Interpretability, Fairness, and Data Scarcity in Machine Learning

Muhang Tian

Duke University

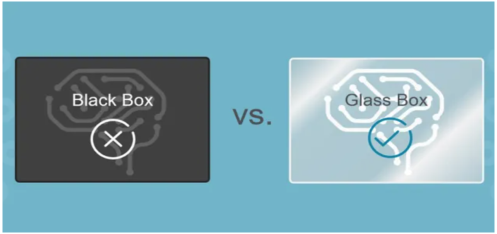
April 2024

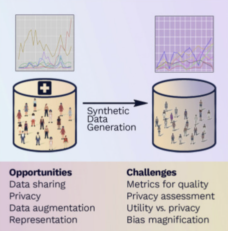
Outline

1. Introduction
2. Fairness in RL
3. Interpretable ML for Critical Care
4. Synthetic EHR Time Series Generation
5. Future Work
6. Q&A

Introduction

- **Field of Study**: Computer Science and Mathematics, minor in Economics
- **Research Interests**: broadly speaking, developing machine learning (ML) techniques to support human tasks. My past research has been focused on the following (in temporal order):
  - **Fairness and Equity**: fairness in reinforcement learning (RL).
  - **Interpretable ML**: risk scores for critical care medicine.
  - **Data Scarcity in Healthcare**: synthetic electronic health records (EHRs) generation.





Outline

1. Introduction
2. Fairness in RL
3. Interpretable ML for Critical Care
4. Synthetic EHR Time Series Generation
5. Future Work
6. Q&A

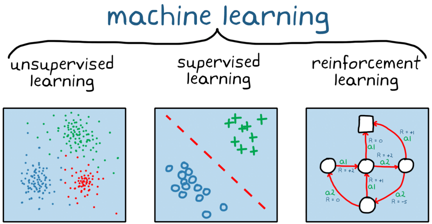
Fairness in RL

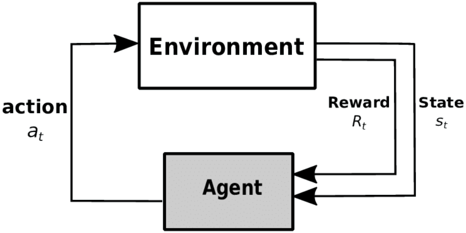
Motivation

When an RL agent's actions could affect multiple people, how can we enable it to produce a socially fair outcome so that people are treated equitably?

Example

Recommendation systems, clinical trials, and patient care.





Fairness in RL – Methodology

Formulation

Denote $\boldsymbol{G}(\tau) = \sum_{t=1}^{T} \gamma^{t-1} \boldsymbol{R}(s_t, a_t) \in \mathbb{R}^d$ as the long-term return for trajectory $\tau = \{(s_1, a_1), (s_2, a_2), ..., (s_T, a_T)\}$ and $W : \mathbb{R}^d \to \mathbb{R}$ be some nonlinear welfare function, where $\boldsymbol{R}(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is the reward function, γ is the discount factor, and d is the number of objectives (people). We aim to find an optimal fair policy π^* that maximizes the *expected welfare*:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[W(\boldsymbol{G}(\tau))\right] \tag{1}$$

- **Intuition**: Originally designed to rank societies, W allows us *scalarize* the return and incorporate fairness concepts defined by the specific function.
  - Examples: $W_{\text{Nash}}(\boldsymbol{G}(\tau)) = (\prod_{i=1}^{d} G(\tau)_i)^{1/d}$ and $W_{\text{egalitarian}}(\boldsymbol{G}(\tau)) = \min\{G(\tau)_i\}_{i=1}^{d}$.
- **Related Work**: [2, 26] focused on optimizing for the *welfare of expectation*, $\max_\pi W\left(\mathbb{E}_{\tau\sim\pi}[\boldsymbol{G}(\tau)]\right)$. This alternative objective could tolerate unfair outcomes within an individual trajectory τ.

Fairness in RL – Methodology

Challenge

Intractability: the proposed objective is difficult to optimize, specifically *APX-hard*, even in the tabular setting (such as in a grid world) due to the nonlinearity of W.

Solution

Proposed an approximate algorithm based on Q-learning [29] to optimize for *expected welfare*. The key components of the algorithm are:

- Nonlinear updates of the Q-table, where η is the learning rate:

$$\boldsymbol{Q}^{\pi}(s,a) \leftarrow \boldsymbol{Q}^{\pi}(s,a) + \eta[\boldsymbol{R}(s,a) + \gamma \boldsymbol{Q}^{\pi}(s',a^*) - \boldsymbol{Q}^{\pi}(s,a)], \tag{2}$$

$$a^* = \arg\max_{a} W(\gamma \boldsymbol{Q}^{\pi}(s',a)). \tag{3}$$

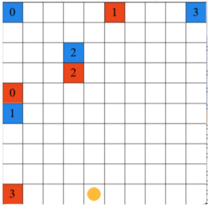
- Non-stationary policy that considers the past history, where $\boldsymbol{R}_{acc} = \sum_{k=1}^{t} \gamma^{k-1} \boldsymbol{R}(s_k, a_k)$:

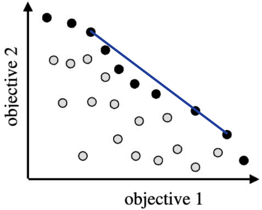
$$a = \arg\max_{a'} W(\boldsymbol{R}_{acc} + \gamma^t \boldsymbol{Q}^{\pi}(s,a')). \tag{4}$$

