Dynamic Survival Transformers for Causal Inference with Electronic Health Records

Prayag Chatha
Department of Statistics
University of Michigan
pchatha@umich.edu

Yixin Wang
Department of Statistics
University of Michigan

Zhenke Wu
Department of Biostatistics
University of Michigan

Jeffrey Regier
Department of Statistics
University of Michigan

Abstract

In medicine, researchers often seek to infer the effects of a given treatment on patients' outcomes. However, the standard methods for causal survival analysis make simplistic assumptions about the data-generating process and cannot capture complex interactions among patient covariates. We introduce the Dynamic Survival Transformer (DynST), a deep survival model that trains on electronic health records (EHRs). Unlike previous transformers used in survival analysis, DynST can make use of time-varying information to predict evolving survival probabilities. We derive a semi-synthetic EHR dataset from MIMIC-III to show that DynST can accurately estimate the causal effect of a treatment intervention on restricted mean survival time (RMST). We demonstrate that DynST achieves better predictive and causal estimation than two alternative models.

1 Introduction

Medical practitioners are often interested in the effect of a treatment on a patient's survival time until an event of interest. For instance, if a patient is prescribed a certain antibiotic, how will that affect their risk of experiencing sepsis in the next 24 hours? The field of causal survival analysis is concerned with estimating treatment effects on time-to-event outcomes given incomplete (censored) data; classical techniques such as the Kaplan-Meier curves [1] and the Cox regression model [2] are extensively used despite their limitations. Kaplan-Meier curves are a descriptive tool that do not model individual survival trajectories, while the Cox model assumes proportionality of hazard functions, which may be unrealistic. Meanwhile, the rise of electronic health records (EHRs) has led to an abundance of multi-concept longitudinal data: a setting for *observational* causal inference, if randomized controlled trials prove impractical or unethical.

With this observational setting in mind, we propose the Dynamic Survival Transformer (DynST), a deep-learning survival model that estimates individual survival probabilities over time from EHR data. DynST is built on the Transformer [3], a recent neural network architecture that has achieved state-of-the-art results in sequence-to-sequence learning, particularly in NLP [4]. Transformers can flexibly model individual survival trajectories without making simplifying parametric assumptions about the data-generating process. Unlike previous survival transformers [5, 6, 7] DynST exploits both static and time-varying features to capture how a patient's event risk evolves over time. Several works have applied transformers to prediction problems in EHR data [8, 9, 10, 11], motivated by similarities between EHRs and text, but DynST is the first transformer used to estimate the average effect of a treatment intervention on survival outcomes. Using a semi-synthetic dataset derived from

Accepted to the NeurIPS 2022 Workshop on Learning from Time Series for Health.

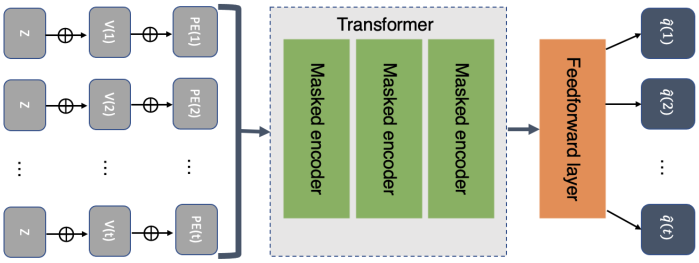

Figure 1: A diagram of DynST modeling the hazard function from a single patient's EHR data.

MIMIC-III [12], we show that DynST can improve on baseline methods in survival time prediction and causal inference.

