Weill Cornell Medicine

Weill Cornell Medicine
Institute of Artificial Intelligence for Digital Health

KDD 2023 Tutorial – LS09
Mining Electronic Health Records for Real-World Evidence

Tuesday, August 8th,10:00 am-13:00 pm PDT, Room 202A

Chengxi Zang, PhD, Weishen Pan, PhD, & Fei Wang, PhD
Department of Population Health Sciences
Institute of Artificial Intelligence for Digital Health (AIDH)
Weill Cornell Medicine, Cornell University

www.calvinzang.com/ehr4rwe_kdd2023.html

Outline

- Generating Real-World Evidence for Understanding Long COVID
- Advancements in Risk Prediction using EHRs
- Discussion & QA

Weill Cornell Medicine

Weill Cornell Medicine
Institute of Artificial Intelligence for Digital Health

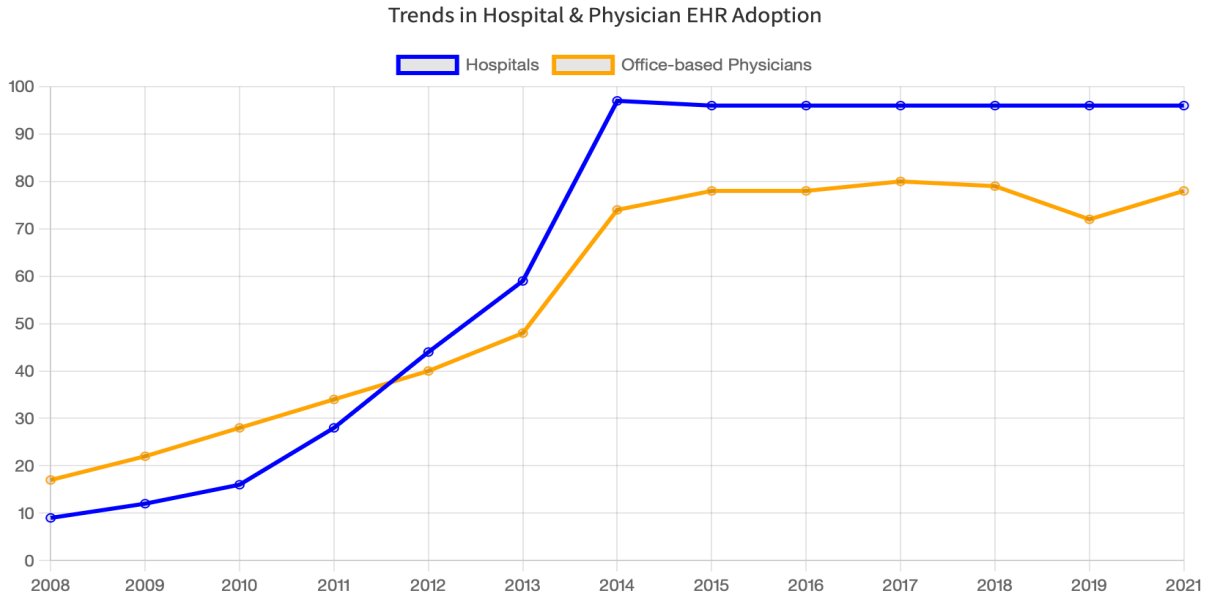
Part-2: Advancements in Risk Prediction for Healthcare

Tuesday, August 8th,10:00 am-13:00 pm PDT, Room 202A

Chengxi Zang, PhD, Weishen Pan, PhD, & Fei Wang, PhD
Postdoctoral Associate @ Department of Population Health Sciences
Institute of Artificial Intelligence for Digital Health (AIDH)
Weill Cornell Medicine, Cornell University
www.calvinzang.com/ehr4rwe_kdd2023.html
wep4001@med.cornell.edu

Electronic Health Record (EHR)

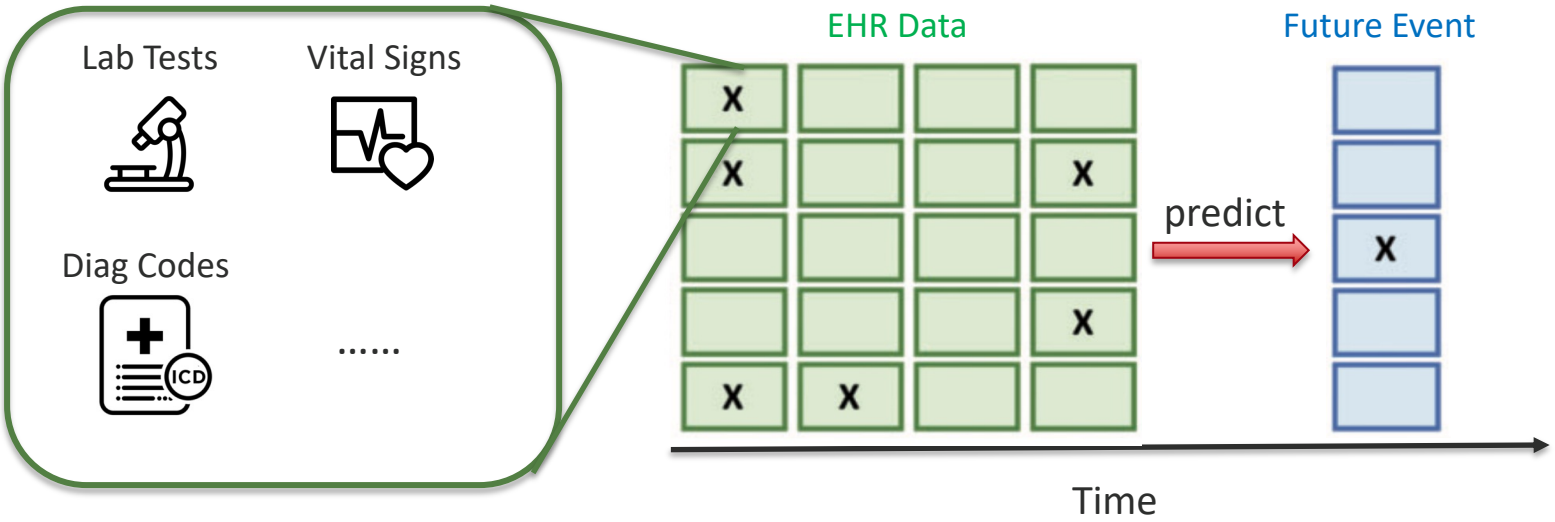
- EHR has been one of the most common data sources to provide patient health information for analysis



(Figure source: https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records)

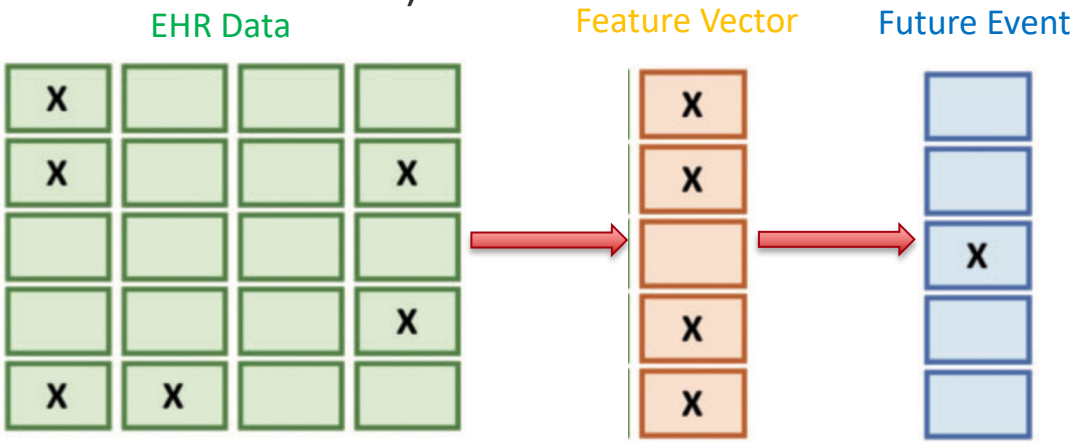
Important Task with EHR : Risk Prediction

- Predict the probability (risk) that an adverse event will happen in the future



Risk Prediction with EHR

- Transform EHR data into a feature vector, input the vector to a predictor for risk prediction
  - Directly summarize the statistics of each event or test within the collection window
  - Organize the EHR data as a sequence and directly input it to an encoder (e.g., recurrent neural networks)



Workflow of Risk Prediction with EHR

1 

Data Gathering and Cleaning

2 

Data Pre-processing

3 

Building a Predictive Model

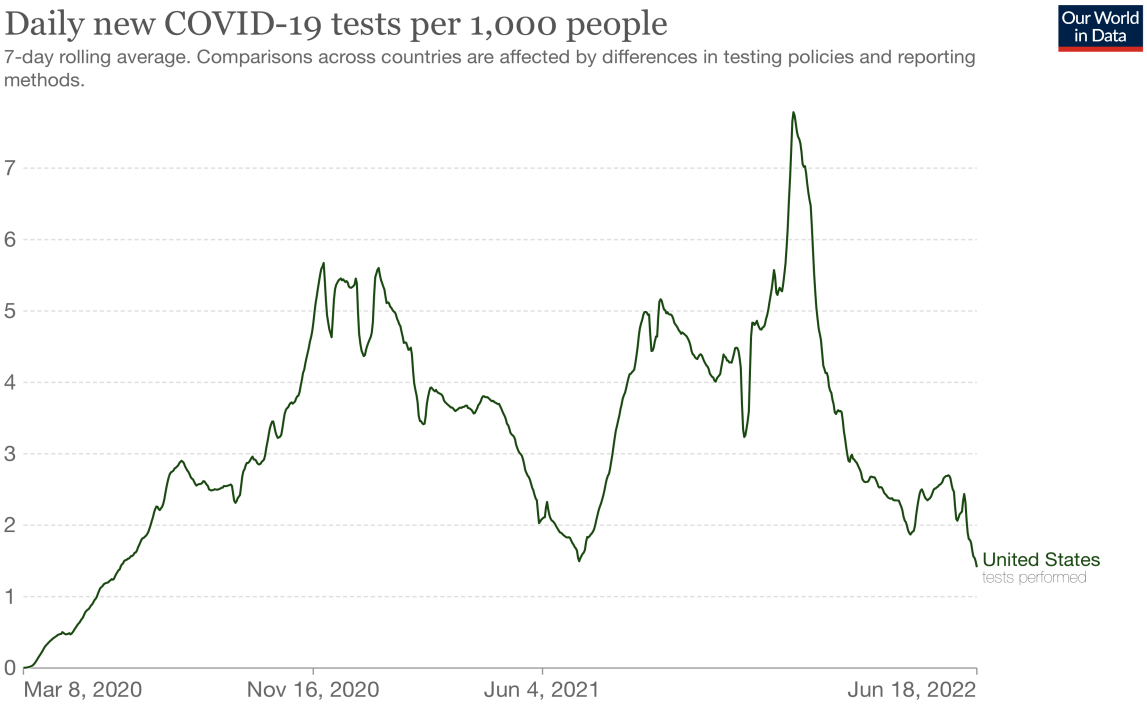
4 

Model Evaluation

5 

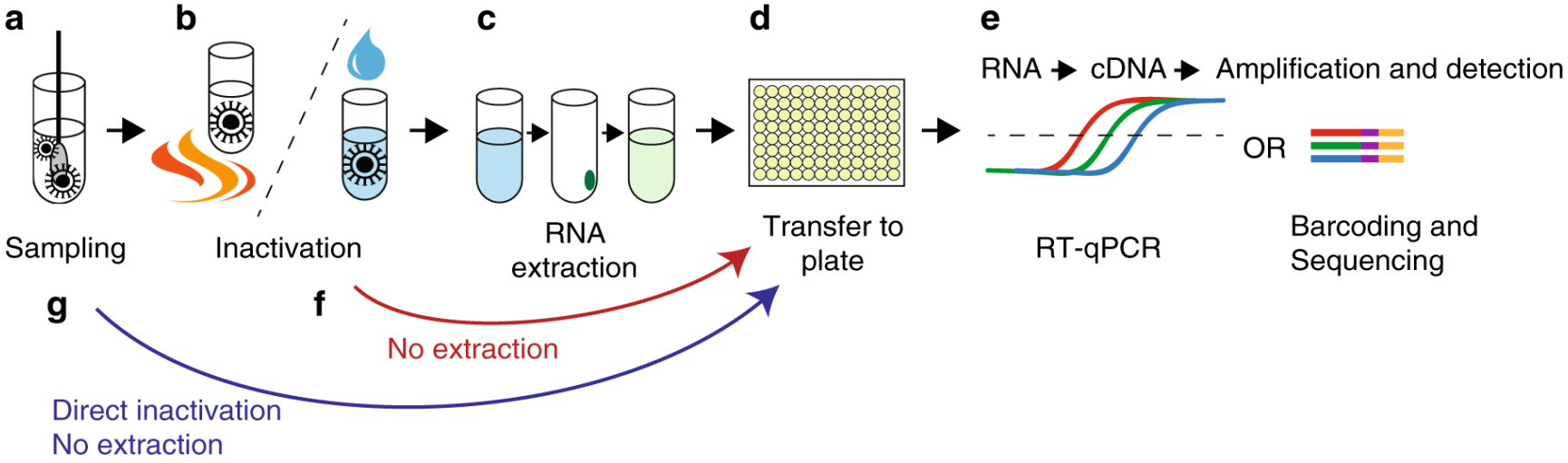
Incorporate the Model in Healthcare Applications

Testing of COVID Infection



RT-PCR Testing of COVID Infection

- Reverse transcription polymerase chain reaction (RT-PCR) test
- The turn-around time (TAT) of RT-PCR testing is usually within 2days, may be longer due to many reasons



Smyrlaki, Ioanna, Martin Ekman, Antonio Lentini, Nuno Rufino de Sousa, Natali Papanicolaou, Martin Vondracek, Johan Aarum et al. "Massive and rapid COVID-19 testing is feasible by extraction-free SARS-CoV-2 RT-PCR." *Nature communications* 11, no. 1 (2020): 4812.

