical line). Using such a PDF to calculate $Pr(M \geq 0)$ would certainly lead to an inaccurate probability.

In addition to boundary issues at intervals, a much-less studied issue but fairly prevalent in health domain is the pulse-like PDFs which are often estimated for discrete random variables, such as number of multimorbidities of patients. Technically, they are associated with small bandwidth values learned for KDE models. The right figure in Figure 2 is such an example, which is a PDF estimated for numbers of multimorbidities for a White patient group from MIMIC-III. The grid-searched optimal bandwidth for this random sample was 0.0526. It presents a pulse-like PDF with peaks around possible discrete values. To calculate above defined deterioration indices (e.g., f_{Pr}^k) using $t_m = 3$, one would need to get the probability $Pr(M \geq 3)$. Using the PDF as it is would lead to an inaccurate result because the nonzero probability mass right to the left of the cutoff (the red dotted line in the figure) is relevant but would be ignored.

While the boundary issue associated with closed intervals has been studied in the literature for more than several decades, existing approaches are either not very generalisable or difficult to implement [Colbrook *et al.*, 2020]. In fact, very few have been implemented in R or Python libraries. Moreover, few of them tackle the issue associated with pulse-like PDFs as described above. To address these issues, we propose a pragmatic, automated boundary adjustment approach. Algorithm 1 in the appendix (https://knowlab.github.io/preprints/DAapdx.pdf) describes the adjustment for the left boundaries including pulse-like PDFs. This is needed for accurately estimating $Pr(M \geq t_m)$. The similar logic could be applied for the right boundary adjustment for $Pr(M \leq t_m)$. The blue dashed lines in Figure 2 denote the adjusted values for given thresholds.

2.3 Area Under Allocation-Deterioration Curve

The first step to get the area is to generate the allocation-deterioration curve (A-D curve for short). To do that, we start with a *resource allocator*, which in this context, as you will recall, is essentially an AI model used for decision making. Technically, a *resource allocator* is $a(p) \in [0, 1]$ assigning a score for quantifying the degree of a patient p needs some service/resource.

Definition 2.3 (Allocation-Deterioration Curve)**.** Given a measurement m, an allocator a and a deterioration index d, the allocation-deterioration curve is defined as

$$\forall x \in [0, 1], \big(x, d(\{p | p \in P, a(X(p)) = x\}; m)\big).$$

In reality, for a particular dataset, the set of patients having one particular allocation score x might be empty or too few to obtain a reliable estimation of their deterioration status. To address this, we propose an approximation method as described in Algorithm 2 in the appendix.

For those missed points in our approximation, interpolation techniques [Muschelli, 2020] could be applied to fill the blank. However, the missing data does not affect inequality quantification as it is calculated as the relative difference between two groups of patients in the same dataset. Keeping the missingness reflects the actual characteristics of the cohorts, leading to an accurate result.

The area under the A-D curve can be estimated using numerical integration, using simple geometric shapes to approximate the area under the curve. We choose Simpson's rule [2].

2.4 Inequality Quantification

Finally, we define the inequality between two groups of patients. Definition 2.4 defines the inequality embedded in a 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 $\frac{d(P_1; m)}{d(P_2; m)} - 1$.

For AI induced inequality, let $AUC(a, P, d, m)$ be the area under the A-D curve of model a for P using $d(P; m)$ as deterioration index. Let $AUC(a, P, d, m; \tau)$ be the area of the sub-region where the allocation index $\geq \tau$. The AI induced inequality can then be defined in Definition 2.5.

