Causal Transformer for Estimating Counterfactual Outcomes

Valentyn Melnychuk, Dennis Frauen, Stefan Feuerriegel

LMU Munich, Munich, Germany

November 22, 2024

Presented by Yuankang Zhao

Introduction: Estimating Counterfactual Outcomes Over Time

Problem Formulation:

- Observational dataset:
  - Time-varying covariates (e.g., blood pressure)
  - Static covariates (e.g., age)
  - Treatments (e.g., ventilation)
  - Factual outcomes (e.g., respiratory frequency)
- Goal: Estimate counterfactual outcomes over time starting from prediction origin for a given sequence of treatment interventions.

Why Important?

- Counterfactual prediction allows to answer individualized "what if" questions: what will happen to the patient, if I apply alternative sequence of treatments, counterfactual 1 to a standard treatment policy
- Overcomes limitations of RCTs. (High cost)
- Utilizes abundant observational data (e.g., EHRs).

Assumptions

Why Estimation is Hard?

- Counterfactual outcomes are not directly observed.
- Observed history grows over time.
- Time-varying confounding introduces bias.

Identifiability Assumptions:

- 1. Consistency. If $\bar{A}_t = \bar{a}_t$ is a given sequence of treatments for some patient, then $Y_{t+1}[\bar{a}_t] = Y_{t+1}$. This means that the potential outcome under treatment sequence $\bar{a}_t$ coincides for the patient with the observed (factual) outcome, conditional on $\bar{A}_t = \bar{a}_t$.
- 2. Sequential Overlap.
  There is always a non-zero probability of receiving/not receiving any treatment for all the history space over time:
  $$0 < \mathbb{P}(A_t = a_t \mid \bar{H}_t = \bar{h}_t) < 1, \quad \text{if } \mathbb{P}(\bar{H}_t = \bar{h}_t) > 0,$$
  where $\bar{h}_t$ is some realization of a patient history.
- 3. Sequential Ignorability The current treatment is independent of the potential outcome, conditioning on the observed history:
  $$A_t \perp Y_{t+1}[\bar{a}_t] \mid \bar{H}_t, \quad \forall \bar{a}_t.$$
  This implies that there are no unobserved confounders that affect both treatment and outcome.

Statistical Framework

IPTW

- $$e(z_1, X_0) = \Pr(Z_1 = z_1 \mid X_0)$$

  and

  $$e(z_2, Z_1, X_1, X_0) = \Pr(Z_2 = z_2 \mid Z_1, X_1, X_0)$$

  as the propensity scores at time points 1 and 2, respectively.

- $$E\{Y(z_1, z_2)\} = E\left\{ \frac{\mathbf{1}(Z_1 = z_1)\mathbf{1}(Z_2 = z_2)Y}{e(z_1, X_0)e(z_2, Z_1, X_1, X_0)} \right\}.$$

- This identification reveals the omitted overlap assumption:

$$0 < e(z_1, X_0) < 1, \quad 0 < e(z_2, Z_1, X_1, X_0) < 1.$$

Marginal structural model

- **MSM with baseline covariates**
  The mean of $Y(z_1, z_2)$ conditional on X_0 equals

$$E\{Y(z_1, z_2) \mid X_0\} = f(z_1, z_2, X_0; \beta).$$

  A leading example is

$$E\{Y(z_1, z_2) \mid X_0\} = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3^\top X_0.$$

- Parameter estimation
  If we observe all the potential outcomes, we can solve β from the following minimization problem:

$$\beta = \arg\min_b \sum_{z_2} \sum_{z_1} E\{Y(z_1, z_2) - f(z_1, z_2, X_0; b)\}^2.$$

- IPTW under MSM

$$\beta = \arg\min_b \sum_{z_2} \sum_{z_1} E\left[\frac{\mathbf{1}(Z_1 = z_1)\mathbf{1}(Z_2 = z_2)}{e(z_1, X_0)e(z_2, Z_1, X_1, X_0)}\{Y - f(z_1, z_2, X_0; b)\}^2\right].$$

G-computation

- G-formula

$$E\{Y(z_1, z_2)\} = \sum_{x_0} \sum_{x_1} E(Y \mid z_2, z_1, x_1, x_0) \Pr(x_1 \mid z_1, x_0) \Pr(x_0);$$