2 Problem setup

We observe survival data taking the form $(X_i, O_i, \delta_i)_{i=1}^n$, where X_i represents the i-th patient's features, O_i is the observed (and possibly censored) time to the event, and δ_i is a binary variable indicating whether the event was observed or not, due to censoring. If $\delta_i = 0$, the i-th patient is right-censored, so the event takes place *after* O_i. Let T_i represent the uncensored survival time and let C_i be the censoring time. Then, $O_i = \min\{T_i, C_i\}$, and $\delta_i = \mathbf{1}(T_i \leq C_i)$. In this paper, we assume conditionally independent censoring, i.e., $T_i \perp\!\!\!\perp C_i \mid X_i$. We also assume a discrete survival setup, where $T_i \in \{1, 2, \ldots, ...t_{\max}\}$ and time steps are evenly spaced. The *hazard function*

$$h(t \mid X) = P_X(T = t \mid T \geq t) \tag{1}$$

is the risk of failure at time t given that the patient has survived thus far. The *survival probability* S at time t is

$$S(t \mid X) = P_X(T > t) = \prod_{\tau=1}^{t}(1 - h(\tau \mid X)). \tag{2}$$

The expected survival time is defined as

$$\mathbb{E}[T \mid X] = \sum_{t=1}^{t_{\max}} S(t \mid X). \tag{3}$$

Lastly, given a cutoff time τ, the restricted mean survival time (RMST) [13, 14] is defined as

$$Y_\tau = \mathbb{E}_X[\min\{T, \tau\}] = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{t=1}^{\tau} S(t \mid X_i)\right). \tag{4}$$

RMST can be thought of as the expected survival time up to time τ, averaged over the population of all patients.

3 Methods

3.1 Model architecture

Let X denote the features of a single patient; we suppress the patient index for readability. X consists of p static features $Z_1, \ldots Z_p$, collectively denoted as Z, and q time-varying features, $V_1, \ldots V_q$, collectively denoted as V. Here each feature V_j is a time series vector $(V_j^{(1)}, V_j^{(2)}, \ldots, V_j^{(t_{\max})})$. Static variables may include initial diagnoses, whereas a sequence of lab measurements is an example of a time-varying feature. Let $\overline{V^{(t)}} = (V_j^{(1)}, \cdots V_j^{(t)})$. At each time step t, the Dynamic Survival Transformer models $q(t; Z, \overline{V^{(t)}}) = 1 - h(t \mid Z, \overline{V^{(t)}})$. That is, DynST predicts the complement of each patient's hazard function using static features and the available history of time-varying features.

Figure 1 illustrates DynST's architecture. The model transforms each patient's medical records through the following procedure:

1. **Input embeddings**: A linear layer transforms time-varying input features V into a $t_{\max} \times d_{model}$ matrix W_V. Another linear transformation on static input features Z yields a length d_{model} vector W_Z. We add W_Z to each row of W_V, obtaining an embedded input sequence $W = (W_1, \ldots, W_{t_{\max}})$.
2. **Positional encodings** (P_E): Following the approach in [3], a deterministic sinusoidal function encodes time steps $1, \ldots, t_{\max}$ as vectors of length d_{model}, $P_E(1), \ldots, P_E(t_{\max})$. We add these encodings to the corresponding rows of W.
3. **Autoregressive Transformer Encoder**: W is fed through m transformer encoder layers, which combine multihead self-attention (eight heads) with a feedforward network and dropout. We apply autoregressive masking so that the t-th position of W cannot attend to (i.e., cannot learn from) future positions. We end up with the transformed output sequence $\phi(W) = (\phi(W_1), \ldots, \phi(W_{t_{\max}}))$.
4. **Hazard function output**: Following [5], we apply a two-layer feedforward network with a final sigmoid layer to obtain a vector of probabilities $\hat{q}(1), \ldots, \hat{q}(t_{\max})$. These estimate the complement of the hazard function.

Let $\hat{q}_i(t)$ denote the estimated hazard function for the i-th patient. From the predicted hazards, we estimate patient survival probabilities as $\hat{S}_i(t) = \prod_{\tau}^{t} \hat{q}(t; Z_i, \overline{V_i^{(t)}})$ and expected survival time as $\hat{T}_i = \sum_{t=1}^{t_{\max}} \hat{S}_i(t)$. We note that our prediction $\hat{S}_i(t)$ has no access to future covariates occurring after time step t.