[2] https://personal.math.ubc.ca/~pwalls/math-python/integration/simpsons-rule/

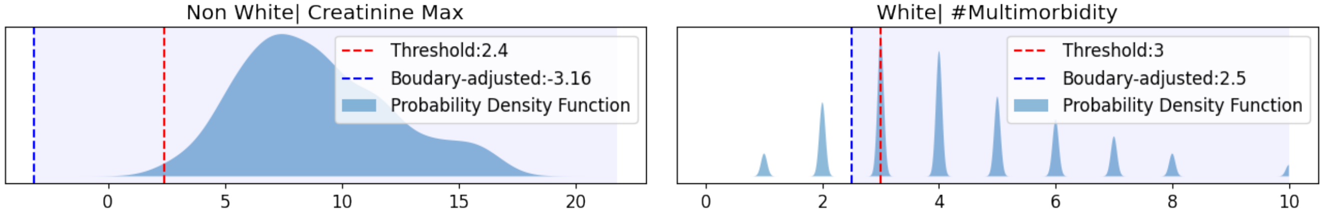

Figure 2: Kernel Density Estimation Issues. The left figure illustrates the boundary bias. It plots a PDF estimated for maximum Creatinine readings (ranged from 0 to 50) of a patient cohort from the MIMIC-III dataset. There is a nonzero probability mass region to the left of legitimate minimal value of zero. The right figure illustrates the issue with pulse-like PDFs for discrete random variables. The PDF was estimated from the number of multimorbitidies of a white patient cohort. Clearly, each discrete value has two regions with nonzero probability mass to either direction.

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

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

3 Results

3.1 Datasets and Cohorts

Two real-world intensive care unit (ICU) datasets were used for experiments, namely: (1) 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 [Faltys *et al.*, 2021]; (2) 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 [Johnson *et al.*, 2016]. Refer to the appendix for ethic statements.

Two case-control cohorts were extracted from MIMIC-III, each of which was for analysing a resource allocation scenario of deciding the need for surgery: (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.

3.2 Inequality Quantification Evaluation

We conducted a few experiments to check whether and how our inequality model works. Specifically, we wanted to evaluate: (a) when there was no bias or inequality, would our model correctly detect it? (b) could our model accurately quantify the known percentages of inequalities? To mimic a near real-world situation, we used the HiRID dataset to generate synthetic data. The generation process was composed of (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.

We evaluated the inequality associated with ICU admissions. Specifically, we used Definition 2.4 to assess inequality of **female vs male** at a *resource allocation scenario* of **ICU admission**. Three measurements (prognosis markers) were chosen for quantifying deterioration indices: *Creatinine max value*, *Creatinine min value* and *ALT min value*. We selected 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 and 52.2 to 91.9 micromoles/L for men. *ALT* measures liver functions and normal ranges chosen were: $\leq$ 30 U/L for men and $\leq$ 19 U/L for women [Kunde *et al.*, 2005]. The deterioration index used a *probability on 20-step cut-offs*.

For answering the above question (a), i.e., detecting no inequality, we generated 10 synthetic datasets using the above-mentioned process and ran inequality assessments on these datasets. Note that the synthetic data were actual data of male patients. Therefore, with sufficient numbers of sampling, there should NOT be any significant amount of inequality between male and (synthetically created) female patients overall. Table 1 shows the overall results of 10 runs on 10 such datasets. The p-value was generated for a T-test for the null hypothesis that the mean value was equal to 0, meaning NO inequality. We observed all p-values were not significant, meaning we could not reject the null hypothesis. This means in all three measurements, the mean values of 10 runs were equal to 0, indicating our model quantified no significant inequalities.

For the above question (b), i.e., whether our model could quantify inequality proportionally to actual inequality, we

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

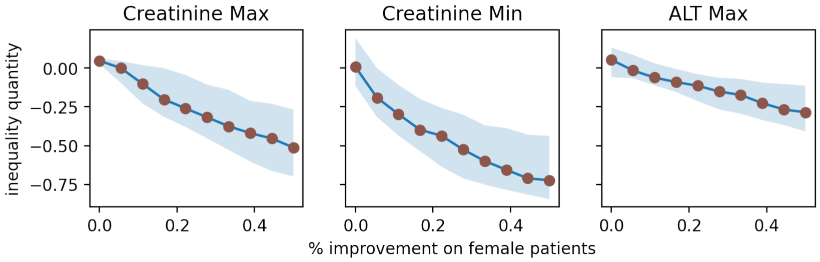

Figure 3: Inequality Quantification Evaluation on synthetic data: y-axis is the inequality quantity of female vs male. x-axis is the percentage of controlled improvements on readings of the female subcohort. Y-value of each point is the mean value of 10 runs on the same x-value, i.e., % of improvement. Shaded areas denote 25-75% quantile regions.