- Intuition

  Compare G-formula with the formula based on the law of total probability to gain more insights:

$$E(Y) = \sum_{x_0} \sum_{z_1} \sum_{x_1} \sum_{z_2} E(Y \mid z_2, z_1, x_1, x_0) \Pr(z_2 \mid z_1, x_1, x_0) \Pr(x_1 \mid z_1, x_0) \Pr(z_1 \mid x_0) \Pr(x_0).$$

  Erasing the probabilities of Z_2 and Z_1, we can obtain the G-formula. This is intuitive because the potential outcome $Y(z_1, z_2)$ has the meaning of fixing Z_1 and Z_2 at z_1 and z_2, respectively.

- Drawback: Under linear model, modeling misspecification may falsely reject the null hypothesis of zero causal effect of (Z_1 , Z_2) on Y even when the true effect is zero in the data-generating process. They called it the g-null paradox.

Related Work

Table 5. Overview of methods for estimating counterfactual outcomes over time.

Method	Setting	Model type (backbone)	Time	Treatments	Framework
HITR (Xu et al., 2016)	DGM (✗)	NP (GP)	Disc & Cont	Seq, Cat	G-computation
CGP (Schulam & Saria, 2017)	C, SO, SI, CSI (✗)	NP (GP)	Cont	Seq, Cat	G-computation
MOGP (Soleimani et al., 2017)	DGM (✗)	SP (GP)	Disc & Cont	Seq, Cont	G-computation
SyncTwin (Qian et al., 2021)	DGM (✗)	SP (GRU-D, LSTM)	Disc	Single-time, Bin	Synthetic control
DCRN (Berrevoets et al., 2021)	C, SO, Cov (✗)	P (3 LSTMs)	Disc	Seq, Bin	Disentangled representation
* MSMs (Robins et al., 2000)	C, SO, SI (✓)	P (Logistic & linear regressions)	Disc	Seq, Cat	IPTW weighted loss
* RMSNs (Lim et al., 2018)	C, SO, SI (✓)	P (LSTM)	Disc	Seq, Cat	IPTW weighted loss
* CRN (Bica et al., 2020)	C, SO, SI (✓)	P (LSTM)	Disc	Seq, Cat	BR (gradient reversal)
* G-Net (Li et al., 2021)	C, SO, SI (✓)	P (LSTM)	Disc	Seq, Cat	G-computation
* *Causal Transformer* (this paper)	C, SO, SI	P (3 transformers)	Disc	Seq, Cat	BR (CDC)

* = Methods with the same assumptions as ours (and thus included in our baselines)

Legend:

- Setting: consistency (C), sequential overlap (SO), sequential ignorability (SI), sequential ignorability but conditional on covariates (Cov), continuous sequential ignorability (CSI), assumed data generating model (DGM)
- Model: parametric (P), semi-parametric (SP), and non-parametric (NP)
- Time: discrete (Disc) or continuous (Cont) time steps
- Treatments: sequential (Seq), binary (Bin), categorical (Cat), continuous (Cont).
- Framework: inverse probability of treatment weights (IPTW), balanced representations (BR)

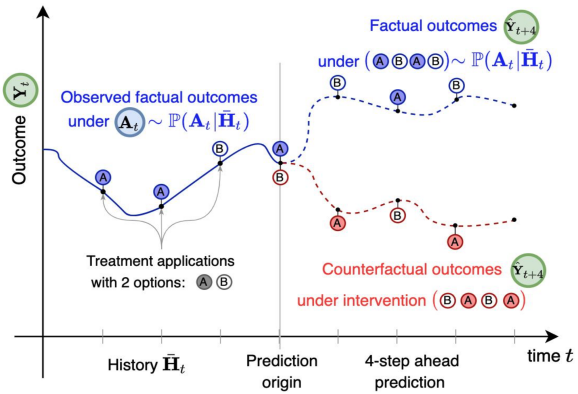
Problem formulation

- X_t: time-varying covariates (e.g., blood pressure)
- V: static covariates (e.g., age)
- A_t: treatments (e.g., ventilation)
- Y_t: (factual[2]) outcomes (e.g., respiratory frequency)
- The patient trajectory summarize by