Predict RT-PCR Testing with Routine Lab Tests

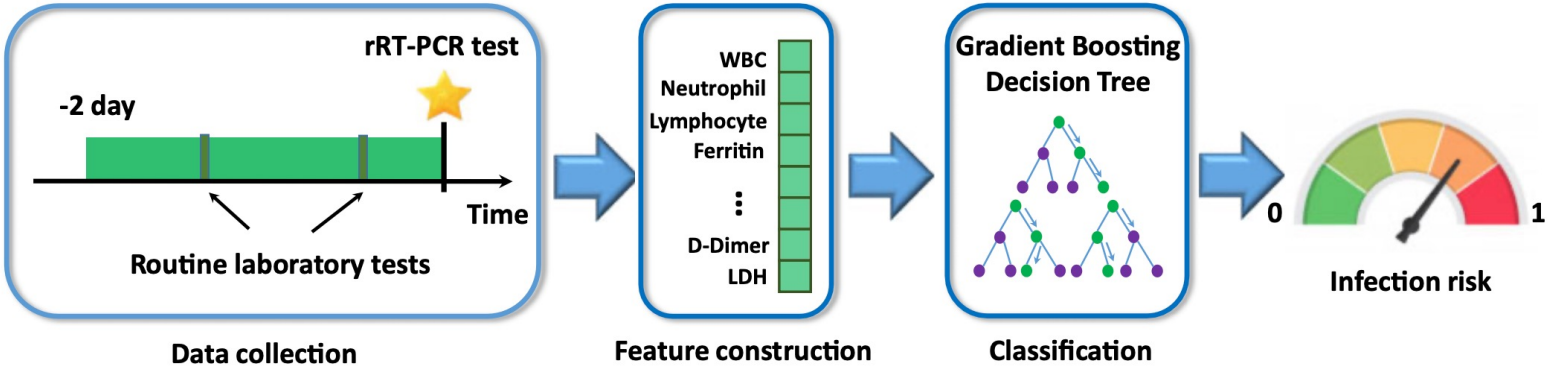
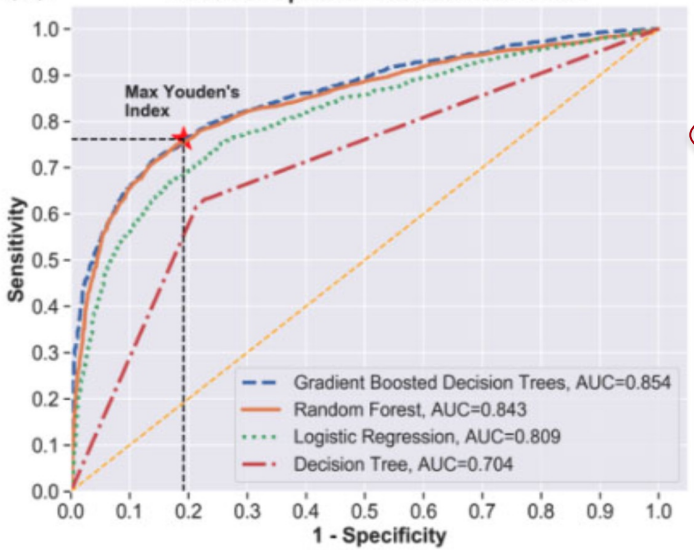


Fig. 2. Illustration of the modeling pipeline. Routine laboratory testing results completed within 2 days prior to the release of RT-PCR results were used to construct a vector, upon which a classifier was built to predict the RT-PCR positive or negative result. Each dimension of the vector corresponds to a specific laboratory test, and its value corresponds to the average of all results of this laboratory test taken during the collection window. The model outputs a probability score ranging from 0–1, indicating the risk of SARS-CoV-2 infection.

Yang, He S., Yu Hou, Ljiljana V. Vasovic, Peter AD Steel, Amy Chadburn, Sabrina E. Racine-Brzostek, Priya Velu et al. "Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning." *Clinical chemistry* 66, no. 11 (2020): 1396-1404.

Prediction Performance

(a)



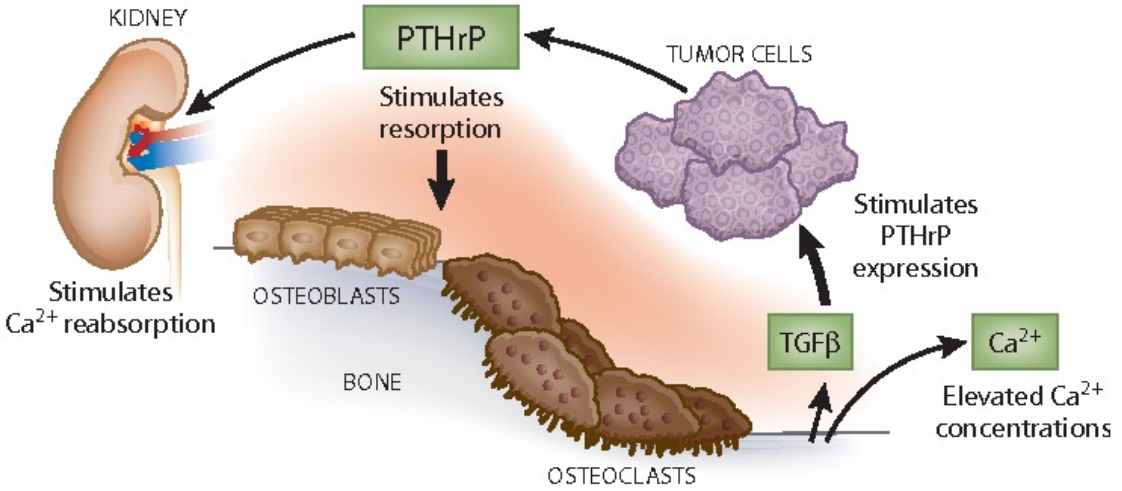
(b)

	AUC	Sensitivity	Specificity	Agreement with RT-PCR
GBDT	0.854 (0.829-0.878)	0.761 (0.744-0.778)	0.808 (0.795-0.821)	0.791 (0.776-0.805)
Random Forest	0.843 (0.823-0.874)	0.735 (0.715-0.754)	0.818 (0.804-0.832)	0.787 (0.771-0.802)
Logistic Regression	0.809 (0.781-0.836)	0.711 (0.689-0.732)	0.756 (0.741-0.771)	0.739 (0.722-0.756)
Decision Tree	0.704 (0.669-0.731)	0.618 (0.599-0.637)	0.732 (0.718-0.746)	0.689 (0.673-0.705)

Fig. 3. Performance of 4 models using 5-fold cross validation on the test set. (a) Comparison of the ROC curves for the gradient boosting decision tree (GBDT) model, random tree model, logistic regression model, and decision tree model. (b) Comparison of the AUC, sensitivity, specificity, and agreement with SARS-CoV-2 RT-PCR (at the operating point determined by the Youden Index) achieved by the 4 models.

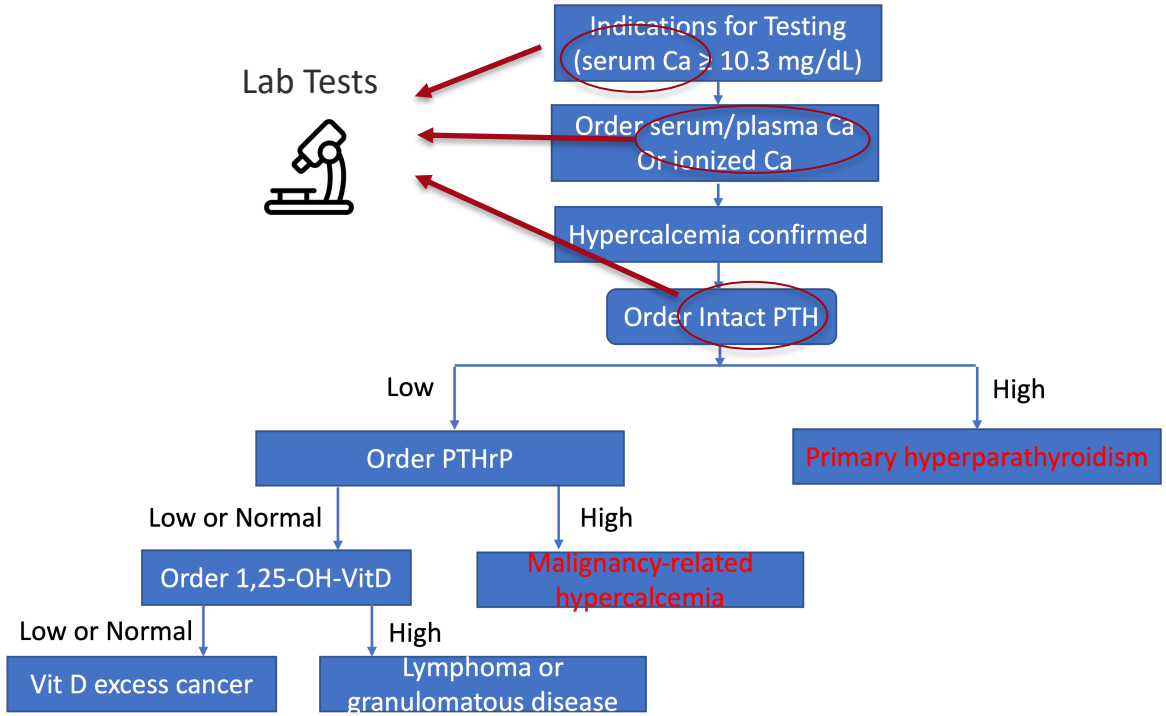
Parathyroid Hormone-related Peptide (PTHrP)

- PTHrP is the most common cause of humoral malignancy-related hypercalcemia
- PTHrP testing can aid in diagnosing hypercalcemia of malignancy



Mundy, Gregory R., and James R. Edwards. "PTH-related peptide (PTHrP) in hypercalcemia." Journal of the American Society of Nephrology 19, no. 4 (2008): 672-675.