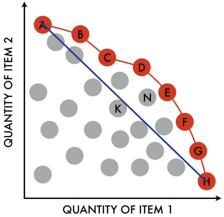
Fairness in RL – Results

Experimental

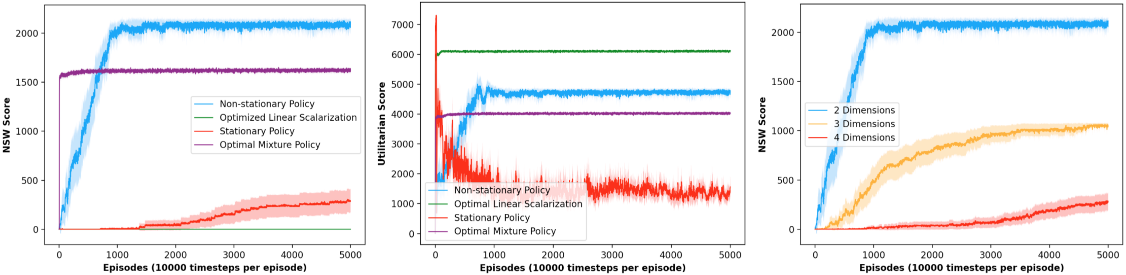
- Designed simulation environments for evaluations (taxi).
- Demonstrated the proposed approach outperforms baselines such as *linearly scalarized* [28], *stationary*, and *mixture policies* [27].
  - *Linearly scalarized*: optimize each objective with Q-learning, take action $a = \arg\max_{a'} \boldsymbol{w}^\top \boldsymbol{Q}(s, a')$ for each state s.
  - *Stationary*: our proposed method, without using $\boldsymbol{R}_{acc}$ for action selection.
  - *Mixture*: use the optimal policy for i^{th} objective for J time steps.







Fairness in RL – Results



(a) Comparisons (Nash welfare).

(b) Comparisons (utilitarian welfare).

(c) Effect of dimensionality.

Theoretical

- Maximizing $W_{\text{Nash}}(\boldsymbol{G}(\tau))$ is APX-hard, even in a deterministic environment. This is found by reducing the problem of allocating indivisible goods.
- The algorithm converges (*Banach's Fixed Point Theorem* [3]).

Outline

1. Introduction
2. Fairness in RL
3. Interpretable ML for Critical Care
4. Synthetic EHR Time Series Generation
5. Future Work
6. Q&A

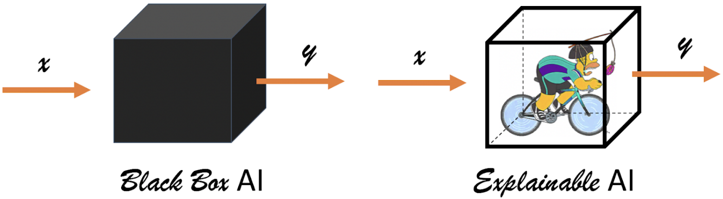
Interpretable ML for Critical Care

Motivation

When ML models are used for high-stakes decisions, trustworthiness is vital due to issues of accountability and transparency. An interpretable model could enable users to understand how model predictions are made.

Example

Applications of ML models in settings that greatly influence people. Mortality risk prediction is important for efficiency and quality of critical care.



Interpretable ML for Critical Care – Methodology

Formulation

Denote $\mathcal{D}/m = \{1/m, \boldsymbol{x}_i/m, y_i\}_{i=1}^{n}$ as a scaled dataset. The set of feature indices $\{1, ..., p\}$ is arbitrarily partitioned into Γ disjoint sets (groups), denoted as $\{G_k\}_{k=1}^{\Gamma}$. The objective is to solve sparse logistic regression with integer, sparsity, box, and group sparsity constraints:

$$\min_{\boldsymbol{w}, w_0, m} \mathcal{L}(\boldsymbol{w}, w_0, \mathcal{D}/m) = \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \frac{\boldsymbol{w}^\top \boldsymbol{x}_i + w_0}{m}\right)\right)$$

$$\text{s.t. } \|\boldsymbol{w}\|_0 \leq \lambda, \boldsymbol{w} \in \mathbb{Z}^p, w_0 \in \mathbb{Z} \quad \textit{\# at most } \lambda \textit{ integer coefficients} \tag{5}$$

$$w_j \in [a_j, b_j] \quad \forall j \in \{1, ..., p\} \quad \textit{\# control range of coefficients} \tag{6}$$

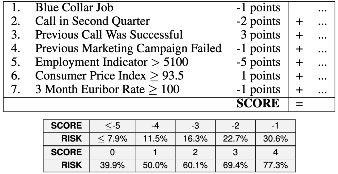
$$m > 0 \quad \textit{\# expand solution space using multiplier} \tag{7}$$

$$\sum_{k=1}^{\Gamma} \mathbb{I}\left\{\boldsymbol{w}_{G_k} \neq \boldsymbol{0}\right\} \leq \gamma. \quad \textit{\# at most } \gamma \textit{ groups, where } G_k \textit{ are the indices of group } k \tag{8}$$

Interpretable ML for Critical Care – Methodology

Intuition for predecessor – FasterRisk [19]

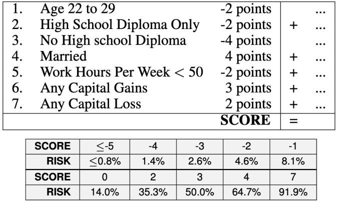
- **Integer constraint**: enables fast calculation of risk in practice, since adding up integers is straightforward.
- **Sparsity constraint**: allows users to understand the final model since the final solution $\boldsymbol{w}^*$ involves at most λ non-zero coefficients.
- **Box constraint**: controls the solution space and acts as regularization.
- **Multiplier** m: expands the solution space.



1.	Blue Collar Job	-1 points		...
2.	Call in Second Quarter	-2 points	+	...
3.	Previous Call Was Successful	3 points	+	...
4.	Previous Marketing Campaign Failed	-1 points	+	...
5.	Employment Indicator $>$ 5100	-5 points	+	...
6.	Consumer Price Index $\geq$ 93.5	1 points	+	...
7.	3 Month Euribor Rate $\geq$ 100	-1 points	+	...
		SCORE	=	

SCORE	$\leq$-5	-4	-3	-2	-1
RISK	$\leq$ 7.9%	11.5%	16.3%	22.7%	30.6%
SCORE	0	1	2	3	4
RISK	39.9%	50.0%	60.1%	69.4%	77.3%

(a) Predicting whether a person opens a bank account.