3.2 Training

Our model jointly minimizes two objective functions. The first, like the loss functions in [15, 5], is an adaption of cross-entropy loss to censored data:

$$\mathcal{L}_1^{(i)} = -\left[\sum_{t=1}^{O_i-1} \log \hat{S}_i(t) + \sum_{t=O_i}^{t_{\max}} \log(1 - \hat{S}_i(t))\right] \cdot \delta_i - \left[\sum_{t=1}^{O_i} \log \hat{S}(t)\right] \cdot (1 - \delta_i). \tag{5}$$

For censored patients ($\delta_i = 0$), this means maximizing survival probabilities over the period of observation. For uncensored patients, survival probabilities up to the failure time are maximized, and subsequent survival probabilities are minimized. The second loss penalizes the error of the predicted survival times:

$$\mathcal{L}_2^{(i)} = \left|O_i - \hat{T}_i\right| \cdot \delta_i + \max\{0, O_i - \hat{T}_i\} \cdot (1 - \delta_i). \tag{6}$$

That is, when failure time is observed, the absolute error of $\hat{T}_i$ adds to the loss. For censored observations, loss is incurred only when the estimated survival time is before the time of censoring. The losses are combined and summed over all patients:

$$\mathcal{L} = \sum_{i=1}^{n} (1 - \alpha)\mathcal{L}_1^{(i)} + \alpha\mathcal{L}_2^{(i)}, \tag{7}$$

with α a tuning hyperparameter. We used the Adam optimizer with weight decay and trained DynST in minibatches on an NVIDIA GTX 2080 Ti GPU.

4 Experiments

4.1 Semi-synthetic dataset

We derived a semi-synthetic longitudinal survival dataset from MIMIC-III, a freely accessible database of de-identified EHRs from intensive care stays [12, 16]. After discarding records for stays shorter than 16 hours in length, we obtained demographic information, diagnoses, and hourly lab results for 30,323 patients. For a synthetic outcome event simulating in-hospital infection, we generated patient survival trajectories based on a subset of static and time-varying features as well as a predetermined treatment effect. After truncating patient histories longer than 128 hours, 39% of all patients had censored outcomes. For further details of the data simulation process, see Appendix A.

Table 1: Performance of DynST and baselines in predicting patient survival times (mean ± SD)

Model	Mean Abs. Error
Cox Oracle	16.04 ± 0.25
Static ST	11.42 ± 0.23
DynST (ours)	$\mathbf{11.19 \pm 0.24}$

Table 2: Comparison of estimators of average treatment effect on RMST (average bias ± SD)

Method	$\tau = 8$	$\tau = 12$	$\tau = 16$
Unadjusted Difference in Means	-0.502 ± 0	-1.11 ± 0	-1.88 ± 0
Cox Regression	-0.260 ± 0.0010	-0.523 ± 0.0012	-0.750 ± 0.0080
Logistic IPW	0.257 ± 0	0.35 ± 0	0.416 ± 0
DynST Regression	-0.146 ± 0.041	-0.160 ± 0.078	-0.190 ± 0.12
AIPW (DynST + Logistic)	$\mathbf{-0.0390 \pm 0.013}$	$\mathbf{0.0131 \pm 0.056}$	$\mathbf{0.127 \pm 0.082}$

4.2 Predictive survival analysis

To highlight DynST's advantages as a component in causal inference, we demonstrate its improved performance in individualized survival prediction compared to two baselines: a Cox Oracle model and a survival transformer using only static features (Static ST). We compare how well these models estimate patient survival times in terms of mean absolute error (MAE). Table 1 reports model performances on held-out test data, averaged over six trials. We used a 70/15/15 random split into training, validation and test sets; hyperparameter spaces and other training details are given in Appendix C. Both transformer-based models achieve a lower MAE than the Cox model, which assumes a misspecified model of the hazards. DynST shows a small but significant improvement over its static counterpart, indicating that DynST can capture the effect of time-varying features on patients' survival probabilities.

4.3 Causal survival analysis

We treat restricted mean survival time (RMST) as the outcome variable. Given a binary treatment variable A, our causal estimand of interest is the *average treatment effect* (ATE) on RMST,

$$\psi = \mathbb{E}[Y_\tau(1) - Y_\tau(0)]. \tag{8}$$

ATE is the expected difference between potential outcomes under treatment and control. See Appendix B for a review of the potential outcomes framework for causal inference.