Hypercalcemia Testing Algorithm

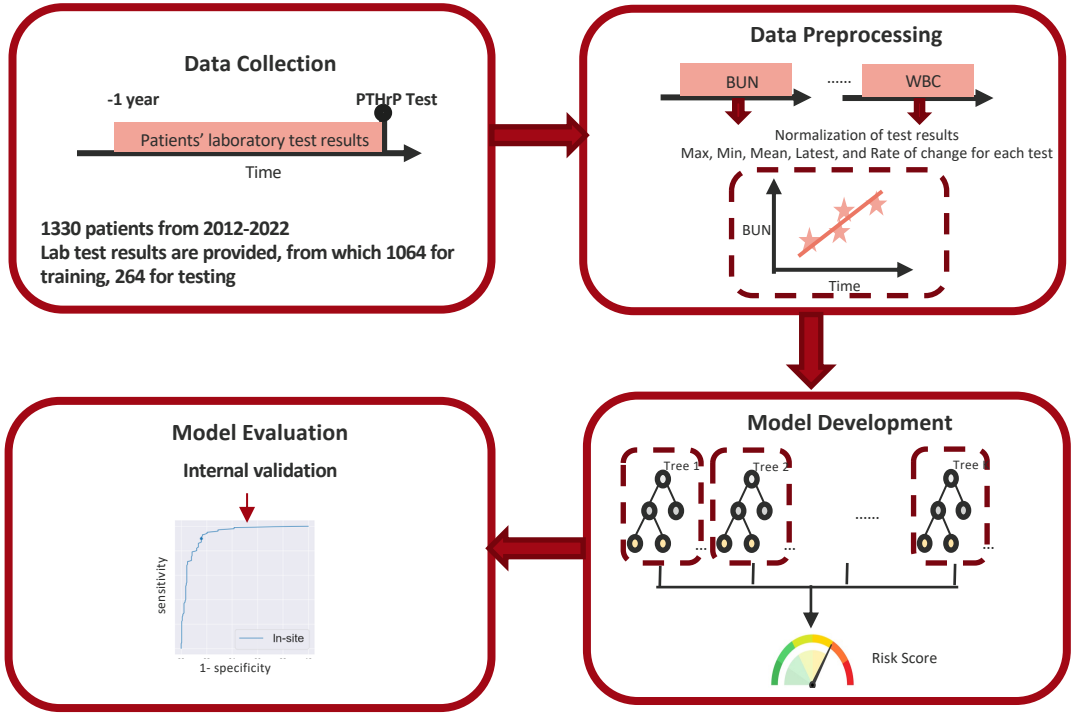


(Algorithm adopted from ARUP Consult)

"The purpose of this competition was to see if a machine-learning approach could better predict test outcomes compared to the traditional manual approach that many clinical laboratories use, reviewing calcium and PTH results to identify potential inappropriate PTHrP orders, thereby improving the PTHrP test utilization. "

(https://www.kaggle.com/competitions/aacc-2022-predicting-pthrp-results)

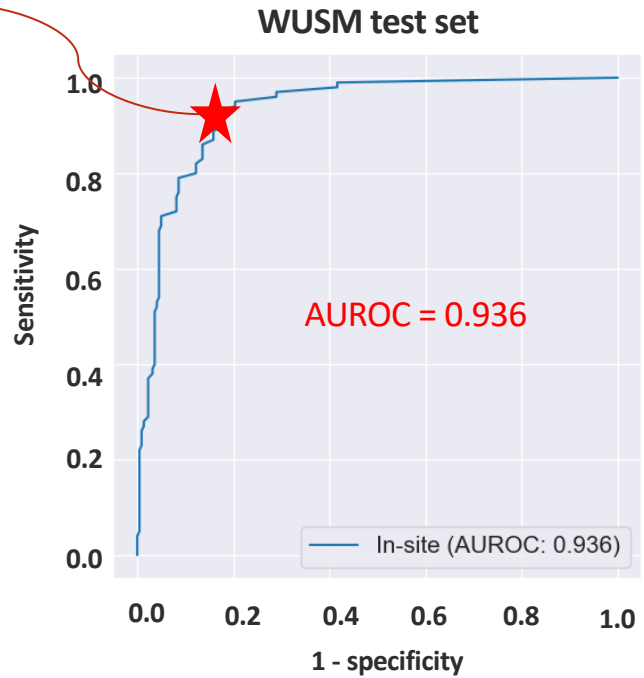
Our Workflow of PTHrP Model Development and Evaluation



Yang et al., manuscript under review

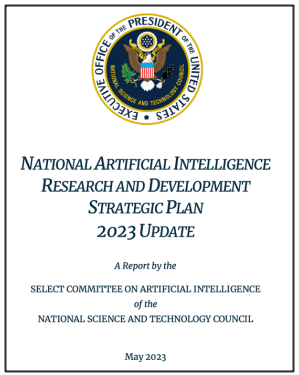
Winner Solution of the Competition

Sensitivity = 0.900
Specificity = 0.842
Precision = 0.539



Concerns on Healthcare AI

Data Gathering and Cleaning → Data Pre-processing → Building a Predictive Model → Model Evaluation → Incorporate the Model in Healthcare Applications



Regulation of the government

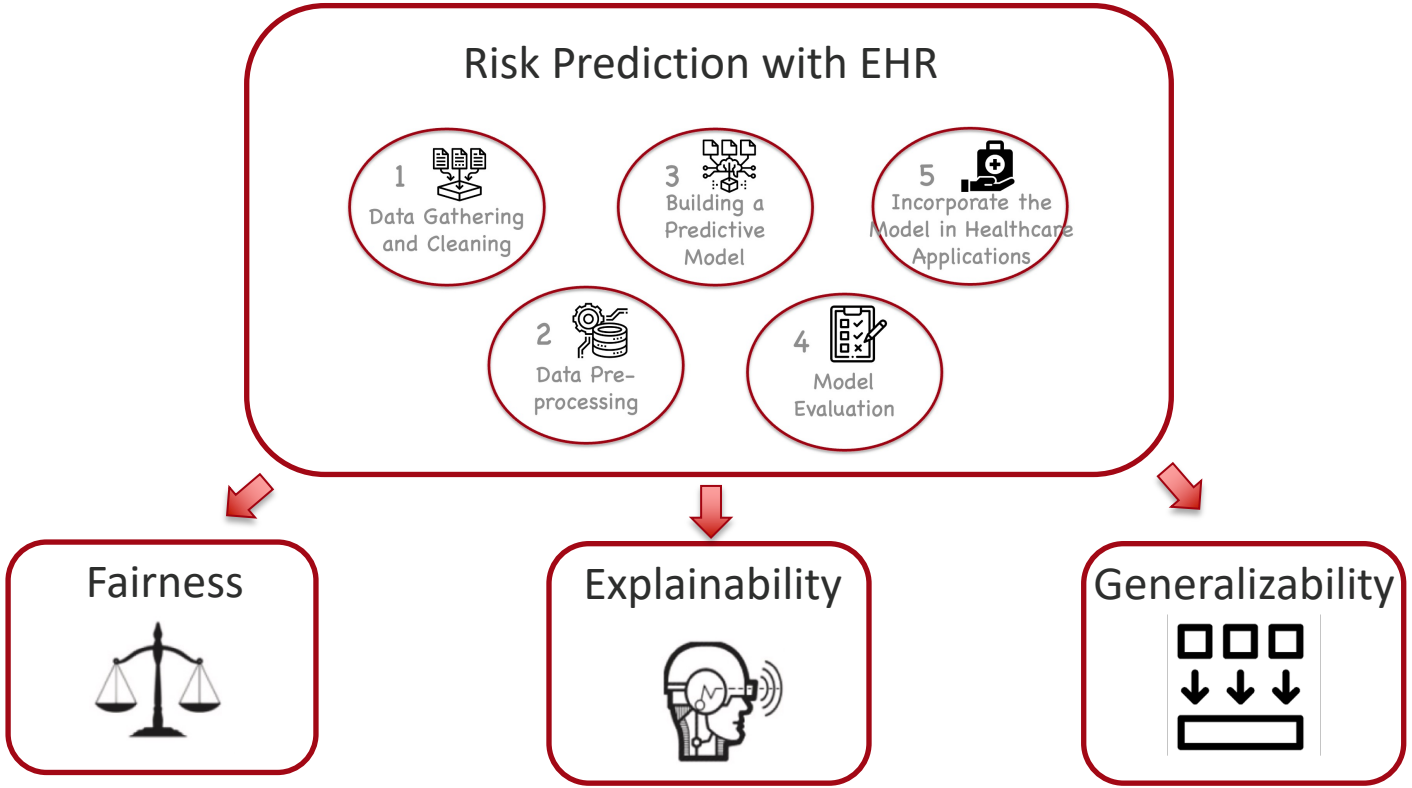
GDPR

Lack of trust from the clinicians

91% of healthcare insiders see artificial intelligence boosting access to care, but 75% believe it could threaten the security and privacy of patient data.

(Stat source: https://hitinfrastructure.com/news/challenges-of-artificial-intelligence-adoption-in-healthcare)

Essential Issues



Risk Prediction with EHR

Fairness

Explainability

Generalizability

- Definition and measurement of algorithmic disparity
- Methods to address algorithmic disparity
- Address disparity across multiple sites

What is Algorithmic Disparity

- Algorithmic bias/Algorithmic Disparity
- Describes systematic and repeatable errors in a computer system that create “unfair” outcomes, such as "privileging" one category (group) over another.(From wikipedia)
  - Different from data bias
  - Focus on group-based algorithmic disparity in this tutorial
  - Group (category) usually refer to gender/race/age in the healthcare problems

Algorithmic Disparity in Healthcare

- The algorithm used predicted healthcare costs rather than illness to calculate the risk scores and decided whether a patient should be enrolled in a healthcare management program
- Within patients at the same percentile risk score (97%), black patients went on to be far less healthy than white patients

Science Contents News Careers Journals

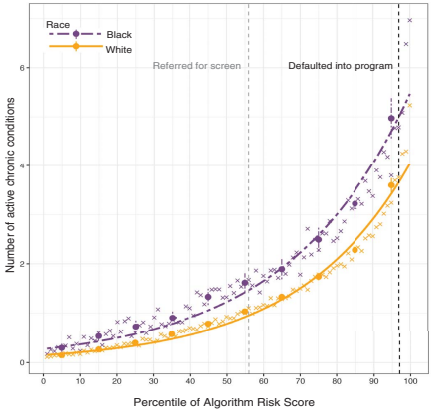
SHARE RESEARCH ARTICLE

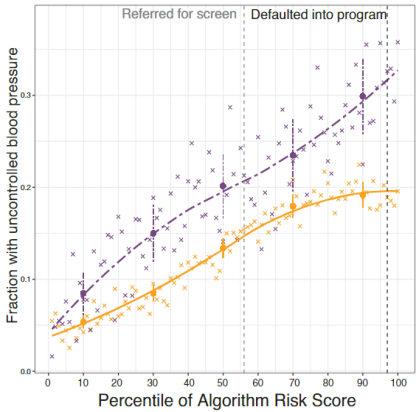
Dissecting racial bias in an algorithm used to manage the health of populations

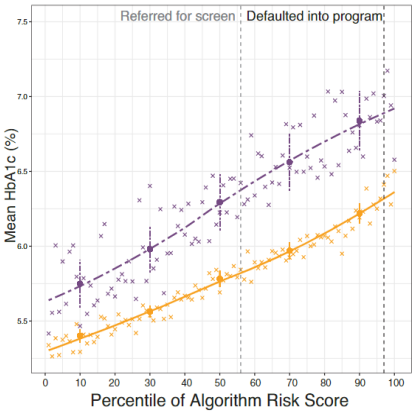
Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]

+ See all authors and affiliations

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







Algorithmic Disparity in Healthcare

- White female pregnant individuals are more likely to be diagnosed with postpartum depression and receive mental health services than black individuals
- Clinical prediction models trained on such biased data may produce unfair outcomes

JAMA Network Open

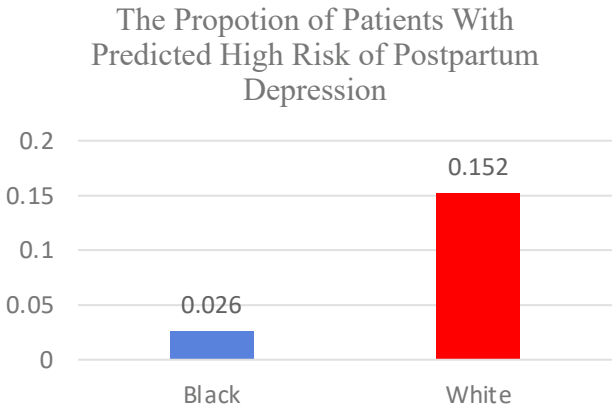
Original Investigation | Health Informatics

Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression

Yoonyoung Park, ScD; Jianying Hu, PhD; Moninder Singh, PhD; Issa Sylla, BA; Irene Dankwa-Mullan, MD, MPH; Eileen Koski, MPhil; Amar K. Das, MD, PhD

Table 2. Selected Characteristics of Pregnant Women Enrolled in Medicaid (2014-2018)

Characteristic	Participants, No. (%) All (N = 573 634)[a]	White (n = 314 903)	Black (n = 217 899)	Standardized difference
Age, mean (SD), y	26.1 (5.5)	26.0 (5.4)	26.2 (5.5)	0.0
Plan type (HMO vs others)	419 483 (73.1)	233 649 (74.2)	162 338 (74.5)	0.0
Pregnancy outcome				
Preterm birth	49 024 (8.5)	24 055 (7.6)	22 051 (10.1)	0.1
Preeclampsia	33 297 (5.8)	15 529 (4.9)	15 841 (7.3)	0.1
Cesarean delivery	167 834 (29.3)	90 554 (28.8)	66 793 (30.7)	0.0
High-risk pregnancy flag	193 437 (33.7)	106896 (33.9)	74 935 (34.4)	0.0
Postpartum period				
Postpartum depression[c]	70 821 (12.3)	52 370 (16.6)	15 410 (7.1)	0.3
Any mental health related visits	48 781 (8.5)	34 044 (10.8)	12 612 (5.8)	0.2
HEDIS-qualifying visits	192 427 (33.5)	107 965 (34.3)	70 216 (32.2)	0.0