1.	Age 22 to 29	-2 points		...
2.	High School Diploma Only	-2 points	+	...
3.	No High school Diploma	-4 points		...
4.	Married	4 points	+	...
5.	Work Hours Per Week $<$ 50	-2 points	+	...
6.	Any Capital Gains	3 points	+	...
7.	Any Capital Loss	2 points	+	...
		SCORE	=	

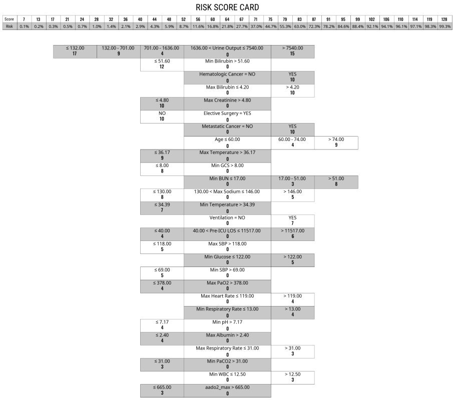
SCORE	$\leq$-5	-4	-3	-2	-1
RISK	$\leq$0.8%	1.4%	2.6%	4.6%	8.1%
SCORE	0	2	3	4	7
RISK	14.0%	35.3%	50.0%	64.7%	91.9%

(b) Predicting salary $>$50K.

Interpretable ML for Critical Care – Methodology

Challenges

Lack of cohesiveness: cannot control the number of group features in the final solution. This is problematic when the sparsity constraint λ is large.



RISK SCORE CARD

Score	7	13	17	21	24	28	32	36	40	44	48	52	56	60	64	67	71	75	79	83	87	91	95	99	102	106	110	114	119	128
Risk	0.1%	0.2%	0.3%	0.5%	0.7%	1.0%	1.4%	2.1%	2.9%	4.3%	5.9%	8.7%	11.6%	16.8%	21.8%	27.7%	37.0%	44.7%	55.3%	63.0%	72.3%	78.2%	84.6%	88.4%	92.1%	94.1%	96.1%	97.1%	98.3%	99.3%

≤ 132.00 17	132.00 - 701.00 9	701.00 - 1636.00 4	1636.00 < Urine Output ≤ 7540.00 0	> 7540.00 15	
		≤ 51.60 12	Min Bilirubin > 51.60 0		
			Hematologic Cancer = NO 0	YES 10	
			Max Bilirubin ≤ 4.20 0	> 4.20 10	
		≤ 4.80 10	Max Creatinine > 4.80 0		
		NO 10	Elective Surgery = YES 0		
			Metastatic Cancer = NO 0	YES 10	
			Age ≤ 60.00 0	60.00 - 74.00 4	> 74.00 9
		≤ 36.17 9	Max Temperature > 36.17 0		
		≤ 8.00 8	Min GCS > 8.00 0		
			Min BUN ≤ 17.00 0	17.00 - 51.00 3	> 51.00 8
		≤ 130.00 8	130.00 < Max Sodium ≤ 146.00 0	> 146.00 5	
		≤ 34.39 7	Min Temperature > 34.39 0		
			Ventilation = NO 0	YES 7	
		≤ 40.00 4	40.00 < Pre-ICU LOS ≤ 11517.00 0	> 11517.00 6	
		≤ 118.00 5	Max SBP > 118.00 0		
			Min Glucose ≤ 122.00 0	> 122.00 5	
		≤ 69.00 5	Min SBP > 69.00 0		
		≤ 378.00 4	Max PaO2 > 378.00 0		
			Max Heart Rate ≤ 119.00 0	> 119.00 4	
			Min Respiratory Rate ≤ 13.00 0	> 13.00 4	
		≤ 7.17 4	Min pH > 7.17 0		
		≤ 2.40 4	Max Albumin > 2.40 0		
			Max Respiratory Rate ≤ 31.00 0	> 31.00 3	
		≤ 31.00 3	Min PaCO2 > 31.00 0		
			Min WBC ≤ 12.50 0	> 12.50 3	
		≤ 665.00 3	aado2_max > 665.00 0		

Solution

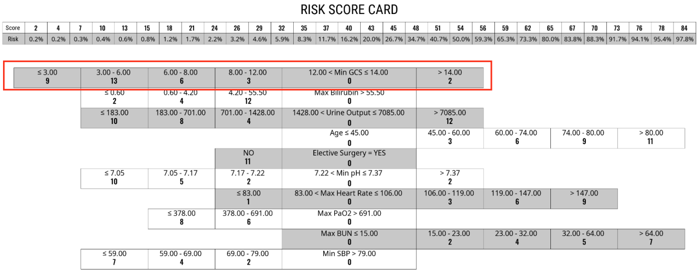
Allow users to define an arbitrary partition of the feature indices $\{1, ..., p\}$ as Γ groups, $\{G_k\}_{k=1}^{\Gamma}$. The user sets group sparsity constraint γ and controls the number of groups used in the final solution.

$$\sum_{k=1}^{\Gamma} \mathbb{I}\{\mathbf{w}_{G_k} \neq \mathbf{0}\} \leq \gamma$$

Interpretable ML for Critical Care – Methodology

Challenges

Domain knowledge: due to data noise, the final model could use counter-intuitive relationships between a variable and risk.



RISK SCORE CARD

Score	2	4	7	10	13	15	18	21	24	26	29	32	35	37	40	43	45	48	51	54	56	59	62	65	67	70	73	76	78	84
Risk	0.2%	0.2%	0.3%	0.4%	0.6%	0.8%	1.2%	1.7%	2.2%	3.2%	4.6%	5.9%	8.3%	11.7%	16.2%	20.0%	26.7%	34.7%	40.7%	50.0%	59.3%	65.3%	73.3%	80.0%	83.8%	88.3%	91.7%	94.1%	95.4%	97.8%