In Table 2, we compare several methods of estimating of ATE in terms of bias, averaged over six trials. We compared (1) the unadjusted difference in means between treated and control groups, (2) a Cox model outcome regression (OR), (3) a inverse-propensity weighted (IPW) estimator using a logistic model of propensity scores, (4) OR using DynST, and (5) an augmented IPW (AIPW) estimator combining (3) and (4). For details of how these models were fitted, see Appendix D. The two methods that used DynST to model patient outcomes achieved the best estimation of ATE. This suggests that DynST can successfully adjust for confounding variables. The Cox model and Logistic IPW performed worse; they assume misspecified models of response and treatment.

5 Discussion

DynST can model patient survival under complex, time-varying interactions among variables. Combining DynST's flexible outcome modeling with knowledge of the treatment assignment mechanism can result in unbiased estimation of treatment effects. Cross-fitting [17] may help with selecting optimal hyperparameters for causal inference. As future work, we can apply DynST to quantify the real-world causal effects of treatments such as COVID-19 vaccines, for which there is ample EHR data. We can also generalize the model to handle the effects of time-varying treatments and outcomes. Lastly, we can experiment with further tailoring the training procedure of DynST toward causal inference by jointly modeling outcome response and treatment assignment [18].

References

[1] Jugal Kishore, Manish Kumar Goel, and Pardeep Khanna. “Understanding Survival Analysis: Kaplan-Meier Estimate.” In: *International Journal of Ayurveda Research* 1.4 (2010), p. 274.

[2] D. R. Cox. “Regression Models and Life-Tables.” In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (Jan. 1972), pp. 187–202.

[3] Ashish Vaswani et al. “Attention is All you Need.” In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[4] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. URL: `https://arxiv.org/abs/1810.04805`.

[5] Shi Hu et al. “Transformer-Based Deep Survival Analysis.” In: *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*. Vol. 146. Proceedings of Machine Learning Research. PMLR, Mar. 2021, pp. 132–148.

[6] Yun Zhao et al. *BERTSurv: BERT-Based Survival Models for Predicting Outcomes of Trauma Patients*. 2021. URL: `https://arxiv.org/abs/2103.10928`.

[7] Zifeng Wang and Jimeng Sun. “SurvTRACE: Transformers for Survival Analysis with Competing Events.” In: *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, Aug. 2022.

[8] Yikuan Li et al. “BEHRT: Transformer for Electronic Health Records.” In: *Scientific Reports* 10.1 (Apr. 2020).

[9] Laila Rasmy et al. “Med-BERT: Pretrained Contextualized Embeddings on Large-scale Structured Electronic Health Records for Disease Prediction.” In: *NPJ Digital Medicine* 4.1 (May 2021).

[10] Chao Pang et al. “CEHR-BERT: Incorporating Temporal Information from Structured EHR Data to Improve Prediction Tasks.” In: *Proceedings of Machine Learning for Health*. Vol. 158. Proceedings of Machine Learning Research. PMLR, Dec. 2021, pp. 239–260.

[11] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. “Causal Transformer for Estimating Counterfactual Outcomes.” In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, July 2022, pp. 15293–15329.

[12] A. Johnson, T. Pollard, and R. Mark. *MIMIC-III Clinical Database (version 1.4)*. 2019.

[13] Patrick Royston and Mahesh KB Parmar. “Restricted Mean Survival Time: an Alternative to the Hazard Ratio for the Design and Analysis of Randomized Trials with a Time-to-event Outcome.” In: *BMC Medical Research Methodology* 13.1 (Dec. 2013).

[14] Lili Zhao. “Deep Neural Networks for Predicting Restricted Mean Survival Times.” In: *Bioinformatics* 36.24 (Dec. 2020), pp. 5672–5677.

[15] Eleonora Giunchiglia, Anton Nemchenko, and Mihaela van der Schaar. “RNN-SURV: A Deep Recurrent Model for Survival Analysis.” In: *Artificial Neural Networks and Machine Learning – ICANN 2018*. 2018, pp. 23–32.