Define and Measure Disparity

Notation	Definition	Example
$A \in \{0,1\}$	**Protected/sensitive attribute**, a grouping variable with respect to which we wish to guarantee fairness	Race(white, non-white)
$\mathbf{X}$	**Features,** all variables except A which are inputs of the model	Other demographic information, lab tests, etc
$Y \in \{0,1\}$	**True label**	Diagnosed as postpartum depression
S	**Predicted score,** indicating the predicted probability of $Y = 1$	A predicted score indicating risk of postpartum depression
$\hat{Y} \in \{0,1\}$	**Predicted label**	$\hat{Y} = 1$: Predicted high risk of postpartum depression

Define and Measure Disparity

Demographic Parity (Statistical Parity)

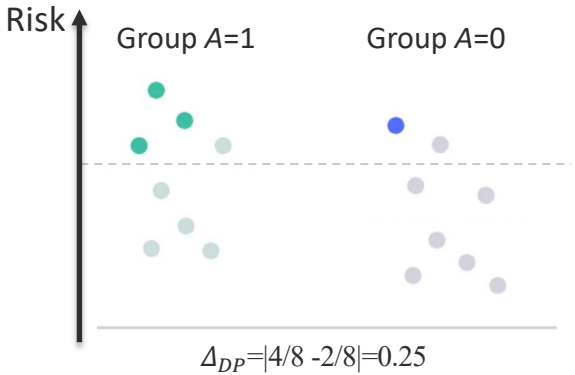
- Formulation

$$\mathrm{P}(\hat{Y}=1|A=1)=\mathrm{P}(\hat{Y}=1|A=0)$$

- Disparity

$$\Delta_{DP}=|\mathrm{P}(\hat{Y}=1|A=1)-\mathrm{P}(\hat{Y}=1|A=0)|$$

- Equalized averaged predictions



Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness." In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214-226. 2012.

Equalized Opportunity

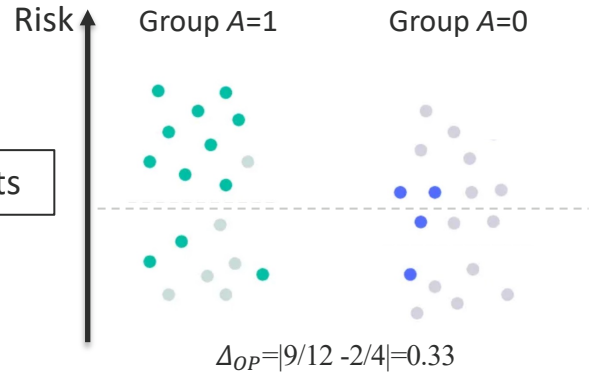
- Formulation

$$\mathrm{P}(\hat{Y}=1|A=1)=\mathrm{P}(\hat{Y}=1|A=0)$$

- Disparity

$$\Delta_{OP}=|\mathrm{P}(\hat{Y}=1|A=1,Y=1)-\mathrm{P}(\hat{Y}=1|A=0,Y=1)|$$

- Equalized true positive rates



Y=1: color dots

Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." *Advances in neural information processing systems* 29 (2016).

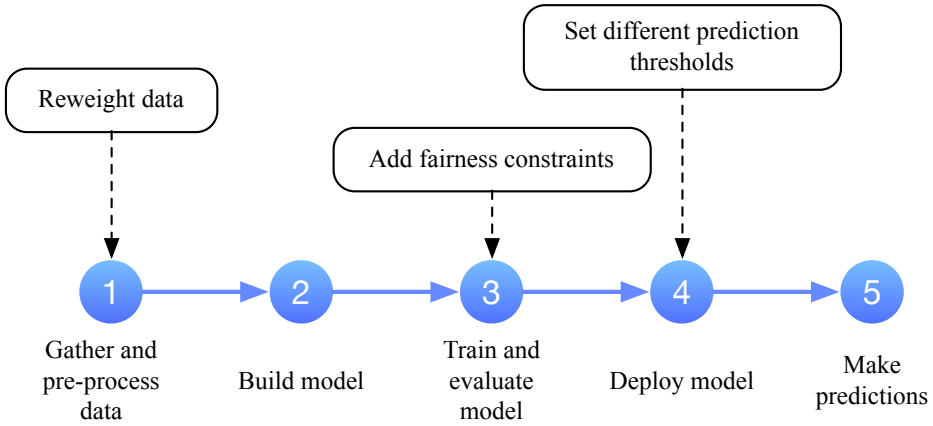
How to Address Disparity

Fairness through Unawareness

- The protected attribute is not explicitly used in the model
- Many variables in the data are proxies for the protected attribute (zip code-race)

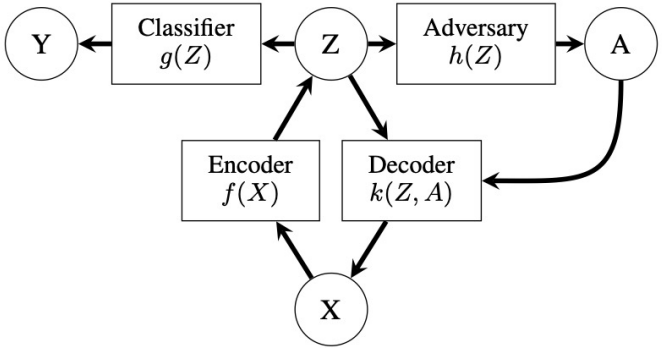
Fair Learning Algorithms

- Pre-process
- In-process
- Post-process



Learning Fair Representation

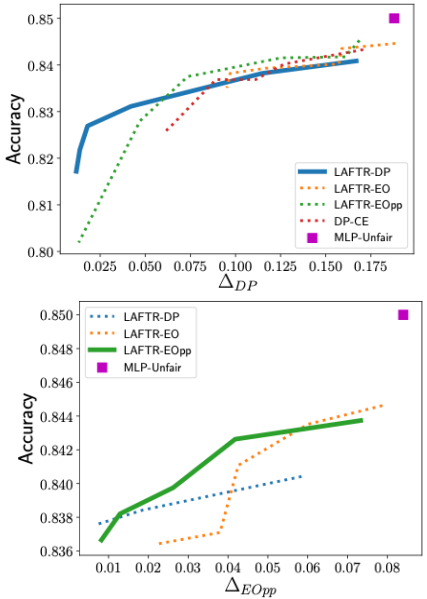
- Learn a fair encoder which maintains useful information in *X* to predict *Y* and hides the sensitive attributes



$$\underset{f,g,k}{\text{minimize}}\ \underset{h}{\text{maximize}}\ \mathbb{E}_{X,Y,A}\left[L(f,g,h,k)\right],$$

$$\mathcal{L}(f,g,h,k) = \alpha\mathcal{L}_{Class} + \beta\mathcal{L}_{Dec} - \gamma\mathcal{L}_{Adv}$$

- Tradeoffs between accuracy and various fairness metrics yielded by different LAFTR loss functions

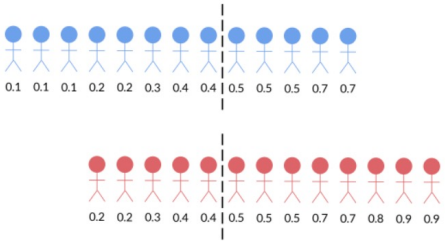


Madras, David, Elliot Creager, Toniann Pitassi, and Richard Zemel. "Learning adversarially fair and transferable representations." In *International Conference on Machine Learning*, pp. 3384-3393. PMLR, 2018.

Post-Processing Risk Scores

- For a given threshold θ, the prediction is decided by the predicted score: $\hat{Y} = \mathbb{I}[S \geq \theta]$
- Original disparity:

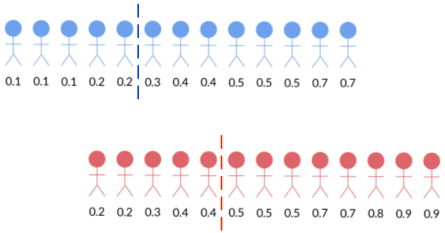
$$\Delta_{DP} = \left|\frac{5}{13} - \frac{8}{13}\right| = 0.23$$



- Set different thresholds for different groups
- Disparity after adjustment:

$$\Delta_{DP} = \left|\frac{8}{13} - \frac{8}{13}\right| = 0$$

- **Advantage: model-agnostic**



Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." *Advances in neural information processing systems* 29 (2016).
Figure Source: https://data102.org/sp20/assets/notes/notes05.pdf

Applications in Healthcare

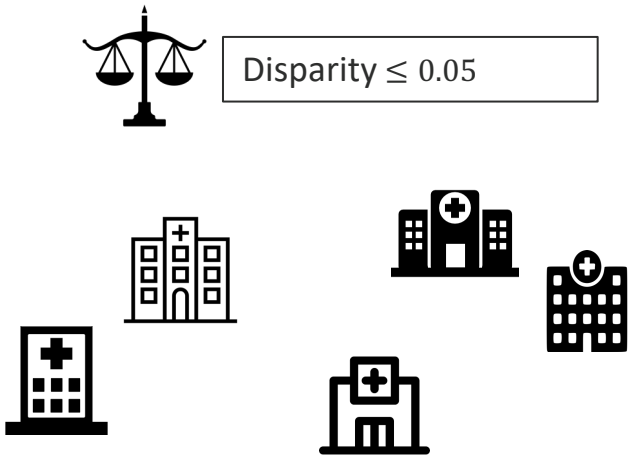
- Numerous works have been published to evaluate and mitigate the algorithmic disparity in healthcare applications
- Build risk prediction model and consider the algorithmic disparity with data from single site
- Problem: the amount of healthcare data from single site may be not enough to train a good model

Author (year)	Clinical objective	How was fairness evaluated?	Was racial bias identified?	How was the AI[a] model biased?	Was racial bias mitigated?	Protected class
Abubakar et al (2020) [29]	Identification of images of burns vs healthy skin	Accuracy	Yes	Poor accuracy of models trained on a Caucasian data set and validated on an African data set and vice versa	Yes	Dark-skinned patients, light-skinned patients
Allen et al (2020) [30]	Intensive care unit (ICU) mortality prediction	Equal opportunity difference (FNR[b] disparity)	N/A[c]	N/A	Yes	Non-White patients
Briggs and Hollmén (2020) [31]	Prediction of future health care expenditures of individual patients	Balanced accuracy, statistical parity, disparate impact, average odds, equal opportunity	N/A	N/A	Yes	Black patients
Burlina et al (2021) [32]	Diagnosis of diabetic retinopathy from fundus photography	Accuracy	Yes	Lower diagnostic accuracy in darker-skinned individuals compared to lighter-skinned individuals	Yes	Dark-skinned patients
Chen et al (2019) [33]	ICU mortality prediction, psychiatric readmission prediction	Error rate (0-1 loss)	Yes	Differences in error rates in ICU mortality between racial groups	No	Non-White patients
Gianattasio et al (2020) [34]	Dementia status classification	Sensitivity, specificity, accuracy	Yes	Existing algorithms varying in sensitivity and specificity between race/ethnicity groups	Yes	Hispanic, non-Hispanic Black patients
Noseworthy et al (2020) [35]	Prediction of left ventricular ejection fraction ≤35% from the electrocardiogram (ECG)	AUROC[d]	No	N/A	No	Non-White patients
Obermeyer et al (2019) [36]	Prediction of future health care expenditures of individual patients	Calibration	Yes	Black patients with a higher burden than White patients at the same algorithmic risk score	Yes	Black patients
Park et al (2021) [37]	Prediction of postpartum depression and postpartum mental health service utilization	Disparate impact, equal opportunity difference (TPR[e] disparity)	Yes	Black women with a worse health status than White women at the same predicted risk level	Yes	Black patients
Seyyed-Kalantari et al (2021) [38]	Diagnostic label prediction from chest X-rays	Equal opportunity difference (TPR disparity)	Yes	Greater TPR disparity in Hispanic patients	No	Non-White patients
Thompson et al (2021) [39]	Identification of opioid misuse from clinical notes	Equal opportunity difference (FNR disparity)	Yes	Greater FNR in the Black subgroup than in the White subgroup	Yes	Black patients
Wissel et al (2019) [40]	Assignment of surgical candidacy score for patients with epilepsy using clinical notes	Regression analysis of the impact of the race variable on the candidacy score	No	N/A	No	Non-White patients

Huang, Jonathan, Galal Galal, Mozziyar Etemadi, and Mahesh Vaidyanathan. "Evaluation and mitigation of racial bias in clinical machine learning models: scoping review." *JMIR Medical Informatics* 10, no. 5 (2022): e36388.