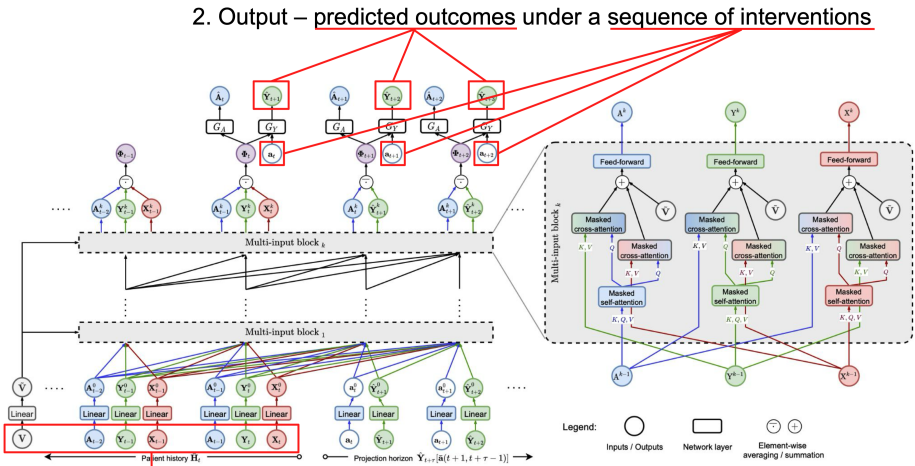
$$\bar{\mathbf{H}}_t = \{\bar{\mathbf{X}}_t, \bar{\mathbf{A}}_{t-1}, \bar{\mathbf{Y}}_t, \mathbf{V}\},$$

where $\bar{\mathbf{X}}_t = (\mathbf{X}_1, \ldots, \mathbf{X}_t)$, $\bar{\mathbf{Y}}_t = (\mathbf{Y}_1, \ldots, \mathbf{Y}_t)$, and $\bar{\mathbf{A}}_{t-1} = (\mathbf{A}_1, \ldots, \mathbf{A}_{t-1})$.

- We want to estimate **counterfactual outcomes over time** starting from prediction origin for a given sequence of treatment interventions.(ie. $\mathbb{E}\left(\mathbf{Y}_{t+\tau}\left[\bar{\mathbf{a}}_{t:t+\tau-1}\right] \mid \bar{\mathbf{H}}_t\right)$)