[16] Alistair E W Johnson et al. “MIMIC-III, a Freely Accessible Critical Care Database.” en. In: *Sci. Data* 3.1 (May 2016), p. 160035.

[17] Victor Chernozhukov et al. “Double/Debiased Machine Learning for Treatment and Structural Parameters.” In: *The Econometrics Journal* 21.1 (Jan. 2018), pp. C1–C68.

[18] Claudia Shi, David Blei, and Victor Veitch. “Adapting Neural Networks for the Estimation of Treatment Effects.” In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[19] Shirly Wang et al. “MIMIC-Extract.” In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. ACM, Apr. 2020.

[20] Peter C. Austin. “Statistical Power to Detect Violation of the Proportional Hazards Assumption When Using the Cox Regression Model.” In: *Journal of Statistical Computation and Simulation* 88.3 (Nov. 2017), pp. 533–552.

[21] M.A. Hernan and J.M. Robins. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. Taylor & Francis, 2023. ISBN: 9781420076165.

[22] Adam N. Glynn and Kevin M. Quinn. “An Introduction to the Augmented Inverse Propensity Weighted Estimator.” In: *Political Analysis* 18.1 (2010), pp. 36–56.

A Semi-synthetic EHR dataset

To preprocess the raw Medical Information Mart for Intensive Care (MIMIC-III) database, we follow the MIMIC-Extract pipeline [19]. Since time-varying features are reported in hourly intervals, in our experiments, each time step corresponds to one hour. We discard patients who have fewer than 16 hours of data and truncate all patient records longer than 128 hours in length. In addition, we scale and center all continuous variables to have mean zero and a standard deviation of one. As estimates of ground-truth causal effects are generally unverifiable, we build a semi-synthetic dataset in which the treatment assignment and outcome response are simulated from real observed features.

First, we develop an assignment mechanism for the synthetic treatment variable A. We assume that this treatment represents an intervention that occurs at the very start of a patient's observed history. We define an indicator variable Z_* denoting whether a patient is "severely ill." We consider three conditions, *hypertension*, *coronary atherosclerosis*, and *atrial fibrillation*. If a patient has been diagnosed with more than one of these illnesses, then $Z_* = 1$; otherwise, $Z_* = 0$. The propensity score is then

$$P(A = 1 \mid X) \equiv \pi(X_i) = \begin{cases} 0.8, & \text{if } Z_* = 1 \\ 0.2, & \text{if } Z_* = 0. \end{cases} \tag{9}$$

Every patient has a chance of being in either the treatment or the control arm, though this chance depends on information that will also influence their time-to-event outcomes.

Next, we simulate patient survival trajectories. Rather than directly modeling a patient's survival time, we model the hazard function at each hourly time step. This way, we ensure that survival probabilities adapt dynamically with patient medical histories. To generate the hazards, we use a subset of the available features: five static and four time-varying. One is the synthetic treatment; the rest were selected because they correlate with treatment. Thus we can induce a confounding structure between treatment and outcome. The synthetic hazard function is defined as follows:

$$h(t) = \underset{(1)}{H_0 \exp(-\lambda t)} \cdot \underset{(2)}{\exp(\theta A)} \cdot \underset{(3)}{\exp(\sum_{j=1}^{4} \beta_j Z_j)} \cdot \underset{(4)}{\exp(\log(1.02) t Z_*)} \cdot \underset{(5)}{\exp(\sum_{j=1}^{4} \gamma_j g(V_j^{(t)}))}. \tag{10}$$