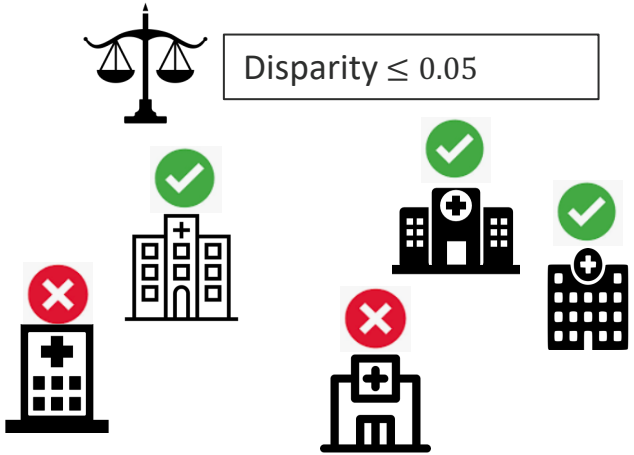
Address Disparity in Multiple Sites

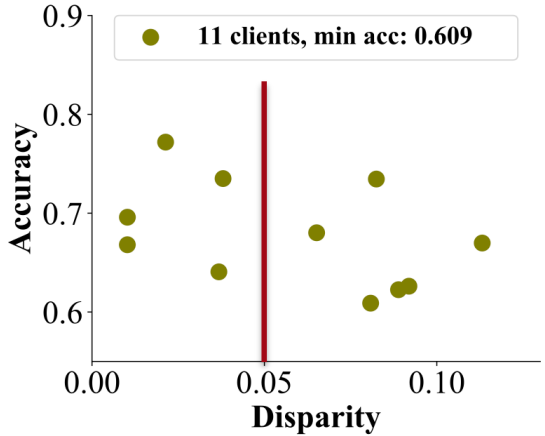
- Training a global model across multiple sites (e.g., hospitals)
- A global standard on disparity that each site should satisfy simultaneously with the global model
- It is challenging to achieve this due to the distribution shift across sites



Address Disparity in Multiple Sites

- Intensive Care Unit data from 11 hospitals
- Fairness: demographic parity (*Y*: Length of Stay > 1 week, *A*: Race)
- With existing methods, the disparities on some clients are still very high

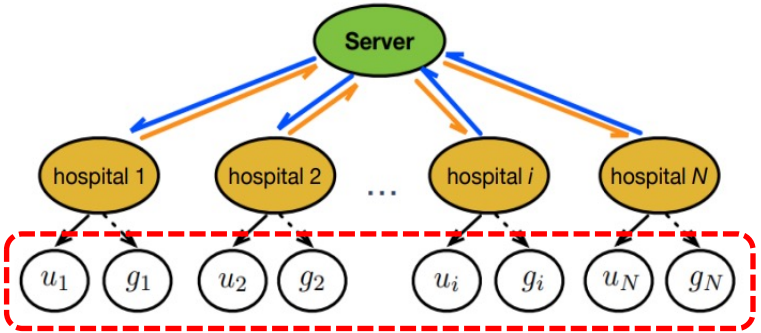




Restrict Disparity for Each Site

Goal

- Address algorithmic disparities in multiple sites



u_i: utility on site i
g_i: disparity on site i

Cui, Sen, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. "Addressing algorithmic disparity and performance inconsistency in federated learning." *Advances in Neural Information Processing Systems* 34 (2021): 26091-26102.

Restrict Disparity for Each Site

Goal

- Address algorithmic disparities in multiple sites

Subgoals

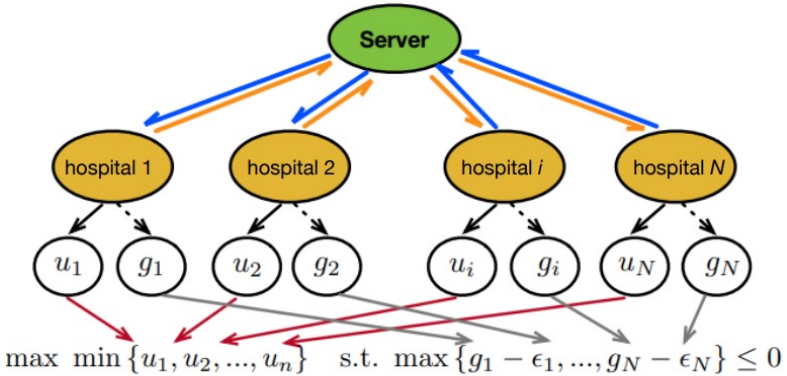
- (Fairness) All sites achieve fairness

$$g_i \leq \varepsilon_i, \forall\, i$$

- (Utility) Maximize the minimal utility across all sites

$$\max \min\{u_1, ..., u_N\}$$

$$\begin{aligned} \min_{h \in \mathcal{H}} \quad & l_{max}(h) \\ & l_{max}(h) = \max\left(l_i(h)\right) \quad i \in \{1, ., N\} \\ \text{s.t.} \quad & g'_{max}(h) \leq 0 \\ & g'_{max}(h) = \max\left(g'_i(h)\right) = \max\left(g_i(h) - \epsilon_i\right) \quad i \in \{1, ., N\} \end{aligned}$$

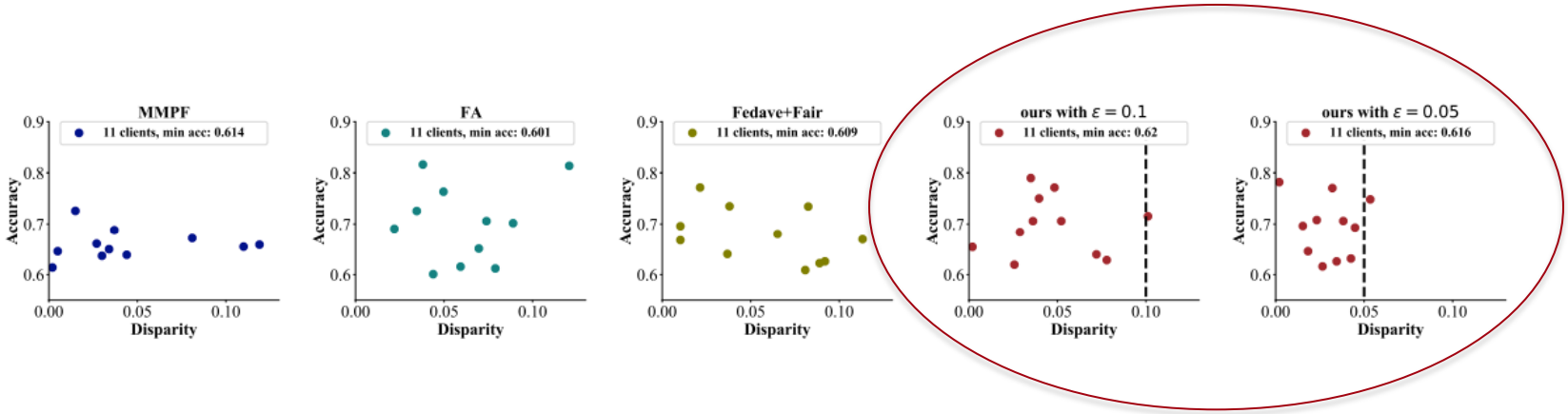


u_i: utility on site i
g_i: disparity on site i
ϵ_i: fairness budge on site i

Cui, Sen, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. "Addressing algorithmic disparity and performance inconsistency in federated learning." *Advances in Neural Information Processing Systems* 34 (2021): 26091-26102.

Fairness Results on All Hospitals

- Satisfy fairness constraints on all hospitals
- Feasible under different levels of fairness constraints



Cui, Sen, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. "Addressing algorithmic disparity and performance inconsistency in federated learning." *Advances in Neural Information Processing Systems* 34 (2021): 26091-26102.

Takeaways

- Risk prediction models applied to healthcare are faced with algorithmic disparity
- The method to address algorithmic disparity should balance the fairness-utility trade-off
- It is also important to consider algorithmic disparity on multiple sites

Risk Prediction with EHR

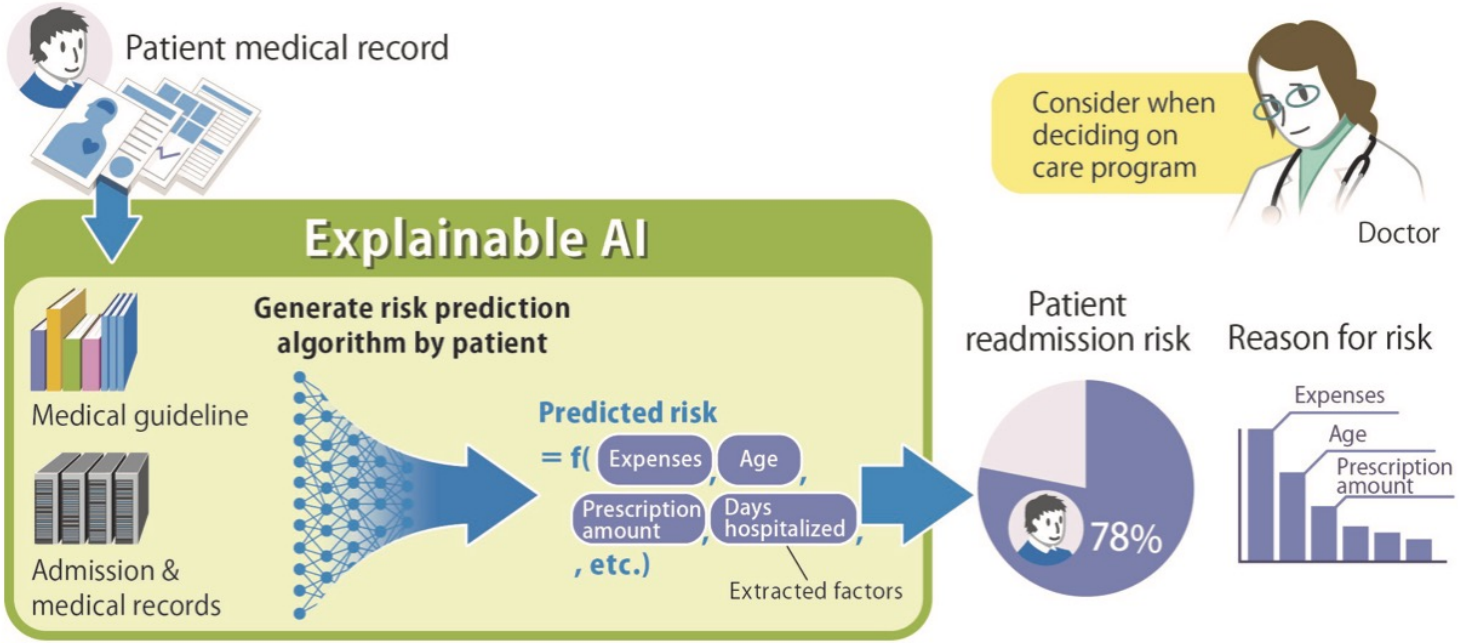
Fairness

Explainability

Generalizability

- The need of explanation in healthcare risk prediction
- Methods on explainable machine learning
  - Learn directly interpretable model
  - Post-hoc explanation
- Explanations of prediction in healthcare applications

Need of Explanation



(Source: https://www.hitachi.us/press/partners-connected-health-and-hitachi-develop-an-explainable-ai-technology-to-help-doctors-predict-readmissions-and-improve-patient-outcomes)

Explainable Machine Learning Methods

Directly interpretable

The oldest AI techniques—such as decision rule sets, decision trees, and decision tables—can be simple enough for people to understand. Supervised learning of these models is directly interpretable.

vs.

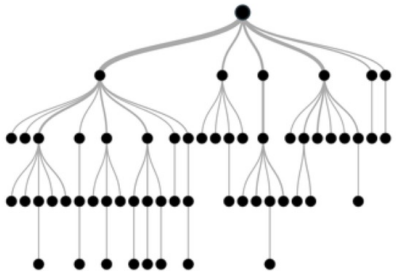
Post hoc interpretation

Starts with a black box model and probes into it with a companion model to create interpretations. The black box model continues to provide the actual prediction while interpretation improves human interactions.

Hind, Michael. "Explaining explainable AI." *XRDS: Crossroads, The ACM Magazine for Students* 25.3 (2019): 16-19.