used the same process as the previous experiment of generating 10 datasets. Then, for female patients, we purposely improved their measurements by changing readings towards the more healthier end, e.g., decrease the Creatinine max readings, increase Creatinine min readings. We selected different levels of improvements - 10 steps evenly spaced between 0.0 and 0.5. For each of them, we quantified the inequality of female vs male. Figure 3 depicts the results of this experiment. In all cases, the model correctly identified the level of inequality changes - inequality trends going downwards consistently when the strength of improvement increases. Specifically, 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, showing near perfect negative correlations.

3.3 Dataset Embedded Inequality Analysis

The first analysis was conducted on a HiRID dataset for detecting and quantifying its embedded inequality, using the same inequality quantification setup as the previous subsection.

Table 2 shows the results on 10 randomly selected cohorts from the total HiRID admissions. Each cohort had 3,390 patients. Among the three deterioration indices, females were shown slightly healthier on *Creatinine max*, but overall marginal with a p-value of 0.02. On the other two indices (both with a much higher statistical significance), females were clearly worse off in those two measurements. In particular, they were significantly more ill than male patients on *Creatinine min*, quantified as 0.337 (intepretable as 33.7% more deteriorated than males at admission). Overall, females admitted to ICU were more deteriorated compared to males within HiRID.

The second analysis was conducted on the MIMIC-III dataset to evaluate the health inequality of **non-White patients vs White patients** at a resource allocation scenario of **Operations on Kidney** - a cohort with 2,336 patients as described above. We were interested in finding the inequality among the 584 patients who underwent kidney operations. Here, we report the *Creatinine Max* based health inequality. Male and female have different normal ranges (MIMIC III uses mg/dL as the unit): 0.74 to 1.35 for men and 0.59 to 1.04 for women. Therefore, we compare male and female separately. Figure 4 depicts the PDF distributions of four sub-cohorts. This experiment also used a deterioration index based on the probability beyond 20-step cut-offs. For those who underwent kidney operations in MIMIC-III, female none-White had a 35.06% inequality over their White female peers, while none-White males were also worse off compared to White males, quantified as 19.94%. This indicates non-White patients were consistently and substantially more deteriorated in terms of their kidney functions in such a resource allocation scenario. In particular, among all the four subgroups, **the inequality of non-White male vs White female** was the most significant: 46.57%.

3.4 Model Induced Inequality Analysis

To assess the impact of AI models on health inequality if they were used for clinical decision making, two case studies were conducted in two resource allocation scenarios: one on **Renal Autotransplantation** and the other on more general **Operations on Kidney**. Tables describing these patient characteristics are available in the appendix (Table 6 and 7)

Logistic Regression (LR) and Random Forest (RF) models were developed for predicting the need for surgeries in both cases with 10-fold cross validation and grid search for hyper-parameter tuning. The two algorithms were chosen because they were widely used in clinical studies. Details of feature selection and hyper-parameters are available in the appendix (Table 5). For prediction performances (ROCAUC), LR achieved 0.795 (IQR:0.784-0.805) and 0.867 (IQR:0.843-

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 2: Inequality analysis of Female vs Male on ten random subcohorts 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.