≤ 3.00 9	3.00 - 6.00 13	6.00 - 8.00 6	8.00 - 12.00 3	12.00 < Min GCS ≤ 14.00 0	> 14.00 2			
	≤ 0.60 2	0.60 - 4.20 4	4.20 - 55.50 12	Max Bilirubin > 55.50 0				
	≤ 183.00 10	183.00 - 701.00 8	701.00 - 1428.00 4	1428.00 < Urine Output ≤ 7085.00 0	> 7085.00 12			
				Age ≤ 45.00 0	45.00 - 60.00 3	60.00 - 74.00 6	74.00 - 80.00 9	> 80.00 11
			NO 11	Elective Surgery = YES 0				
	≤ 7.05 10	7.05 - 7.17 5	7.17 - 7.22 2	7.22 < Min pH ≤ 7.37 0	> 7.37 2			
			≤ 83.00 1	83.00 < Max Heart Rate ≤ 106.00 0	106.00 - 119.00 3	119.00 - 147.00 6	> 147.00 9	
		≤ 378.00 8	378.00 - 691.00 6	Max PaO2 > 691.00 0				
				Max BUN ≤ 15.00 0	15.00 - 23.00 2	23.00 - 32.00 4	32.00 - 64.00 5	> 64.00 7
	≤ 59.00 7	59.00 - 69.00 4	69.00 - 79.00 2	Min SBP > 79.00 0				

Figure: Counter-intuitive scorecard for Glasgow Coma Scale.

Solution

Allow users to define monotonicity constraints for each component function (row of the scorecard) so that the component function of interest obeys domain medical knowledge.

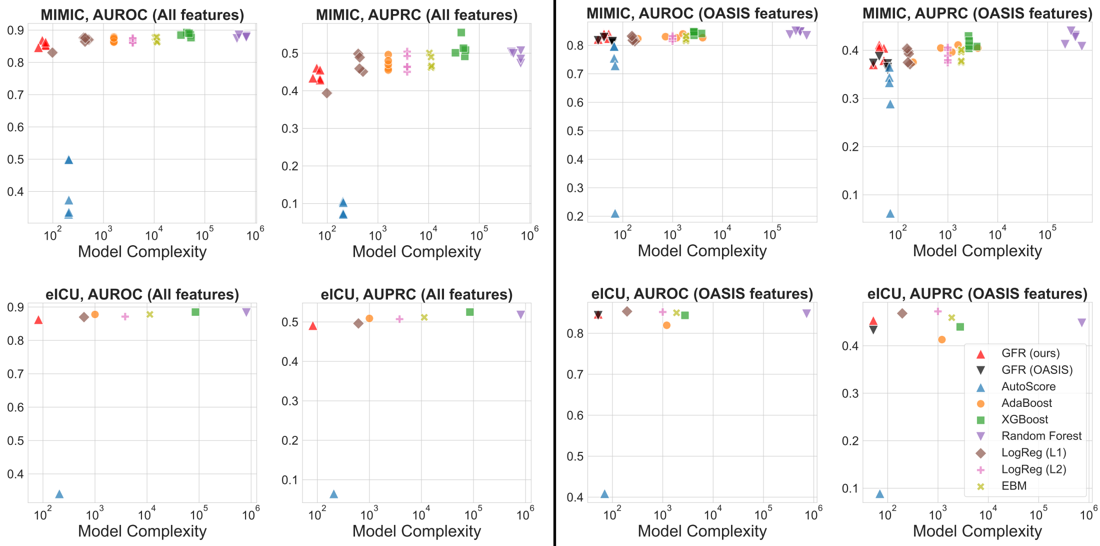
Interpretable ML for Critical Care – Results

- **Datasets**: MIMIC III [16] for internal evaluation, eICU [23] for out-of-distribution testing.
- **Risk Score Baselines**: OASIS [15], SAPS II [17], and APACHE IV/IVa [34].
- **ML Baselines**: Logistic Regression, Explainable Boosting Machine [20], Random Forest [4], AdaBoost [11], XGBoost [7], AutoScore [30], and OASIS+ [9].

		Sparse				Not Sparse		
		GFR-10	OASIS	GFR-15	SAPS II	GFR-40	APACHE IV	APACHE IVa
		$F = 10$	$F = 10$	$F = 15$	$F = 17$	$F = 40$	$F = 142$	$F = 142$
MIMIC III	AUROC	**0.813 ± 0.007**	0.775 ± 0.008	**0.836 ± 0.006**	0.795 ± 0.009	**0.858 ± 0.008**		
Test Folds	AUPRC	**0.368 ± 0.011**	0.314 ± 0.014	**0.403 ± 0.011**	0.342 ± 0.012	**0.443 ± 0.013**		
	HL χ^2	**16.28 ± 2.51**	146.16 ± 10.27	**26.73 ± 6.38**	691.45 ± 18.64	**35.78 ± 11.01**		
	SMR	**0.992 ± 0.022**	0.686 ± 0.008	**0.996 ± 0.015**	0.485 ± 0.005	**1.002 ± 0.017**		
	Sparsity	**42 ± 0**	47	**48 ± 4.9**	58	**66 ± 8.0**		
eICU	AUROC	**0.844**	0.805	**0.859**	0.844	0.864	0.871	**0.873**
Test Set	AUPRC	**0.437**	0.361	**0.476**	0.433	**0.495**	0.487	0.489
	Sparsity	**34**	47	**50**	58	**80**	≥142	≥142

Table: Comparison with baselines, where F is the number of features used.