Learn Directly Interpretable Model

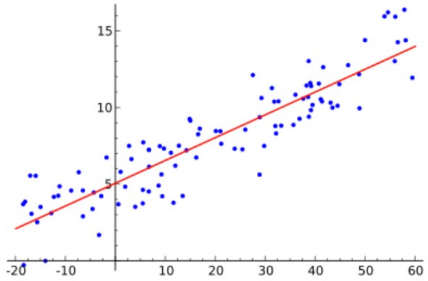
Rule Based Models



Rules for Length of Stay Prediction

IF age > 55 AND gender = male
AND condition = 'COPD' AND
complication = 'YES'
THEN
Length of stay = long (> 7 days)

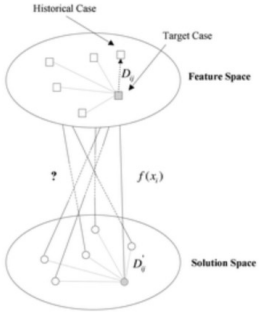
Relative Variable Importance



Top 5 Variables for Length of Stay Prediction

Variable	*Importance*
Age	0.45
Gender	0.37
Diabetic	0.32
Race	0.21
Smoker	0.14

Case Based Models



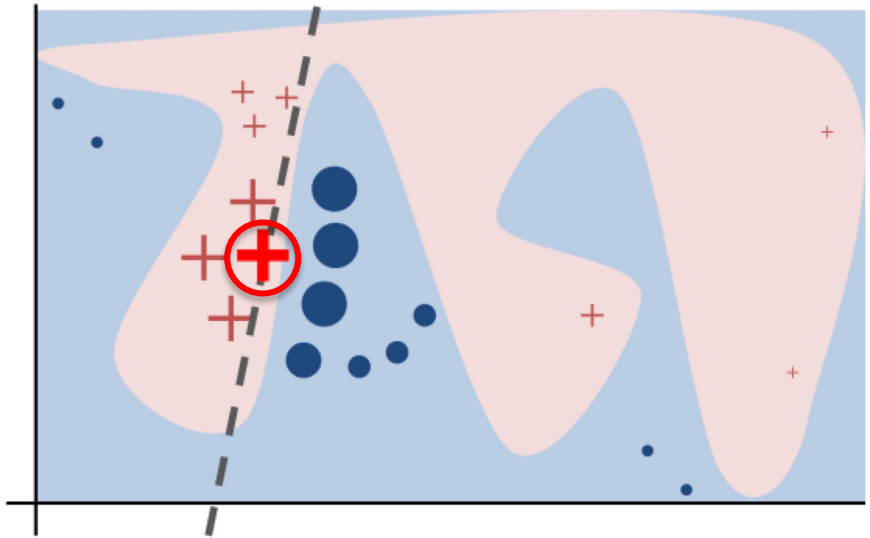
Example cases for Length of Stay Prediction

Patient X is predicted to have a length of stay of 20 days because he is most similar to these 5 patients who on average had length of stay of 5 days

(Source: Tutorial on KDD2018: Explainable AI in Healthcare)

LIME (Local Interpretable Model-agnostic Explanations)

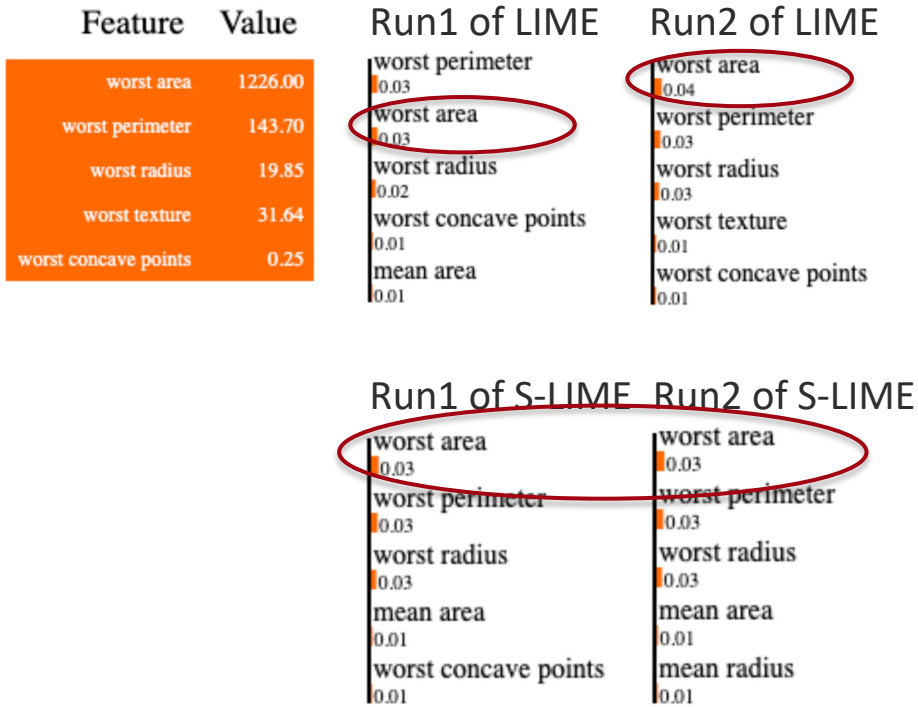
- Explains the predictions of any model *f* by approximating it locally with a linear model
- For a given sample ⊕, samples instances around ⊕, gets their predictions using *f*, and weighs them by the similarity to the instance being explained (represented by size)
- The dashed line is the explanation learned locally



Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

Stabilized LIME

- LIME has been shown to exhibit large instability across different samples, which harms user trust
- Stabilized LIME: Automatically and adaptively determine the number of perturbations needed to guarantee a stable explanation
- Results on Breast Cancer dataset
  - The explanations by stabilized LIME are more stable than LIME across different runs



Zhou, Zhengze, Giles Hooker, and Fei Wang. "S-lime: Stabilized-lime for model explanation." *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 2021.

SHAP (SHapley Additive exPlanations)

- Game theoretic method to calculate feature contributions to the prediction
- Each feature is a player in a game where the prediction is the payout
- The Shapley value - a method from coalitional game theory to attribute the payout among the players

X={age = 55, smoke = yes}, *f(X) = 0.75* *(f(X):* risk of long-term mortality*)*
f(∅) =*E[f(x)]*= 0.2 *f*({smoke = yes}) = 0.4 *f*({age = 55}) = 0.5 *f*({age = 55, smoke = yes}) = 0.75

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

Permutation	Marginal for age = 55
∅	*f*({age = 55}) - *f*({∅}) = 0.5 – 0.2 = 0.3
{smoke = yes}	*f*({age = 55, smoke = yes}) - *f*({smoke = yes}) = 0.75 - 0.4 = 0.35
ϕ_{age}	(0.35 + 0.3) / 2 = 0.325

Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).

Predict RT-PCR Test with Routine Lab Tests

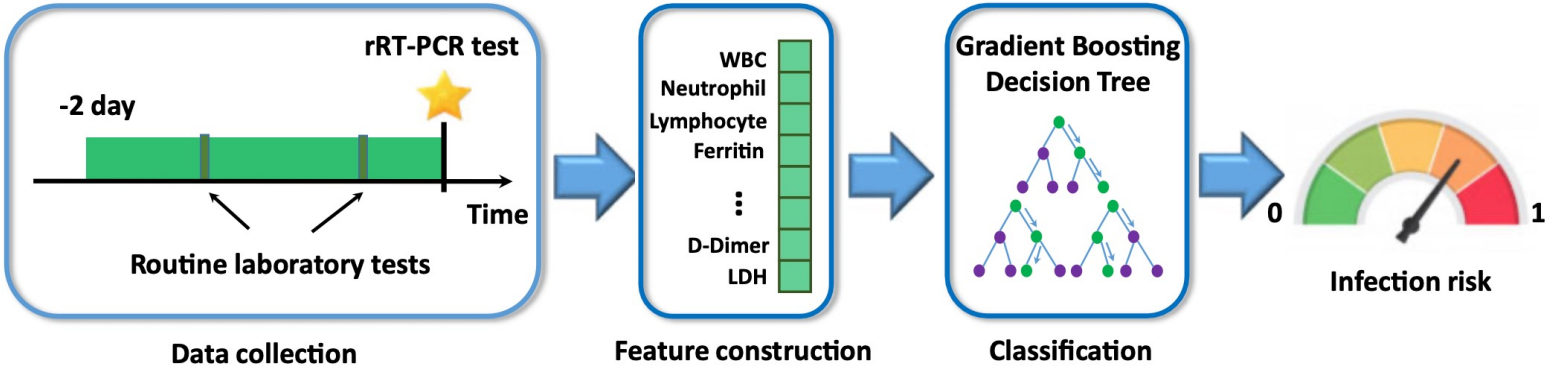
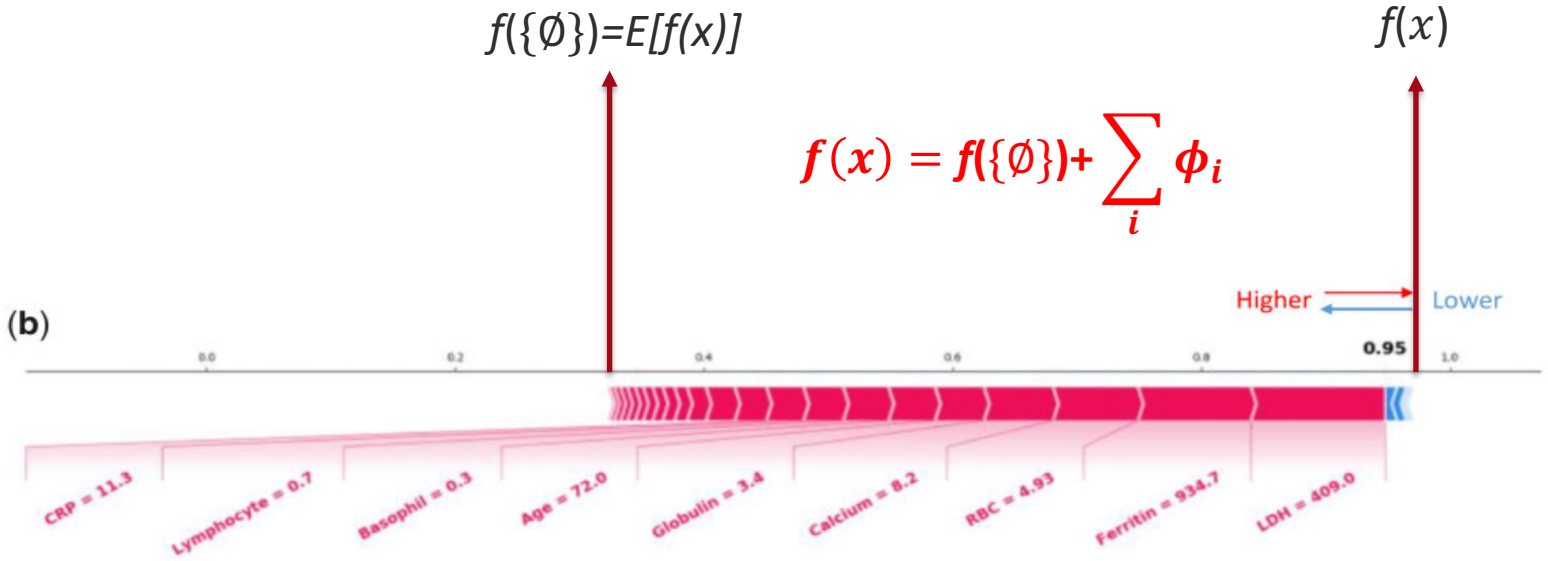


Fig. 2. Illustration of the modeling pipeline. Routine laboratory testing results completed within 2 days prior to the release of RT-PCR results were used to construct a vector, upon which a classifier was built to predict the RT-PCR positive or negative result. Each dimension of the vector corresponds to a specific laboratory test, and its value corresponds to the average of all results of this laboratory test taken during the collection window. The model outputs a probability score ranging from 0–1, indicating the risk of SARS-CoV-2 infection.

Yang, He S., Yu Hou, Ljiljana V. Vasovic, Peter AD Steel, Amy Chadburn, Sabrina E. Racine-Brzostek, Priya Velu et al. "Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning." *Clinical chemistry* 66, no. 11 (2020): 1396-1404.

