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

### Background - disparities in healthcare

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Population</th>
<th>All COVID-19 Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>n (%)</td>
<td>56608985</td>
<td>3469528 (6.1)</td>
</tr>
<tr>
<td>COVID Deaths (%)</td>
<td>140908 (0.2)</td>
<td>140908 (4.1)</td>
</tr>
<tr>
<td>All Deaths (%)</td>
<td>723925 (1.3)</td>
<td>178721 (5.2)</td>
</tr>
<tr>
<td>Male (%)</td>
<td>28031640 (49.5)</td>
<td>1571566 (45.3)</td>
</tr>
</tbody>
</table>

#### Ethnic group (%)

<table>
<thead>
<tr>
<th>Ethnic Group</th>
<th>Population</th>
<th>All COVID-19 Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>45300233 (80.0)</td>
<td>2679541 (77.2)</td>
</tr>
<tr>
<td>Asian or Asian British</td>
<td>4892279 (8.6)</td>
<td>448329 (12.9)</td>
</tr>
<tr>
<td>Black or Black British</td>
<td>2155780 (3.8)</td>
<td>139171 (4.0)</td>
</tr>
<tr>
<td>Chinese</td>
<td>516056 (0.9)</td>
<td>10412 (0.3)</td>
</tr>
<tr>
<td>Mixed</td>
<td>1198711 (2.1)</td>
<td>68536 (2.0)</td>
</tr>
<tr>
<td>Other</td>
<td>1249555 (2.2)</td>
<td>73041 (2.1)</td>
</tr>
<tr>
<td>Unknown</td>
<td>1296371 (2.3)</td>
<td>50498 (1.5)</td>
</tr>
</tbody>
</table>

#### Social deprivation fifths (%)

<table>
<thead>
<tr>
<th>Social Deprivation</th>
<th>Population</th>
<th>All COVID-19 Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (most deprived)</td>
<td>11702944 (20.7)</td>
<td>832546 (24.0)</td>
</tr>
<tr>
<td>5 (least deprived)</td>
<td>10816336 (19.1)</td>
<td>562664 (16.2)</td>
</tr>
<tr>
<td>Unknown</td>
<td>53813 (0.1)</td>
<td>2902 (0.1)</td>
</tr>
</tbody>
</table>

Background - inequality because of discrimination

<table>
<thead>
<tr>
<th>Physician perception that patient is...</th>
<th>Race/ethnicity</th>
<th>Percent</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Not at all likely” to abuse alcohol or other drugs (N=582)</td>
<td>White/Black</td>
<td>79/67</td>
<td>11.65, $p \leq 0.001$</td>
</tr>
<tr>
<td>“Not at all likely” to lack social support$^a$ (N=576)</td>
<td>White/Black</td>
<td>63/45</td>
<td>19.61, $p \leq 0.001$</td>
</tr>
<tr>
<td>‘Very’ intelligent (vs. unintelligent)$^b$ (N=438)</td>
<td>White/Black</td>
<td>26/13</td>
<td>16.32, $p \leq 0.0001$</td>
</tr>
<tr>
<td>‘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)</td>
<td>low SES</td>
<td>53/27</td>
<td>8.26, $p \leq 0.01$</td>
</tr>
<tr>
<td></td>
<td>White/low SES</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Background - inequality because of underrepresentation

<table>
<thead>
<tr>
<th></th>
<th>Population</th>
<th>All COVID-19 Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n (%)</strong></td>
<td>56608985</td>
<td>3469528 (6.1)</td>
</tr>
<tr>
<td><strong>COVID Deaths (%)</strong></td>
<td>140908 (.2)</td>
<td>140908 (4.1)</td>
</tr>
<tr>
<td><strong>All Deaths (%)</strong></td>
<td>723925 (1.3)</td>
<td>178721 (5.2)</td>
</tr>
<tr>
<td><strong>Male (%)</strong></td>
<td>28031640 (49.5)</td>
<td>1571566 (45.3)</td>
</tr>
<tr>
<td><strong>Ethnic group (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>45300233 (80.0)</td>
<td>2679541 (77.2)</td>
</tr>
<tr>
<td>Asian or Asian British</td>
<td>4892279 (8.6)</td>
<td>448329 (12.9)</td>
</tr>
<tr>
<td>Black or Black British</td>
<td>2155780 (3.8)</td>
<td>139171 (4.0)</td>
</tr>
<tr>
<td>Chinese</td>
<td>516056 (0.9)</td>
<td>10412 (0.3)</td>
</tr>
<tr>
<td>Mixed</td>
<td>1198711 (2.1)</td>
<td>68536 (2.0)</td>
</tr>
<tr>
<td>Other</td>
<td>1249555 (2.2)</td>
<td>73041 (2.1)</td>
</tr>
<tr>
<td>Unknown</td>
<td>1296371 (2.3)</td>
<td>50498 (1.5)</td>
</tr>
<tr>
<td><strong>Social deprivation fifths (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (most deprived)</td>
<td>11702944 (20.7)</td>
<td>832546 (24.0)</td>
</tr>
<tr>
<td>5 (least deprived)</td>
<td>10816336 (19.1)</td>
<td>562664 (16.2)</td>
</tr>
</tbody>
</table>

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.
“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.”