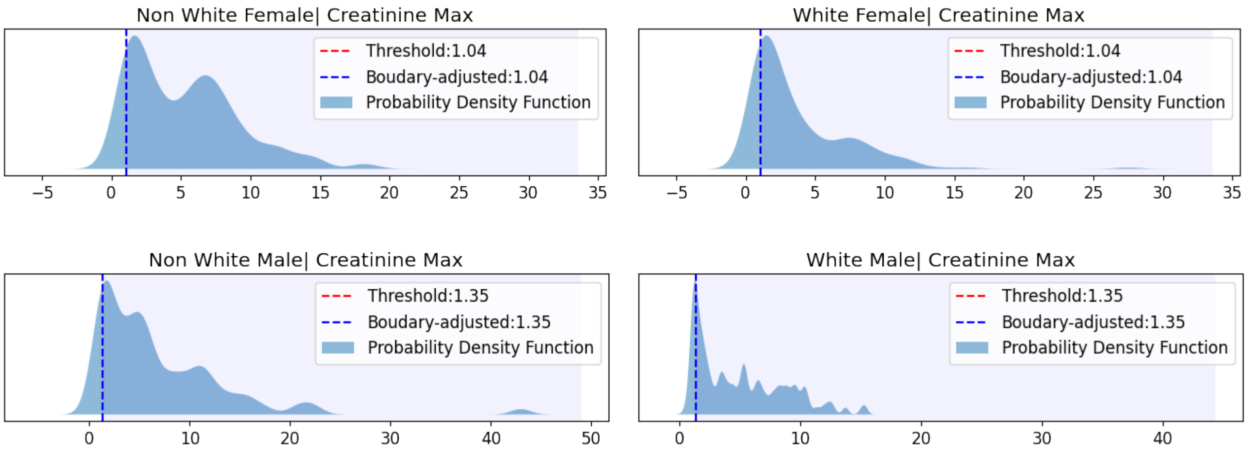

Figure 4: Probability density functions for quantifying inequalities of non-White vs White in the scenario of kidney operations in MIMIC-III dataset. Dashed lines denote thresholds (i.e., boundary values of abnormal readings) for computing deterioration index. Shaded area are regions where the probability integral happens for getting the deterioration index. The above two figures are females, which illustrate an inequality of 35.06% for non-white patients. The bottom two are males, where there is an inequality of 19.94%.

Kidney operation				
Measurement	Creatinine Max		Normalised MM	
DB Inequality	29.10%		7.62%	
Models	LR	RF	LR	RF
DR Inequality	**37.6%**	22.2%	**10.52%**	4.54%

Table 3: Inequality of non-White vs White induced by AI models for predicting kidney operations. DB inequality row gives the DB embedded inequality quantities. DR Inequality is quantified by areas under A-D curves in the region where a model suggests surgery.

0.891) for *Operations on Kidney* and *Renal Autotransplantation*, respectively, while RF achieved 0.830 (0.816-0.844) and 0.878 (0.853-0.904), respectively.

For quantifying the inequality, two deterioration indices were used including *Creatinine Max* and a new measurement of *Noramlised number of multimorbidities*, *Normalised MM* for short. The multimorbidities included those in a list of 17 chronic conditions as defined by [St Sauver *et al.*, 2021]. *Normalised MM* is defined as $\#MM \times \frac{65}{age}$, where $\#MM$ is the number of multimorbidities a patient had.

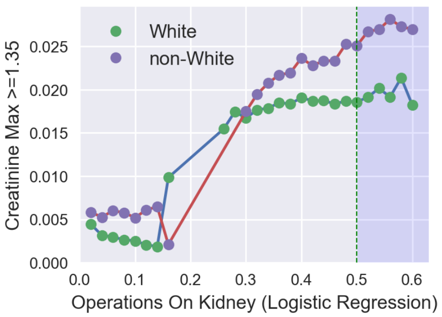

Figure 5: Allocation-Deterioration Curves (non-White vs White) of a model trained for predicting the need for kidney related surgeries. Non-White patients are significantly (37.6%) more severe within the decision region (shaded area, allocation index > 0.5).

Using inequalities quantified by Definition 2.5, Table 3 summarises the inequality of **non-White vs White** induced by AI models for *Kidney Operation*. Table 4 in the appendix shows full details of the two surgeries. Compared to inequality embedded in the database (of those who actually underwent surgeries), LR models exacerbated the inequality in 3 out of 4 assessments. RF tends to perform better in terms of mitigating the inequality in 3 out of 4 assessments, albeit very marginally in most cases. However, RF significantly exacerbated inequality more than 9 times on one occasion (see the last column of Appendix Table 4). Overall, AI models induced inequalities in all cases and exacerbated inequalities severely in 3 out of 8 assessments. Figure 5 illustrates one selected exemplar visualisation out of the eight total assessments (Appendix Figure 6 shows all eight). All these curves demonstrate a clear pattern that non-White patients are more deteriorated at the decision regions in all situations.