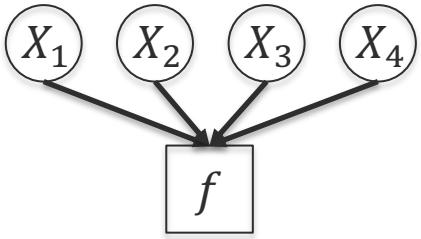
Shapley Values of COVID Infection Prediction

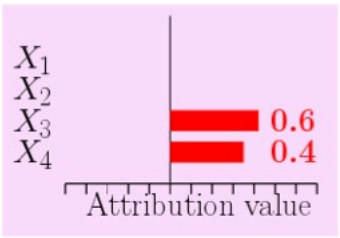


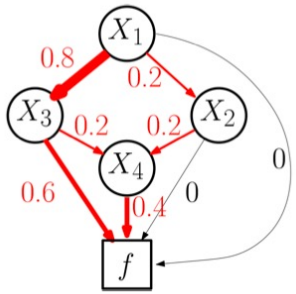
Yang, He S., Yu Hou, Ljiljana V. Vasovic, Peter AD Steel, Amy Chadburn, Sabrina E. Racine-Brzostek, Priya Velu et al. "Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning." *Clinical chemistry* 66, no. 11 (2020): 1396-1404.

Shapley Flow

- Original SHAP assumed the features to be independent, while the features may follow a causal structure in real-world problems
- Shapley Flow: A graph-based approach to explain model prediction by attributing the prediction to the edges on the graph





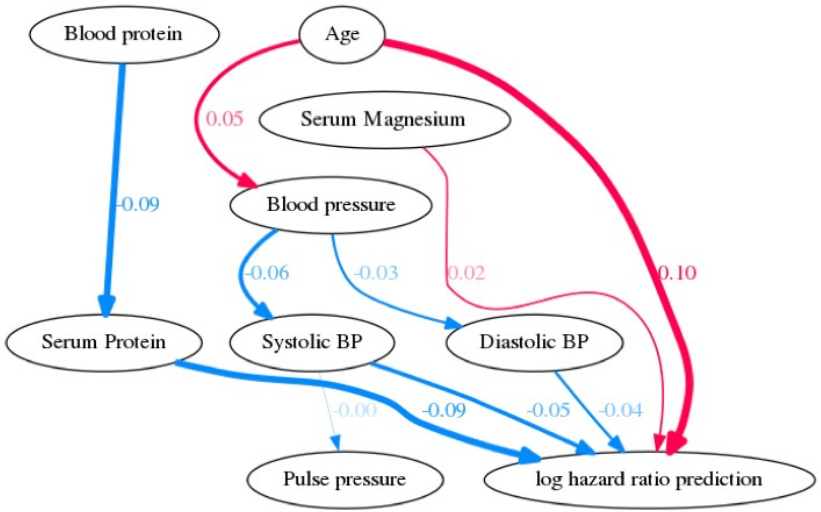


Wang, Jiaxuan, Jenna Wiens, and Scott Lundberg. "Shapley flow: A graph-based approach to interpreting model predictions." In *International Conference on Artificial Intelligence and Statistics*, pp. 721-729. PMLR, 2021.

Shapley Flow

- Dataset based on National Health and Nutrition Examination Survey
- Explain the risk of death over 15-yr followup for a given individual

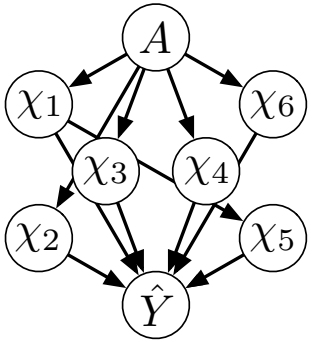
Attributions	Independent
Age	0.1
Serum Magnesium	0.02
Serum Protein	-0.09
Blood pressure	0.0
Systolic BP	-0.05
Diastolic BP	-0.04
Serum Cholesterol	0.0
Serum Albumin	0.0
Blood protein	0.0
White blood cells	0.0
Race	0.0
BMI	-0.0
TIBC	0.0
Sex	0.0
TS	0.0
Pulse pressure	0.0
Poverty index	0.0
Red blood cells	0.0
Serum Iron	0.0
Sedimentation rate	0.0
Iron	0.0
Inflamation	0.0



Wang, Jiaxuan, Jenna Wiens, and Scott Lundberg. "Shapley flow: A graph-based approach to interpreting model predictions." In *International Conference on Artificial Intelligence and Statistics*, pp. 721-729. PMLR, 2021.

FACTS (Fairness-Aware Causal paTh decompoSition)

- Specifically target at explaining the disparity
- Provide more comprehensive explanation of disparity
- **Highlighted in AMIA (American Medical Informatics Association) 2021 Year-in-Review Session**



Feature-based explanations on disparity

Features	ISV
Poverty Idx, Food Program (χ_1)	0.0318
Blood pressure (χ_2)	0.0141
Serum magnesium (χ_3)	0.0076
Blood protein (χ_4)	0.0064
Sedimentation rate (χ_5)	0.0099
White blood cells, Red blood cells (χ_6)	-0.0077

Path-based explanations

Paths	$\Phi_f(p_i)$
$A \to \chi_1 \to \hat{Y}$	0.0324
$A \to \chi_2 \to \hat{Y}$	0.0126
$A \to \chi_3 \to \hat{Y}$	0.0081
$A \to \chi_4 \to \hat{Y}$	0.0077
$A \to \chi_5 \to \hat{Y}$	0.0060
$A \to \chi_1 \to \chi_5 \to \hat{Y}$	0.0031
$A \to \chi_6 \to \hat{Y}$	-0.0082

Pan, Weishen, Sen Cui, Jiang Bian, Changshui Zhang, and Fei Wang. "Explaining algorithmic fairness through fairness-aware causal path decomposition." In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1287-1297. 2021.

Takeaways

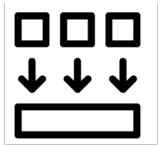
- Explanability is essential for clinicians/patients to trust prediction of machine learning model in healthcare AI
- Post-hoc explanations (e.g., Shapeley values) are model-agonistic and more applicable
- Taking the causal structure between the features into consideration could make the explanations more comprehensive

Risk Prediction with EHR

Fairness

Explainability

Generalizability

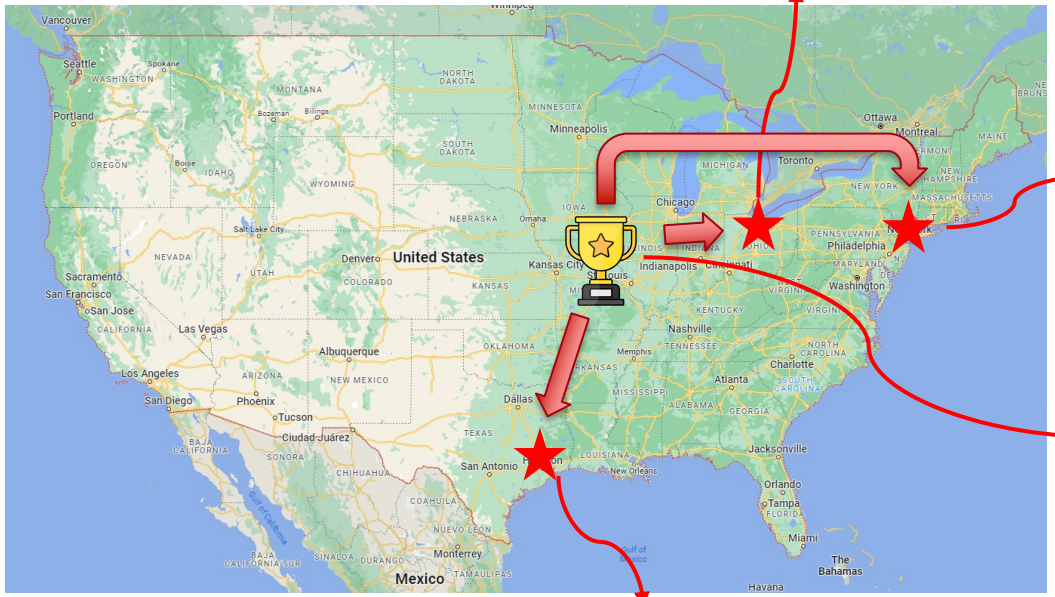


Generalizability/Transferability

- The need to apply a model to external sites
- Factors that affect model generalizability
- Strategies to improve model generalizability

Apply the Champion Model to External Sites

<u>Ohio State University Wexner Medical Center (OSU)</u>
Data under analysis



<u>Weill Cornell Medicine (WCM):</u>
1101 PTHrP orders from 2017 to 2022
PTHrP positive rate 16.9%
Send-out lab: Quest Diagnostics

<u>Washington University School of Medicine in St. Louis (WUSM)</u>
1330 PTHrP orders from 2012 to 2022
Positive rate 17.5%
Send-out lab: Mayo Clinic Laboratories

<u>University of Texas M.D. Anderson Cancer Center (MDA)</u>
1090 PTHrP orders from 2021-2022
PTHrP positive rate 23.9%
Send-out lab: Mayo Clinic Laboratories

Factors that Affect Model Generalizability

- Patient demographic characteristics
- Geographic features
- Instrument platforms
- Sample handling protocols and other pre-analytical factors
- Testing methodologies
- Send-out laboratories

Directly Applying the PTHrP Model

- When the ready-made model from WUSM was directly applied “as-is” to the two independent datasets, its performance moderately deteriorated in MDA but substantially deteriorated in WCM

(Yang, Sarina, Pan, Weishen, et.al. manuscript under review)

Strategies to Improve Model Performance

- Strategy 1: Re-training the XGBoost model using site-specific data with the same model architecture, feature sets, and hyperparameters
- Strategy 2: Re-building the model using site-specific data including feature selection, hyperparameter tuning and model parameter learning

When a ready-made model cannot be directly transported to external datasets due to the shift of data distribution, some local customization strategies can be utilized to improve model performance, such as re-training or re-building the model using site-specific data.

(Yang, Sarina, Pan, Weishen, et.al. manuscript under review)

What If a Hospital Has Limited Data to Re-Train the Model?

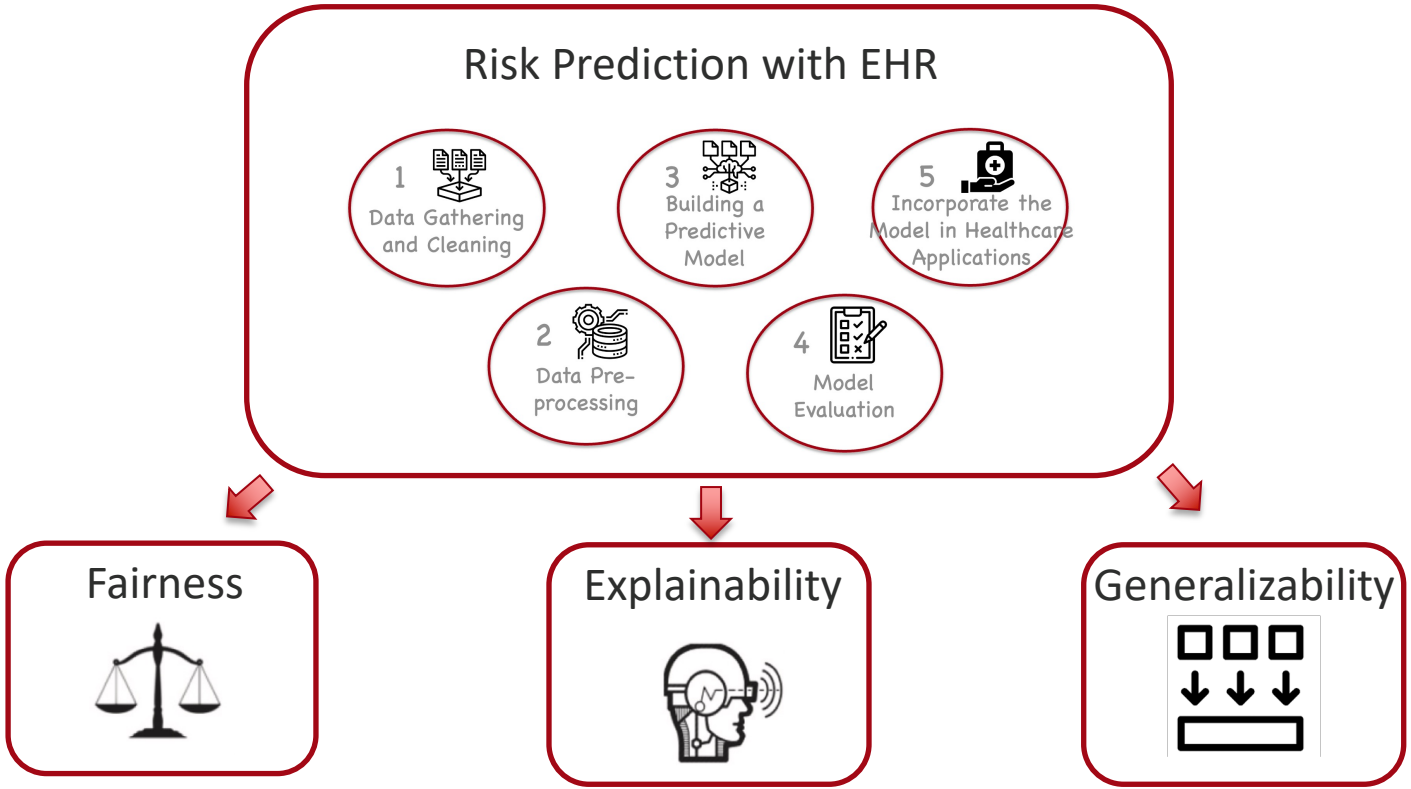
- Explored a model fine-tuning strategy in which the ready-made model is applied to hospitals with limited training data (low-resource scenarios)
- The fine-tuning strategy performed best when the amounts of available samples were relatively small (< 200). However, when the number of available samples exceeded 200, model re-training appeared to be a better option

(Yang, Sarina, Pan, Weishen, et.al. manuscript under review)

Takeaways

- It is essential to evaluate model generalizability in independent, external datasets.
- Directly transporting a ready-made model to external datasets may lead to performance deterioration due to data distribution shift. Model re-training or re-building could improve the performance when there are enough local data, whereas model fine-tuning may be a favorable strategy when site-specific data is limited.