1. **Baseline hazard**: We model the baseline hazard function with exponential decay, setting $H_0 = 0.001$ and $\lambda = 0.25$. All else being equal, a patient's risk decreases the longer they survive (a.k.a. the Lindy effect).
2. **Treatment effect**: A is the binary treatment variable, and $\theta = -0.5$ parameterizes the treatment effect. A treatment intervention will decrease a patient's lifetime risk.
3. **Static variables**: We include the linear effects of four binary, static variables $Z_1, \ldots, Z_4$ along the lines of a Cox regression model. The first variable is whether a patient is male; the remaining three are the presence of ICD-9 codes for hypertension, coronary atherosclerosis, and atrial fibrillation. We generated the coefficients β_j from a Uniform$(0.7, 1.2)$ distribution.
4. **Temporal interaction term**: Z_* is the indicator variable for severely ill patients (those with two or more of the above diagnoses present). This variable has a temporal interaction, thus violating the proportional hazards assumption [20]. When Z_* is 1, the hazard increases by two percent at every time step, modeling a severely ill patient's deteriorating health.
5. **Dynamic variables**: Our time-varying features $V_1, \ldots, V_4$ correspond to vital readings for hematocrit, hemoglobin, platelets, and mean blood pressure. Coefficients γ_j are generated from a Uniform$(0.1, 0.3)$ distribution. We set

   $$g(V_j^{(}t)) = \begin{cases} 0, & \text{if } V_j^{(t)} \geq 0 \\ \max\{(V_j^{(t)})^2, 3\}, & \text{else.} \end{cases} \tag{11}$$

   Thus, whenever V_j drops below the average (i.e., zero) value, it makes a quadratic contribution to the log hazard. We used clipping to avoid inflated hazards.

We constrain hazards to lie in the range $(10e^{-8}, 0.1)$ for reasons of stability. At each time step t, we calculate the survival probability to be $S(t) = \prod_{\tau=1}^{t}(1 - h(t))$. To introduce a degree of unexplained randomness to the system, we add Gaussian noise ($\sigma = 0.5$) to the logits of the survival probabilities. Lastly, we simulate patients' time-to-event trajectories with the modeled survival probabilities. At each time step $t = 1, \dots, t_{\max}$, we use a patient's survival probability to generate a Bernoulli random variable. The first time step that results in a zero-value we take to be a patient's survival time. If no zeros were generated at any t, then the patient is right-censored. In the resulting dataset, 39% of patients have right-censored survival times. The average time to censoring or failure is 27.8 hours.

To derive the "true" ATE on RMST, we create two copies of the MIMIC dataset, one where every patient has $A_i = 1$ and the other where every patient has $A_i = 0$. Let $Y_{i,\tau}(A_i, X_i)$ denote the simulated RMST at cutoff τ for patient i. We calculate true ATE as

$$\psi \approx \frac{1}{n} \sum_{i=1}^{n} \left(Y_{i,\tau}(1, X_i) - Y_{i,\tau}(0, X_i) \right). \tag{12}$$

Since $n = 30323$, this should be an accurate approximation of the population ATE.

B Causal inference

We follow the potential outcomes (Neyman-Rubin) framework for causal inference. We observe data of the form (X_i, A_i, Y_i) for subjects $i = 1, \dots, n$, where X_i denotes subject covariates, A_i denotes a binary treatment indicator, and Y_i is an outcome of interest. The *potential outcome* of Y_i under an intervention $A_i = a$ (i.e., assigning a subject to patient or control) is written as $Y_i(a)$. The causal relationship between A and Y is confounded if $\mathbb{E}[Y(a)] \neq \mathbb{E}[Y \mid A = a]$. That is, certain confounding variables predispose subjects who are observed in one treatment arm toward certain outcomes. This does occur in randomized controlled trials (RCTs), but in observational causal inference, the goal is to adjust for confounding variables to identify the true cause-and-effect relationship between A and Y. We may identify causal effects in an observational study if the following assumptions hold.

1. **Consistency**: The potential outcome for a patient assigned intervention $A = a$ is the same as the outcome for a patient observed "in the wild" with treatment $A = a$. That is, $A_i = a \implies Y_i(a) = Y_i$.
2. **Positivity**: All subjects have a non-zero chance of being in the treatment or control groups. Let $\pi(X_i) = P(A_i = 1 \mid X_i)$ be the *propensity score* for the i-th subject. We then assume that $\pi(X_i) \in (0, 1)$ for all i.
3. **Exchangeability**: Also referred to as ignorability or no unobserved confounders, we assume that potential outcomes are independent of observed treatment status when conditioning on the set of covariates X. Formally, we assume that $Y(a) \perp\!\!\!\perp A \mid X$. For subjects with comparable features, an observational study emulates an RCT.