Model

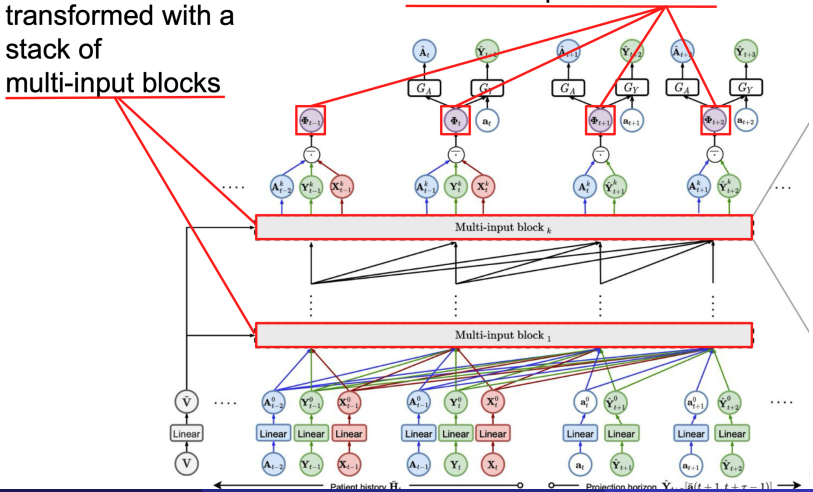


1. Input – observed patient history

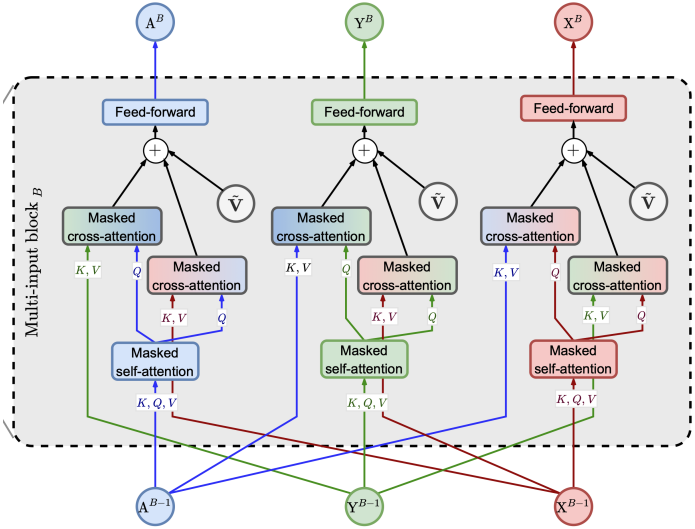
Model

3. Inputs are transformed with a stack of multi-input blocks

4. Outputs of the last block are averaged and form balanced representations

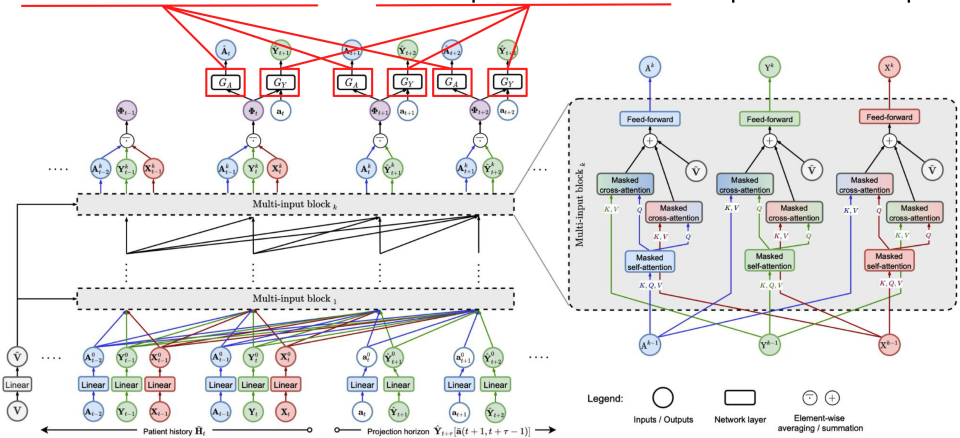


Transformer Blocks



Model

6. We place treatment classifier network and outcome prediction network on top of balanced representation



7. Both treatment classifier and outcome prediction networks are used for the novel counterfactual domain confusion loss (CDC) loss

Positional Encoding

Objective: Preserve order information in hidden states for better sequential modeling.

Setup:

- Use **relative positional encodings** to capture distances between positions, avoiding absolute location dependence.

Formulations:

- **Relative Position Encoding:**

$$a_{ij}^V = W^V_{clip(j-i,-l_{\max})}, \quad a_{ij}^K = W^K_{clip(j-i,-l_{\max})},$$

  where $\text{clip}(x, l_{\max}) = \max(-l_{\max}, \min(l_{\max}, x))$.

- **Attention with PE:**

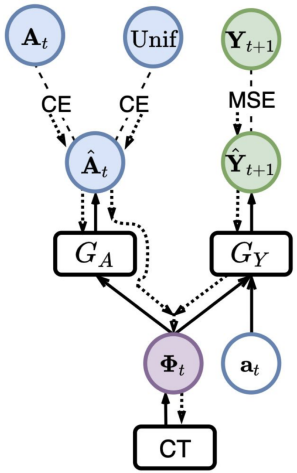
$$\text{Attn}(Q, K, V) = \sum_{j=1}^{t} \alpha_{ij}(V_j + a_{ij}^V),$$

$$\alpha_{ij} = \text{softmax}_j \left(\frac{Q_i^\top (K_j + a_{ij}^K)}{\sqrt{d_{qk}}} \right).$$

Training objective

Setup:

- Two networks:
  - Outcome prediction network G_Y: predicts next outcome Y_{t+1}.
  - Treatment classifier network G_A: classifies current treatment A_t.
- Representation $\Phi_t(\theta_R)$: shared between networks.
- θ_Y and θ_A denote the trainable parameters in G_Y and G_A, respectively. θ_R denotes all trainable parameters in CT for generating the representation $\mathbf{\Phi}_t$.
- Objectives:
  1. Predict the outcome accurately (**factual loss**).
  2. Ensure $\Phi_t(\theta_R)$ is non-predictive of A_t (**domain confusion via CDC loss**).