4 Conclusion

This paper proposes an Allocation-Deterioration framework that, to the best of our knowledge, is the first utility that visualises and quantifies health inequality induced by AI models and embedded in health datasets. Such a utility enables the evaluation, debugging and mitigation of inequality caused by AI technologies. While we focused on motivations and real-world data in the health domain, this framework is clearly generalisable and has much wider applications. An extensive set of experiments were conducted on two large, real-world datasets to assess its performances and reveal the existing (hidden) inequalities in different decision-making scenarios.

References

[Bailey *et al.*, 2017] Zinzi D Bailey, Nancy Krieger, Madina Agénor, Jasmine Graves, Natalia Linos, and Mary T Bassett. Structural racism and health inequities in the USA: evidence and interventions. *Lancet*, 389(10077):1453–1463, April 2017.

[Colbrook *et al.*, 2020] Matthew J Colbrook, Zdravko I Botev, Karsten Kuritz, and Shev MacNamara. Kernel density estimation with linked boundary conditions. *Studies in Applied Mathematics*, 145(3):357–396, 2020.

[de Groot *et al.*, 2003] Vincent de Groot, Heleen Beckerman, Gustaaf J Lankhorst, and Lex M Bouter. How to measure comorbidity. a critical review of available methods. *J. Clin. Epidemiol.*, 56(3):221–229, March 2003.

[Dinh *et al.*, 2019] An Dinh, Stacey Miertschin, Amber Young, and Somya D Mohanty. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med. Inform. Decis. Mak.*, 19(1):211, November 2019.

[Faltys *et al.*, 2021] Martin Faltys, Marc Zimmermann, Xinrui Lyu, Matthias Hüser, Stephanie Hyland, Gunnar Rätsch, and Tobias Merz. Hirid, a high time-resolution icu dataset (version 1.1.1). https://physionet.org/content/hirid/1.1.1/, 2021. Accessed: 2022-06-07.

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

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, 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.

[Kunde *et al.*, 2005] Sachin S Kunde, Audrey J Lazenby, Ronald H Clements, and Gary A Abrams. Spectrum of nafld and diagnostic implications of the proposed new normal range for serum alt in obese women. *Hepatology*, 42(3):650–656, 2005.

[Leslie *et al.*, 2021] David Leslie, Anjali Mazumder, Aidan Peppin, Maria K Wolters, and Alexa Hagerty. Does “ai” stand for augmenting inequality in the era of covid-19 healthcare? *bmj*, 372, 2021.

[Muschelli, 2020] John Muschelli. Roc and auc with a binary predictor: a potentially misleading metric. *Journal of classification*, 37(3):696–708, 2020.

[Nelson, 2002] Alan Nelson. Unequal treatment: confronting racial and ethnic disparities in health care. *J. Natl. Med. Assoc.*, 94(8):666–668, August 2002.

[Obermeyer *et al.*, 2019] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019.

[Passi and Barocas, 2019] Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* ’19, pages 39–48, New York, NY, USA, January 2019. Association for Computing Machinery.

[Rajkomar *et al.*, 2018] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.*, 169(12):866–872, December 2018.

[St Sauver *et al.*, 2021] Jennifer L St Sauver, Alanna M Chamberlain, William V Bobo, Cynthia M Boyd, Lila J Finney Rutten, Debra J Jacobson, Michaela E McGree, Brandon R Grossardt, and Walter A Rocca. Implementing the us department of health and human services definition of multimorbidity: a comparison between billing codes and medical record review in a population-based sample of persons 40–84 years old. *BMJ open*, 11(4):e042870, 2021.

[Topol, 2019] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.*, 25(1):44–56, January 2019.

[van Ryn and Burke, 2000] Michelle van Ryn and Jane Burke. The effect of patient race and socio-economic status on physicians’ perceptions of patients. *Soc. Sci. Med.*, 50(6):813–828, March 2000.

[Vasan, 2006] Ramachandran S Vasan. Biomarkers of cardiovascular disease: molecular basis and practical considerations. *Circulation*, 113(19):2335–2362, May 2006.

[Vyas *et al.*, 2020] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine*, 383(9):874–882, 2020.