Interpretable ML for Critical Care – Results



Interpretable ML for Critical Care – Results

Table: **Fairness and calibration across population subgroups in eICU.**

		Ethnicity (alphabetical order)						Gender	
		African American	Asian	Caucasian	Hispanic	Native American	Other/Unknown	Female	Male
Percentage (%)		11.17	1.49	76.91	3.86	0.68	4.68	45.08	54.90
AUROC (↑)	GFR-10	**0.829**	**0.833**	**0.837**	**0.856**	**0.881**	**0.849**	**0.835**	**0.840**
	OASIS	0.811	0.797	0.803	0.825	0.824	0.809	0.806	0.805
	GFR-15	**0.846**	**0.848**	**0.854**	**0.873**	**0.895**	**0.860**	**0.853**	**0.856**
	SAPS II	0.846	0.828	0.843	0.859	0.893	0.842	0.844	0.845
	GFR-40	0.859	0.861	0.859	0.881	0.902	0.873	0.857	0.865
	APACHE IV	0.873	0.858	0.869	0.890	**0.903**	0.884	0.867	0.875
	APACHE IVa	**0.875**	**0.866**	**0.870**	**0.893**	0.901	**0.886**	**0.869**	**0.876**
AUPRC (↑)	GFR-10	**0.415**	**0.390**	**0.422**	**0.480**	**0.558**	**0.418**	**0.418**	**0.429**
	OASIS	0.345	0.330	0.364	0.410	0.370	0.328	0.356	0.365
	GFR-15	**0.453**	**0.454**	**0.466**	**0.534**	**0.555**	**0.477**	**0.466**	**0.471**
	SAPS II	0.424	0.408	0.435	0.470	0.598	0.395	0.440	0.428
	GFR-40	**0.488**	**0.500**	**0.489**	**0.553**	**0.585**	**0.512**	**0.488**	**0.499**
	APACHE IV	0.488	0.467	0.484	0.536	0.536	0.479	0.478	0.493
	APACHE IVa	0.487	0.492	0.487	0.538	0.522	0.484	0.481	0.496
HL χ^2 (↓)	GFR-10	**27.90**	**11.00**	**113.70**	24.68	**5.48**	12.53	**58.65**	102.74
	OASIS	43.48	21.02	135.52	**5.23**	14.84	**11.75**	82.52	**79.11**
	GFR-15	**23.64**	**9.88**	**63.40**	**10.62**	**4.43**	**3.73**	**13.62**	**57.75**
	SAPS II	1070.09	94.34	6599.71	228.75	62.95	333.65	3575.48	4750.90
	GFR-40	**8.72**	**5.20**	**120.03**	**12.03**	**11.57**	**6.09**	**58.34**	**97.92**
	APACHE IV	308.51	34.51	1257.11	78.93	42.53	114.22	835.14	950.18
	APACHE IVa	167.60	13.04	502.27	42.78	23.21	62.48	372.68	384.89

Outline

1. Introduction
2. Fairness in RL
3. Interpretable ML for Critical Care
4. Synthetic EHR Time Series Generation
5. Future Work
6. Q&A

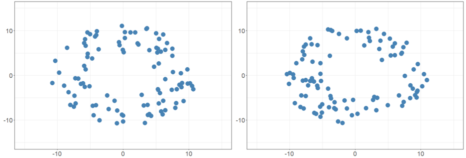
Synthetic EHR Time Series Generation

Overview

Motivation: due to the sensitive nature of EHRs, privacy concerns and confidentiality regulations pose major barriers to data access and sharing [1, 6].
Potential Solution: synthetic data generation can allow us to obtain a larger sample size while protecting privacy. This can be done with deep generative models, given their ability to generate realistic high-dimensional data [12, 31].

Example

Personal anecdote: accessing EHR at Duke University requires CITI training and IRB protocols.



Original data

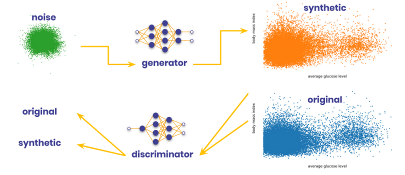
Synthetic data

The synthetic data retains the structure of the original data but is not the same

Synthetic EHR Time Series Generation

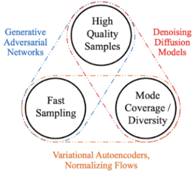
Related Work

- Generative adversarial networks (GANs): RCGAN [10], EHR-Safe [32], EHR-M-GAN [18], and medGAN [8].
- Diffusion models (DMs) for discrete variables such as international classification of diseases (ICD) codes: MedDiff [13], EHRDiff [33], ScoEHR [21], and TabDDPM [5].



Goal

- GANs could suffer from issues of training instability and mode collapse [25].
- EHR time series generation is relatively under-explored.
- Given the state-of-the-art performance of DMs on image generation tasks [14, 22, 24], is it possible to generate realistic EHR time series with diffusion models?


High Quality Samples
Generative Adversarial Networks
Denoising Diffusion Models
Fast Sampling
Mode Coverage / Diversity
Variational Autoencoders, Normalizing Flows

Synthetic EHR Time Series Generation – Methodology

Mixed diffusion with time-conditional bidirectional recurrent neural network (BRNN).

Mixed Diffusion

Denote numerical and discrete multivariate time series as $\boldsymbol{X} \in \mathbb{R}^{P_r \times L}$ and $\boldsymbol{C} \in \mathbb{Z}^{P_d \times L}$, respectively. L is the number of time steps, and P_r and P_d are the number of variables for numerical and discrete data types.

For $\boldsymbol{X}$, apply Gaussian diffusion, the forward process is:

$$q\left(\boldsymbol{X}^{(1:T)} | \boldsymbol{X}^{(0)}\right) = \prod_{t=1}^{T} \prod_{l=1}^{L} q\left(\boldsymbol{X}_{\cdot,l}^{(t)} | \boldsymbol{X}_{\cdot,l}^{(t-1)}\right), \tag{9}$$

where $q(\boldsymbol{X}_{\cdot,l}^{(t)} | \boldsymbol{X}_{\cdot,l}^{(t-1)}) = \mathcal{N}(\boldsymbol{X}_{\cdot,l}^{(t)}; \sqrt{1-\beta^{(t)}}\boldsymbol{X}_{\cdot,l}^{(t-1)}, \beta^{(t)}\boldsymbol{I})$ and $\boldsymbol{X}_{\cdot,l}$ is the l^{th} observation of the numerical time series.