Training objective

1. Factual Outcome Loss:

$$\mathcal{L}_{G_Y}(\theta_Y, \theta_R) = \| Y_{t+1} - G_Y\left(\Phi_t(\theta_R), A_t; \theta_Y\right)\|^2 .$$

2. CDC Loss:

- Treatment classification loss:

$$\mathcal{L}_{G_A}(\theta_A, \theta_R) = -\sum_{j=1}^{d_a} \mathbb{1}[A_t = a_j] \log G_A\left(\Phi_t(\theta_R); \theta_A\right).$$

- Domain confusion loss:

$$\mathcal{L}_{\text{conf}}(\theta_A, \theta_R) = -\sum_{j=1}^{d_a} \frac{1}{d_a} \log G_A\left(\Phi_t(\theta_R); \theta_A\right).$$

3. Overall Adversarial Objective:

$$\hat{\theta}_Y, \hat{\theta}_R = \arg\min_{\theta_Y, \theta_R} \mathcal{L}_{G_Y}(\theta_Y, \theta_R) + \alpha \mathcal{L}_{\text{conf}}(\hat{\theta}_A, \theta_R),$$

$$\hat{\theta}_A = \arg\min_{\theta_A} \alpha \mathcal{L}_{G_A}(\theta_A, \hat{\theta}_R).$$

Experiment Dataset

Fully-Synthetic Data:

- Based on the pharmacokinetic-pharmacodynamic model for tumor growth
- Simulates effects of lung cancer treatments.
- Two settings: Single sliding treatment; Random trajectories with one or more treatments.
- Different levels of confounding (γ).

Semi-Synthetic Data:

- Created from real-world MIMIC-III dataset
- Generates patient trajectories under endogenous and exogenous dependencies.
- Includes treatment effects and confounding control.

Real-World Data:

- Uses MIMIC-III dataset with 25 vital signs and 3 static features.
- Focuses on diastolic blood pressure as the outcome.
- Considers two treatments:
  - Vasopressors.
  - Mechanical ventilation.
- Highlights confounding due to treatment interactions.

encoder-decoder causal transformer

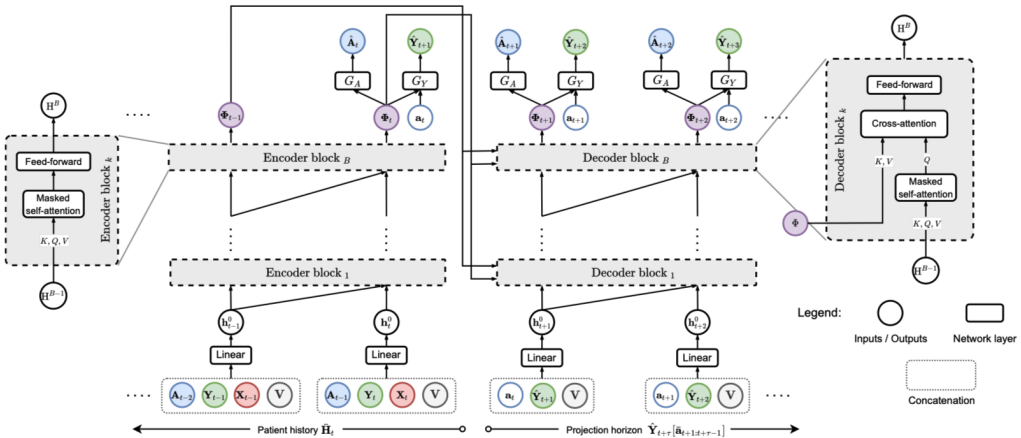
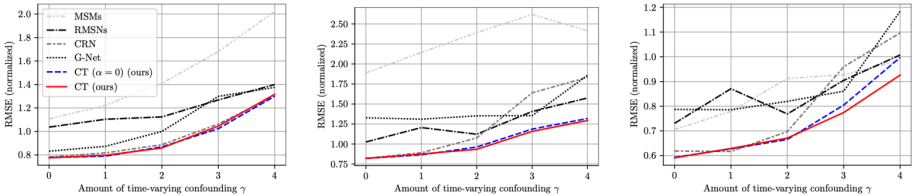


Figure 4. Architecture of the encoder-decoder causal transformer (EDCT). Residual connections with layer normalizations are omitted for clarity. The encoder is trained to perform one-step-ahead prediction $\hat{\mathbf{Y}}_{t+1}[\mathbf{a}_t]$, whereas the decoder uses the pretrained balanced representations of history from the encoder. Based on them, the decoders makes predictions for the projection horizon $\tau \geq 2$ via $\hat{\mathbf{Y}}_{t+\tau}[\bar{\mathbf{a}}_{t+1:t+\tau-1}]$.