Motivation of this work

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

Why quantification?

*like* Precision/Recall/F1 *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

with a numeric measurement function

The deterioration index is

\[ P = \{p_1, p_2, \ldots, p_n\} \]

\[ m: P \rightarrow \mathbb{R}. \]

\[ d: \mathbb{P}(P) \xrightarrow{m} [0, 1] \]

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

\[ \{m(p) | p \in P\} \]

\[ d(P; m) = f(\{M_1, M_2, \ldots, M_n\}; t_m) \]

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 $\text{fpr}(M;1.35)$, 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 $f_{Pr}^k$ 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 $\delta = \left\lfloor \frac{\max_m - t_m}{k} \right\rfloor$, $\max_m$ 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 $f_{Pr}^2$ values of 0.21 and 0.09, respectively.
Use kernel density estimation to estimate the probability density function (PDF) of $Pr$

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

$K$ is a Gaussian kernel

$$\exp\left(-\frac{v^2}{2}\right)/\sqrt{2\pi}$$
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.

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*: $\text{arg max}(\{v|v \in M : v < t\})$ when $M$ is discrete and $t$ is not boundary, otherwise $t$; 
$\epsilon$: a small constant like $1^{-10}$; 
$V$: an empty array.

output: $t$: the adjusted value for $t$

if $\text{len}(V) = 0$ then
    /* get an evenly spaced numbers between $lb$ and $ub$ with a relatively big number $n$, e.g., $n = 20 \times (ub - lb)$. */
    a ← gen($lb$, $ub$, $n$);
    s ← ($ub - lb$)/$n$;
    for $i$ ← 1 to $\text{len}(a)$ do
        $x_p$ ← $lb$;
        if $i > 1$ then
            $x_p$ ← $a[i - 1]$;
        end
        $x$ ← $a[i]$;
        $p$ ← $\text{exp}(E(x))$;
        while $p \geq \epsilon$ and $x > x_p$ do
            $x$ ← ($x - s$);
            $p$ ← $\text{exp}(E(x))$;
        end
        if $\text{exp}(E(x)) < \epsilon$ then
            $V$.add($x$);
        end
    end
end

$t$ ← $\text{arg max}(\{v|v \in V : v < t\})$;
if $t$ ≤ $t*$ then
    $t$ ← $t$;
end
return $t$;
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 $P_1 \text{ vs } P_2$) is quantified as $\frac{d(P_1; m)}{d(P_2; m)} - 1$.

Definition 2.5 (Inequality induced by a model). In a decision making scenario with an allocation threshold $\tau$, 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


**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


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**  
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.

**Creatinine min value**

**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**  
\[ \#M \times \frac{65}{age} \]

The deterioration index used a probability on 20-step cut-offs.
Inequality quantification evaluation

do 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?

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.
- ICU admission to HiRID: female vs male

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

<table>
<thead>
<tr>
<th>Measurement</th>
<th>mean [95% CI]</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creatinine max</td>
<td>-0.079 [-0.207, 0.034]</td>
<td>0.0219</td>
</tr>
<tr>
<td>Creatinine min</td>
<td>0.337 [0.181, 0.472]</td>
<td>0.0000</td>
</tr>
<tr>
<td>ALT max</td>
<td>0.093 [0.018, 0.197]</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

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.
**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

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature List</td>
<td>['age', 'Chronic kidney disease', 'gender', 'Leukemia', 'cirrhosis', 'Infection']</td>
</tr>
</tbody>
</table>
| Random Forest       | tuned_parameters = {
| Hyper-parameters    | 'n_estimators': [50, 100, 200],
|                     | 'max_depth': [5, 10, 20, 50]               |
| Logistic Regression | tuned_parameters = {
| Hyper-parameters    | 'penalty': ['11', '12'],
|                     | '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

<table>
<thead>
<tr>
<th>DB inequality</th>
<th>Kidney operation</th>
<th>Renal Autotransplantation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Creatinine Max</td>
<td>Normalised MM</td>
</tr>
<tr>
<td></td>
<td>29.10%</td>
<td>7.62%</td>
</tr>
<tr>
<td>Models</td>
<td>LR</td>
<td>RF</td>
</tr>
<tr>
<td>Inequality at Decision Region</td>
<td>37.58%</td>
<td>22.15%</td>
</tr>
<tr>
<td>Inequality at the whole area</td>
<td>16.17%</td>
<td>30.21%</td>
</tr>
<tr>
<td></td>
<td>9.13%</td>
<td>3.51%</td>
</tr>
<tr>
<td></td>
<td>14.73%</td>
<td>22.70%</td>
</tr>
</tbody>
</table>

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.
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 induce inequalities, in most cases making them worse.