Synthetic EHR Time Series Generation – Methodology

Mixed Diffusion (continued)

The reverse process is $p_\theta(\boldsymbol{X}^{(0:T)}) = p_\theta(\boldsymbol{X}^{(T)}) \prod_{t=1}^{T} p_\theta\big(\boldsymbol{X}^{(t-1)}|\boldsymbol{X}^{(t)}\big)$, and

$$p_\theta\big(\boldsymbol{X}^{(t-1)}|\boldsymbol{X}^{(t)}\big) := \mathcal{N}\big(\boldsymbol{X}^{(t-1)}; \boldsymbol{\mu}_\theta(\boldsymbol{X}^{(t)}, t), \tilde{\beta}^{(t)}\boldsymbol{I}\big),$$

$$\boldsymbol{\mu}_\theta(\boldsymbol{X}^{(t)}, t) = \frac{1}{\sqrt{\alpha^{(t)}}}\left(\boldsymbol{X}^{(t)} - \frac{\beta^{(t)}}{\sqrt{1-\bar{\alpha}^{(t)}}}\boldsymbol{s}_\theta(\boldsymbol{X}^{(t)}, t)\right), \quad \tilde{\beta}^{(t)} = \frac{1-\bar{\alpha}^{(t-1)}}{1-\bar{\alpha}^{(t)}}\beta^{(t)}, \tag{10}$$

where $\boldsymbol{s}_\theta$ is the BRNN. For $\boldsymbol{C}$, the forward process is:

$$q\big(\tilde{\boldsymbol{C}}^{(1:T)}|\tilde{\boldsymbol{C}}^{(0)}\big) = \prod_{t=1}^{T}\prod_{p=1}^{P_d}\prod_{l=1}^{L} q\big(\tilde{\boldsymbol{C}}_{p,l}^{(t)}|\tilde{\boldsymbol{C}}_{p,l}^{(t-1)}\big), \tag{11}$$

$$q\big(\tilde{\boldsymbol{C}}_{p,l}^{(t)}|\tilde{\boldsymbol{C}}_{p,l}^{(t-1)}\big) := \mathcal{C}\big(\tilde{\boldsymbol{C}}_{p,l}^{(t)}; (1-\beta^{(t)})\tilde{\boldsymbol{C}}_{p,l}^{(t-1)} + \beta^{(t)}/K\big), \tag{12}$$

where $\mathcal{C}$ is a categorical distribution, $\tilde{\boldsymbol{C}}_{p,l}^{(0)} \in \{0, 1\}^K$ is a one-hot encoding of $C_{p,l}$.

Synthetic EHR Time Series Generation – Methodology

Mixed Diffusion (continued)

The forward process posterior distribution is defined as follows:

$$q(\tilde{\boldsymbol{C}}_{p,l}^{(t-1)}|\tilde{\boldsymbol{C}}_{p,l}^{(t)}, \tilde{\boldsymbol{C}}_{p,l}^{(0)}) := \mathcal{C}\left(\tilde{\boldsymbol{C}}_{p,l}^{(t-1)}; \phi/\sum_{k=1}^{K}\phi_k\right), \tag{13}$$

$$\phi = \left(\alpha^{(t)}\tilde{\boldsymbol{C}}_{p,l}^{(t)} + (1-\alpha^{(t)})/K\right) \odot \left(\bar{\alpha}^{(t-1)}\tilde{\boldsymbol{C}}_{p,l}^{(0)} + (1-\bar{\alpha}^{(t-1)})/K\right). \tag{14}$$

The reverse process $p_\theta(\tilde{\boldsymbol{C}}_{p,l}^{(t-1)}|\tilde{\boldsymbol{C}}_{p,l}^{(t)})$ is parameterized as $q(\tilde{\boldsymbol{C}}_{p,l}^{(t-1)}|\tilde{\boldsymbol{C}}_{p,l}^{(t)}, \boldsymbol{s}_\theta(\tilde{\boldsymbol{C}}_{p,l}^{(t)}, t))$.
$\boldsymbol{s}_\theta$ is trained using both Gaussian and multinomial diffusion processes:

$$\mathcal{L}_{\mathcal{N}}(\theta) := \mathbb{E}_{\boldsymbol{X}^{(0)},\epsilon,t}\left[\left\|\epsilon - \boldsymbol{s}_\theta\left(\sqrt{\bar{\alpha}^{(t)}}\boldsymbol{X}^{(0)} + \sqrt{1-\bar{\alpha}^{(t)}}\epsilon, t\right)\right\|^2\right], \tag{15}$$

$$\mathcal{L}_{\mathcal{C}}(\theta) := \mathbb{E}_{p,l}\left[\sum_{t=2}^{T} D_{\mathrm{KL}}\left(q(\tilde{\boldsymbol{C}}_{p,l}^{(t-1)}|\tilde{\boldsymbol{C}}_{p,l}^{(t)}, \tilde{\boldsymbol{C}}_{p,l}^{(0)}) \,\middle\|\, p_\theta(\tilde{\boldsymbol{C}}_{p,l}^{(t-1)}|\tilde{\boldsymbol{C}}_{p,l}^{(t)})\right)\right]. \tag{16}$$

Synthetic EHR Time Series Generation – Results

Mixed Diffusion (continued)

The objective is to $\min_\theta \lambda \mathcal{L}_{\mathcal{C}}(\theta) + \mathcal{L}_{\mathcal{N}}(\theta)$, where λ is a hyperparameter.

Evaluation Metrics: discriminative/predictive scores, train on synthetic test on real (TSTR), nearest neighbor adversarial accuracy (NNAA), and membership inference risk (MIR).

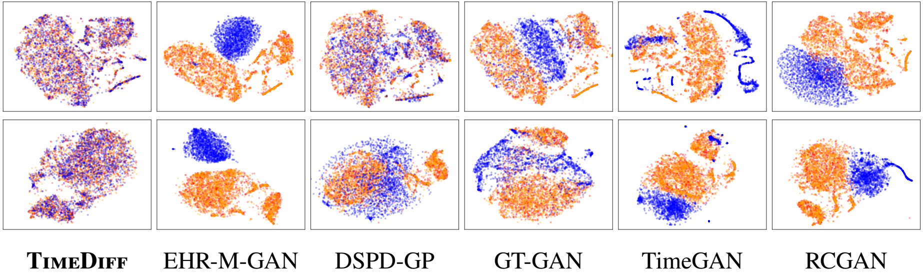


Figure: t-SNE for eICU (1st row) and MIMIC-IV (2rd row). Synthetic samples in **blue**, real training samples in **red**, and real testing samples in **orange**.