Experiment result



(a) One-step-ahead prediction

(b) τ-step-ahead prediction (single sliding treatment).

(c) τ-step-ahead prediction (random trajectories)

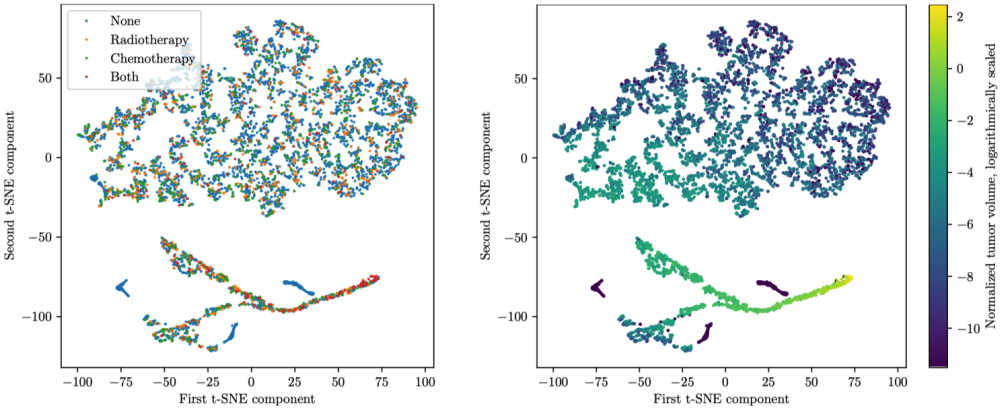
Figure 2. Results for fully-synthetic data based on tumor growth simulator (lower values are better). Shown is the mean performance averaged over five runs with different seeds. Here: $\tau = 6$.

Table 1. Results for semi-synthetic data for τ-step-ahead prediction based on real-world medical data (MIMIC-III). Shown: RMSE as mean $\pm$ standard deviation over five runs. Here: random trajectory setting. MSMs struggle for long prediction horizons with values $>$ 10.0 (due to linear modeling of IPTW scores).

	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$	$\tau = 6$	$\tau = 7$	$\tau = 8$	$\tau = 9$	$\tau = 10$
MSMs (Robins et al., 2000)	0.37 ± 0.01	0.57 ± 0.03	0.74 ± 0.06	0.88 ± 0.03	1.14 ± 0.10	1.95 ± 1.48	3.44 ± 4.57	> 10.0	> 10.0	> 10.0
RMSNs (Lim et al., 2018)	0.24 ± 0.01	0.47 ± 0.01	0.60 ± 0.01	0.70 ± 0.02	0.78 ± 0.04	0.84 ± 0.05	0.89 ± 0.06	0.94 ± 0.08	0.97 ± 0.09	1.00 ± 0.11
CRN (Bica et al., 2020)	0.30 ± 0.01	0.48 ± 0.02	0.59 ± 0.02	0.65 ± 0.02	0.68 ± 0.02	0.71 ± 0.01	0.72 ± 0.01	0.74 ± 0.01	0.76 ± 0.01	0.78 ± 0.02
G-Net (Li et al., 2021)	0.34 ± 0.01	0.67 ± 0.03	0.83 ± 0.04	0.94 ± 0.04	1.03 ± 0.05	1.10 ± 0.05	1.16 ± 0.05	1.21 ± 0.06	1.25 ± 0.06	1.29 ± 0.06
EDCT w/ GR ($\lambda = 1$) (*ours*)	0.29 ± 0.01	0.46 ± 0.01	0.56 ± 0.01	0.62 ± 0.01	0.67 ± 0.01	0.70 ± 0.01	0.72 ± 0.01	0.74 ± 0.01	0.76 ± 0.01	0.78 ± 0.01
CT ($\alpha = 0$) (*ours*) *	$\mathbf{0.20 \pm 0.01}$	$\mathbf{0.38 \pm 0.01}$	$\mathbf{0.45 \pm 0.01}$	0.50 ± 0.02	$\mathbf{0.52 \pm 0.02}$	0.55 ± 0.02	0.56 ± 0.02	0.58 ± 0.02	0.60 ± 0.02	0.61 ± 0.02
CT (*ours*)	$\mathbf{0.20 \pm 0.01}$	$\mathbf{0.38 \pm 0.01}$	$\mathbf{0.45 \pm 0.01}$	$\mathbf{0.49 \pm 0.01}$	$\mathbf{0.52 \pm 0.02}$	$\mathbf{0.53 \pm 0.02}$	$\mathbf{0.55 \pm 0.02}$	$\mathbf{0.56 \pm 0.02}$	$\mathbf{0.58 \pm 0.02}$	$\mathbf{0.59 \pm 0.02}$