Finlayson, Samuel G., Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. "The clinician and dataset shift in artificial intelligence." *New England Journal of Medicine* 385, no. 3 (2021): 283-286.


Risk Prediction with EHR
1
Data Gathering and Cleaning
2
Data Pre-processing
3
Building a Predictive Model
4
Model Evaluation
5
Incorporate the Model in Healthcare Applications
Fairness
Explainability
Generalizability

Risk Prediction with EHR

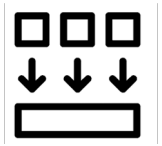
Risk Scores

Ranking Patients for Resource Allocation

Fairness

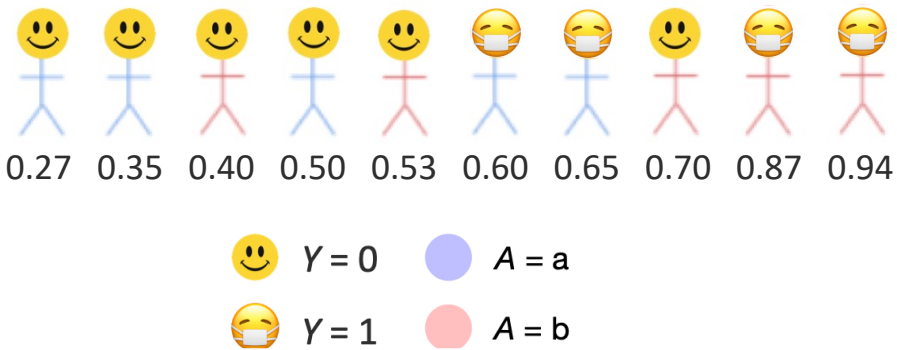
Explainability

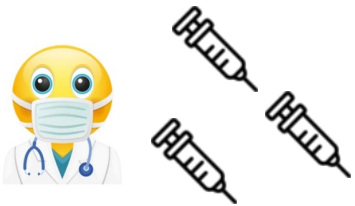
Generalizability



Disparity when Ranking by Risk Scores

- Allocating limited medical resources
- Ranking the people based on the risk score
- Ranking disparity !





Measuring Ranking Disparity

- AUC is calculated as the probability that a randomly drawn predicted score from the positive class (S_1) is higher than a predicted score from the negative class (S_0) :

$$\mathrm{AUC} = \Pr[S_1 > S_0]$$

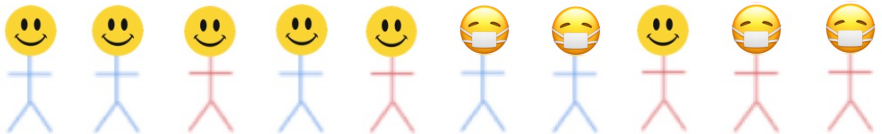
AUC:  vs 

- The xAUC of group a over b (Kallus et.al. 2019):

$$\mathrm{xAUC}(a,b) = \Pr\left[S_1^a > S_0^b\right]$$

xAUC(a,b):  vs 

  - S_1^a: predicted score from the positive class in group *a*
  - S_0^b: predicted score from the negative class in group *b*



Kallus, Nathan, and Angela Zhou. "The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric." *Advances in neural information processing systems* 32 (2019).

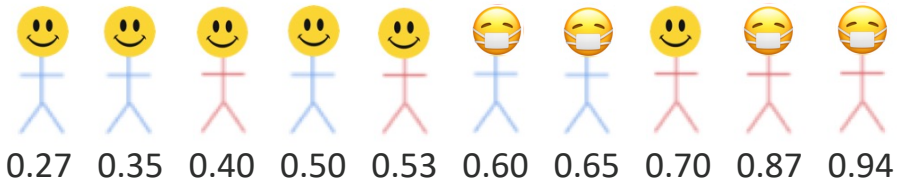
Measuring Ranking Disparity

- The xAUC of group a over b:

$$\mathrm{xAUC}(a,b) = \Pr\left[S_1^a > S_0^b\right]$$

- Ranking Disparity:

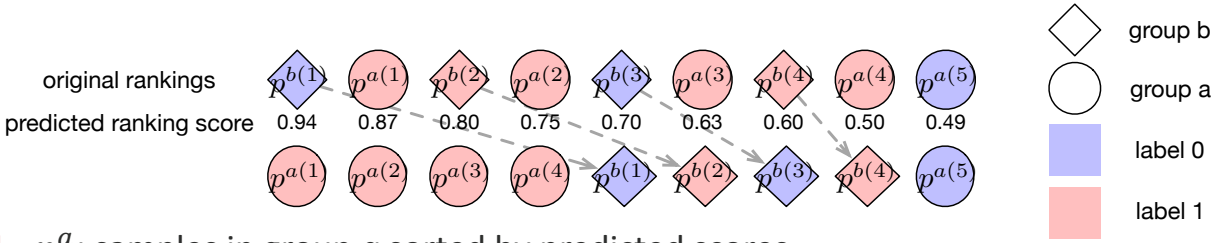
$$\begin{aligned}\Delta\mathrm{xAUC}(a,b) &= |\mathrm{xAUC}(a,b) - \mathrm{xAUC}(b,a)| \\ &= \left|\Pr\left[\mathrm{S}_1^a > \mathrm{S}_0^b\right] - \Pr\left[\mathrm{S}_1^b > \mathrm{S}_0^a\right]\right|\end{aligned}$$



Kallus, Nathan, and Angela Zhou. "The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric." *Advances in neural information processing systems* 32 (2019).

Post-hoc Ranking Adjustment

- Optimize the ordering directly



  - p^a: samples in group *a* sorted by predicted scores
  - p^b: samples in group *b* sorted by predicted scores
- Change cross-group ordering while keeping the inner-group ordering fixed (xOrder)
- Find the optimal ordering operation *o* while considering the fairness-utility trade-off explicitly

utility fairness

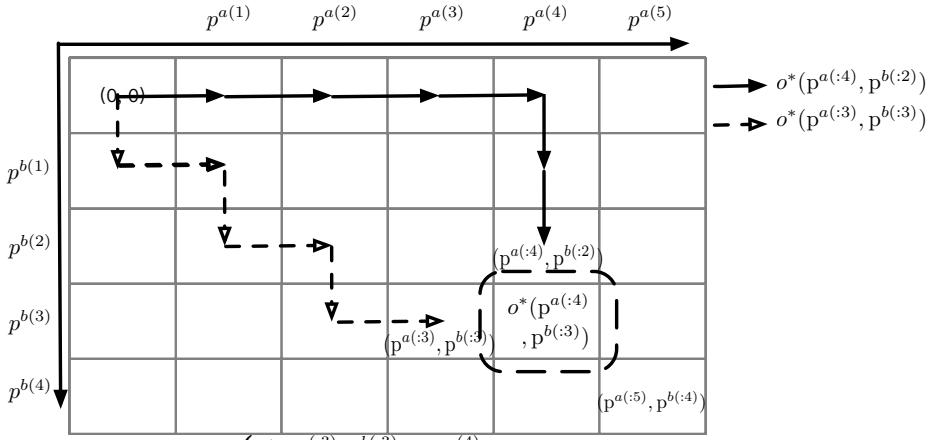
$$\mathcal{L}(o(p^a, p^b)) = \mathrm{AUC}(o(p^a, p^b)) - \lambda \cdot \Delta\mathrm{xAUC}(o(p^a, p^b))$$

Cui, Sen, Weishen Pan, Changshui Zhang, and Fei Wang. "Towards model-agnostic post-hoc adjustment for balancing ranking fairness and algorithm utility." In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 207-217. 2021.

Post-hoc Ranking Adjustment

$$\mathcal{L}(o(p^a, p^b)) = \underbrace{\mathrm{AUC}(o(p^a, p^b))}_{\text{utility}} - \lambda \cdot \underbrace{\Delta\mathrm{xAUC}(o(p^a, p^b))}_{\text{fairness}}$$

- Solve the through a dynamic programming process
- The efficiency of the proposed method can be proved theoretically

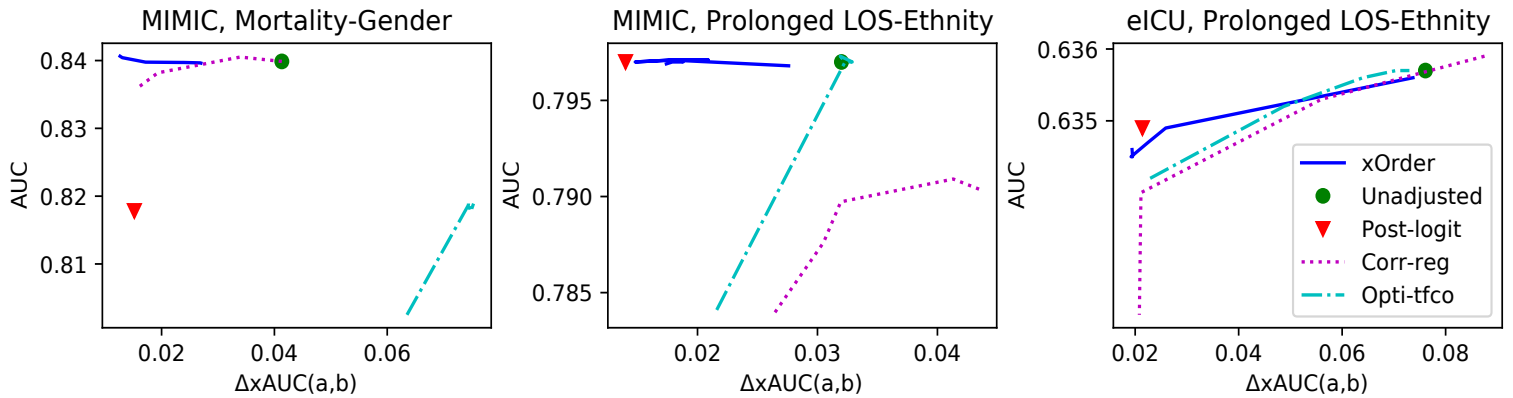


$$o^*(\mathrm{p}^{a(:4)}, \mathrm{p}^{b(:3)}) = \begin{cases} o^*(\mathrm{p}^{a(:3)}, \mathrm{p}^{b(:3)}) \oplus \mathrm{p}^{a(4)}, \\ \quad if\ \widehat{\mathrm{G}}(o^*(\mathrm{p}^{a(:3)}, \mathrm{p}^{b(:3)}) \oplus \mathrm{p}^{a(4)}) > \widehat{\mathrm{G}}(o^*(\mathrm{p}^{a(:4)}, \mathrm{p}^{b(:2)}) \oplus \mathrm{p}^{b(3)}), \\ o^*(\mathrm{p}^{a(:4)}, \mathrm{p}^{b(:2)}) \oplus \mathrm{p}^{b(3)}, \quad otherwise \end{cases}$$

Cui, Sen, Weishen Pan, Changshui Zhang, and Fei Wang. "Towards model-agnostic post-hoc adjustment for balancing ranking fairness and algorithm utility." In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 207-217. 2021.

Better Trade-off between Ranking Disparity and Utility

- xOrder always achieves low disparities
- xOrder obtains better utility-fairness trade-offs than baselines



upper-left corner is preferred

Thank you!

Q & A

Weill Cornell Medicine