Synthetic EHR Time Series Generation – Results

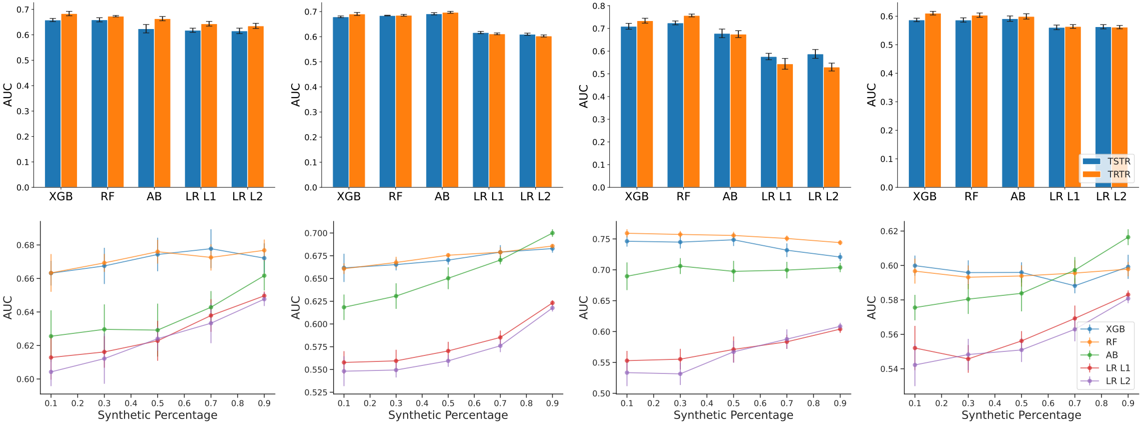
Metric	Method	Stocks	Energy	MIMIC-III	MIMIC-IV	HiRID	eICU
Discriminative Score (↓)	**TimeDiff**	**.048±.028**	**.088±.018**	**.028±.023**	**.030±.022**	**.333±.056**	**.015±.007**
	EHR-M-GAN	.483±.027	.497±.006	.499±.002	.499±.001	.496±.003	.488±.022
	DSPD-GP	.081±.034	.416±.016	.491±.002	.478±.020	.489±.004	.327±.020
	DSPD-OU	.098±.030	.290±.010	.456±.014	.444±.037	.481±.007	.367±.018
	CSPD-GP	.313±.061	.392±.007	.498±.001	.488±.010	.485±.007	.489±.010
	CSPD-OU	.283±.039	.384±.012	.494±.002	.479±.005	.489±.004	.479±.017
	GT-GAN	.077±.031	.221±.068	.488±.026	.472±.014	.455±.015	.448±.043
	TimeGAN	.102±.021	.236±.012	.473±.019	.452±.027	.498±.002	.434±.061
	RCGAN	.196±.027	.336±.017	.498±.001	.490±.003	.499±.001	.490±.023
	C-RNN-GAN	.399±.028	.499±.001	.500±.000	.499±.000	.499±.001	.493±.010
	T-Forcing	.226±.035	.483±.004	.499±.001	.497±.002	.480±.010	.479±.011
	P-Forcing	.257±.026	.412±.006	.494±.006	.498±.002	.494±.004	.367±.047
	Real Data	*.019±.016*	*.016±.006*	*.012±.006*	*.014±.011*	*.014±.015*	*.004±.003*
Predictive Score (↓)	**TimeDiff**	**.037±.000**	**.251±.000**	**.469±.003**	**.432±.002**	**.292±.018**	**.309±.019**
	EHR-M-GAN	.120±.047	.254±.001	.861±.072	.880±.079	.624±.028	.913±.179
	DSPD-GP	.038±.000	.260±.001	.509±.014	.586±.026	.404±.013	.320±.018
	DSPD-OU	.039±.000	.252±.000	.497±.006	.474±.023	.397±.024	.317±.023
	CSPD-GP	.041±.000	.257±.001	1.083±.002	.496±.034	.341±.029	.624±.066
	CSPD-OU	.044±.000	.253±.000	.566±.006	.516±.051	.439±.010	.382±.026
	GT-GAN	.040±.000	.312±.002	.584±.010	.517±.016	.386±.033	.487±.033
	TimeGAN	.038±.001	.273±.004	.727±.010	.548±.022	.729±.039	.367±.025
	RCGAN	.040±.001	.292±.005	.837±.040	.700±.014	.675±.074	.890±.017
	C-RNN-GAN	.038±.000	.483±.005	.933±.046	.811±.048	.727±.082	.769±.045
	T-Forcing	.038±.001	.315±.005	.840±.013	.641±.017	.364 ±.018	.547±.069
	P-Forcing	.043±.001	.303±.006	.683±.031	.557±.030	.445±.018	.345±.021
	Real Data	*.036±.001*	*.250±.003*	*.467±.005*	*.433±.001*	*.267±.012*	*.304±.017*

Synthetic EHR Time Series Generation – Results

Table: Privacy score evaluations.

Metric	Method	MIMIC-III	MIMIC-IV	HiRID	eICU
AA_{test} ($\sim$0.5)	**TimeDiff**	**.574±.002**	**.517±.002**	**.531±.003**	**.537±.001**
	EHR-M-GAN	.998±.000	1.000±.000	1.000±.000	.977±.000
	RCGAN	.983±.001	.999±.000	1.000±.000	1.000±.000
AA_{train} ($\sim$0.5)	**TimeDiff**	**.573±.002**	**.515±.002**	**.531±.002**	**.531±.002**
	EHR-M-GAN	.999±.000	1.000±.000	1.000±.000	.965±.002
	RCGAN	.984±.001	.999±.000	1.000±.000	1.000±.000
NNAA (↓)	**TimeDiff**	.002±.002	.002±.002	.004±.003	.006±.002
	EHR-M-GAN	.000±.000	.000±.000	.000±.000	.012±.003
	RCGAN	.001±.000	.000±.000	.000±.000	.000±.000
MIR (↓)	**TimeDiff**	.191±.008	.232±.048	.236±.179	.227±.021
	EHR-M-GAN	.025±.007	.435±.031	.459±.161	.049±.006
	RCGAN	.013±.002	.277±.049	.063±.013	.000±.000

Synthetic EHR Time Series Generation – Results



MIMIC-III eICU HiRID MIMIC-IV

Figure: (Top) TSTR/TRTR; (Bottom) TSRTR.