Lower = better (best in bold)

Balance Representation



(a) t-SNE embeddings of balanced representations with indicated current treatments

(b) t-SNE embeddings of balanced representations with indicated next outcomes

Figure 5. t-SNE embeddings of the balanced representations of CT. We display $N = 100$ patients from the fully-synthetic data (tumor growth simulator). Here: representations of the validation set ($\gamma = 4$), where each patient trajectory contains 60 time steps, thus displaying 6,000 embeddings. Note the logarithmic scale for the outcomes (in color).

Result for MIMIC-III

Table 2: RMSE Results (Mean $\pm$ Standard Deviation)

Method	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
MSMs	6.37 ± 0.26	9.06 ± 0.41	11.89 ± 1.28	13.12 ± 1.25	14.44 ± 1.12
RMSNs	5.20 ± 0.15	9.79 ± 0.31	10.52 ± 0.39	11.09 ± 0.49	11.64 ± 0.62
CRN	4.84 ± 0.08	9.15 ± 0.16	9.81 ± 0.17	10.15 ± 0.19	10.40 ± 0.21
G-Net	5.13 ± 0.05	11.88 ± 0.20	12.91 ± 0.26	13.57 ± 0.30	14.08 ± 0.31
CT	$\mathbf{4.59 \pm 0.09}$	$\mathbf{8.99 \pm 0.21}$	$\mathbf{9.59 \pm 0.22}$	$\mathbf{9.91 \pm 0.26}$	$\mathbf{10.14 \pm 0.29}$

ablation study

Table 3. Ablation study for proposed CT (with CDC loss, $\alpha = 0.01$, $\beta = 0.99$). Reported: normalized RMSE of CT with relative changes.

		$\tau = 1$		$\tau = 6$	
		$\gamma = 1$	$\gamma = 4$	$\gamma = 1$	$\gamma = 4$
	CT (proposed)	0.80	1.32	0.63	0.93
a	w/ non-trainable PE*	±0.00	−0.02	+0.01	−0.03
a	w/ absolute PE*	+0.04	+0.16	+0.15	+1.00
a	w/o attentional dropout*	±0.00	+0.07	+0.00	+0.09
a	w/o cross-attention*	+0.03	+0.16	+0.06	+0.10
b	w/o EMA ($\beta = 0$)*	+0.03	+0.38	+0.03	+0.33
b	w/o balancing ($\alpha = 0$; $\beta = 0.99$)*	−0.01	−0.02	±0.00	+0.07
b	w/ GR ($\lambda = 1$)	+0.02	+0.17	+0.08	+0.33
c	EDCT w/ GR ($\lambda = 1$)	+0.16	+0.08	+0.05	+0.23
c	EDCT w/ DC ($\alpha = 0.01$; $\beta = 0.99$)	−0.03	+0.10	−0.03	+0.23

Lower = better;
Improvement over CT in green, worse performance in red
* Identical hyperparameters as proposed CT for comparability

Recommendation

- Good paper with all the details and codes of the implementation.
- Solid proof of balanced representation under CDC loss
- Some critical thoughts
  - seems overlook the overlap assumption
  - validation for balanced representation in real-world setting
  - challenge of ITE estimation