Outline

1. Introduction
2. Fairness in RL
3. Interpretable ML for Critical Care
4. Synthetic EHR Time Series Generation
5. Future Work
6. Q&A

Future Work

Fair RL

- Proposed algorithm, *Welfare Q-learning*, does not have strong convergence guarantees.
- Adapting to deep learning techniques for complex state space and environments (non-grid-world).

Interpretable ML

- Interpretability for knowledge discovery and verification, i.e., helping doctors to understand whether a diagnosis methodology is useful or not.
- Applications in supporting healthcare in real-world settings.

Synthetic EHR

- Adaptive diffusion model for class-aware generation, so that the trained model can be used to generate synthetic samples for different population.
- Privacy protection guarantees and interpretability of diffusion models.

Outline

1. Introduction
2. Fairness in RL
3. Interpretable ML for Critical Care
4. Synthetic EHR Time Series Generation
5. Future Work
6. Q&A

References I

[1] Karim Abouelmehdi et al. “Big data security and privacy in healthcare: A Review”. In: *Procedia Computer Science* 113 (2017), pp. 73–80.

[2] Mridul Agarwal, Vaneet Aggarwal, and Tian Lan. “Multi-objective reinforcement learning with non-linear scalarization”. In: *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 2022, pp. 9–17.

[3] Stefan Banach. “Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales”. In: *Fundamenta mathematicae* 3.1 (1922), pp. 133–181.

[4] Leo Breiman. “Random forests”. In: *Machine Learning* 45 (2001), pp. 5–32.

[5] Taha Ceritli et al. “Synthesizing Mixed-type Electronic Health Records using Diffusion Models”. In: *arXiv preprint arXiv:2302.14679* (2023).

[6] Richard J Chen et al. “Synthetic data in machine learning for medicine and healthcare”. In: *Nature Biomedical Engineering* 5.6 (2021), pp. 493–497.

[7] Tianqi Chen et al. “Xgboost: extreme gradient boosting”. In: *R package version 0.4-2* 1.4 (2015), pp. 1–4.

[8] E. Choi et al. “Generating Multi-label Discrete Patient Records using Generative Adversarial Networks”. In: *Proceedings of the 2nd Machine Learning for Healthcare Conference*. Vol. 68. 2017, pp. 286–305.

References II

[9] Yasser El-Manzalawy et al. "OASIS+: leveraging machine learning to improve the prognostic accuracy of OASIS severity score for predicting in-hospital mortality". In: *BMC Medical Informatics and Decision Making* 21.1 (2021), p. 156.

[10] Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. *Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs*. 2017. arXiv: `1706.02633 [stat.ML]`.

[11] Yoav Freund and Robert E Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139.

[12] Jie Gui et al. "A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications". In: *IEEE Transactions on Knowledge and Data Engineering* 35.4 (2023), pp. 3313–3332.

[13] Huan He et al. "MedDiff: Generating Electronic Health Records using Accelerated Denoising Diffusion Model". In: *arXiv preprint arXiv:2302.04355* (2023).

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

[15] Alistair EW Johnson, Andrew A Kramer, and Gari D Clifford. "A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy". In: *Critical Care Medicine* 41.7 (2013), pp. 1711–1718.

[16] Alistair EW Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* 3.1 (2016), pp. 1–9.

References III

[17] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study". In: *JAMA* 270.24 (1993), pp. 2957–2963.

[18] Jin Li et al. "Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications". In: *NPJ Digital Medicine* 6.1 (2023), p. 98.

[19] Jiachang Liu et al. "FasterRisk: fast and accurate interpretable risk scores". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 17760–17773.

[20] Yin Lou et al. "Accurate intelligible models with pairwise interactions". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2013, pp. 623–631.

[21] Ahmed Ammar Naseer et al. "ScoEHR: Generating Synthetic Electronic Health Records using Continuous-time Diffusion Models". In: (2023).

[22] Alex Nichol and Prafulla Dhariwal. "Improved Denoising Diffusion Probabilistic Models". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. 2021.

[23] Tom J Pollard et al. "The eICU Collaborative Research Database, a freely available multi-center database for critical care research". In: *Scientific Data* 5.1 (2018), pp. 1–13.

[24] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10684–10695.

[25] Divya Saxena and Jiannong Cao. "Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions". In: *ACM Computing Surveys* 54.3 (2021), p. 63.

References IV

[26] Umer Siddique, Paul Weng, and Matthieu Zimmer. "Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8905–8915.

[27] Peter Vamplew et al. "Constructing stochastic mixture policies for episodic multiobjective reinforcement learning tasks". In: *AI 2009: Advances in Artificial Intelligence: 22nd Australasian Joint Conference, Melbourne, Australia, December 1-4, 2009. Proceedings 22*. Springer. 2009, pp. 340–349.

[28] Kristof Van Moffaert, Madalina M Drugan, and Ann Nowé. "Scalarized multi-objective reinforcement learning: Novel design techniques". In: *2013 IEEE symposium on adaptive dynamic programming and reinforcement learning (ADPRL)*. IEEE. 2013, pp. 191–199.

[29] Christopher John Cornish Hellaby Watkins. "Learning from delayed rewards". In: (1989).

[30] Feng Xie et al. "AutoScore: a machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records". In: *JMIR Medical Informatics* 8.10 (2020), e21798.

[31] Xin Yi, Ekta Walia, and Paul S. Babyn. "Generative Adversarial Network in Medical Imaging: A Review". In: *Medical image analysis* 58 (2018), p. 101552.

[32] Jinsung Yoon et al. "EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records". In: *NPJ Digital Medicine* 6 (2023), p. 141.

[33] Hongyi Yuan, Songchi Zhou, and Sheng Yu. "EHRDiff: Exploring Realistic EHR Synthesis with Diffusion Models". In: *arXiv preprint arXiv:2303.05656* (2023).

References V

[34] Jack E Zimmerman et al. “Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today’s critically ill patients”. In: *Critical Care Medicine* 34.5 (2006), pp. 1297–1310.