Under these assumptions, $E[Y(a) \mid X] = E[Y \mid A = a, X]$, allowing us to infer potential outcomes from observed data. Still, a key difficulty is that only one of two potential outcomes is observed per subject. (If a patient was treated, we do not observe their potential outcome under control assignment.) The other is a *counterfactual* outcome. The potential outcomes framework is akin to reformulating causal inference as a missing data problem. In this paper, we assume all of the above assumptions. In non-synthetic datasets, only the second assumption is generally testable. For a textbook-level treatment of the potential outcomes framework, we refer the reader to [21].

A common causal estimand is *average treatment effect*, defined as

$$\psi = \mathbb{E}_X[Y(1) - Y(0)], \tag{13}$$

for some outcome variable Y. This quantifies the average difference in outcomes between two hypothetical populations, one in which all subjects were assigned treatment and another in which everyone was made a control. In general, due to confounding,

$$E[Y \mid X = 1] - E[Y \mid X = 0] \neq \psi. \tag{14}$$

We refer to the observed difference in outcomes between treated and control groups as the *unadjusted difference in means*.

Two common approaches to estimating ATE are inverse propensity weighing (IPW) and outcome regression. An IPW estimator involves fitting a model (e.g., a logistic regression) that estimates the treatment probabilities $\pi(X)$ and using these propensity scores to create a pseudo-population in which all individuals are equally likely to receive treatment or not. Outcome regression (a.k.a. standardization, g-computation) requires estimating the conditional expectation $E[Y \mid A, X]$ to determine the effect of A on Y while adjusting for potential confounders X. Synthesizing these methods, the Augmented IPW (AIPW) estimator fits nuisance models for both treatment assignment and outcome response. The AIPW has the desirable double robustness property, whereby it gives an unbiased estimate of ATE if either of its components are unbiased. Furthermore, even when one or both of the nuisance models have a relatively slow rate of convergence (e.g., neural networks), under certain conditions, the AIPW can achieve the linear semiparametric efficiency bound [17]. For an overview of techniques for estimating ATE and double robustness, see [22].

C Prediction experiment

We compare how well DynST, a static survival transformer (i.e., one with no access to time-varying features), and an oracle Cox model (i.e, one that uses only the features relevant to the true hazard function) make predictions of patients' expected survival times. We measure this performance in terms of mean absolute error. We define our MAE as follows:

$$\mathcal{C}_{\text{MAE}} = \frac{1}{n} \sum_{i=1}^{n} \left[\left| O_i - \hat{T}_i \right| \cdot \delta_i + \max\{0, O_i - \hat{T}_i\} \cdot (1 - \delta_i) \right] . \tag{15}$$

Similar to the loss function in equation (4), this metric accounts for patients with observed and censored survival times alike. When a patient's time-to-event is observed, the absolute difference between estimated and true survival times contributes to error. When a patient's survival is censored, an estimated survival time that falls short of the censoring time counts as error.

We used the following hyperparameter space to train DynST and the static transformer:

- Latent dimension $d_{model} \in \{32, 48, 64\}$
- Number of transformer blocks $m \in 2, 3, 4$
- Batch size $b \in \{16, 32\}$
- Joint loss ratio $\alpha \in \{0, 0.1, 0.2\}$
- Number of epochs $k_{epoch} \in \{1, 2, 3, 4, 5\}$

We used MAE as the criterion for early stopping and fixed the dropout proportion at 0.1. Using a 70/15/15 ratio to randomly split the data into training, validation, and test sets, we selected optimal hyperparameters using the validation set and reported MAE for the held-out test data.

We used Python's `lifelines` package[1] for an implementation of the Cox proportional hazards model and tuned over a range of L1 and L2 penalties, using the same training/validation/test splits. While `lifelines` has the basic implementation of a time-varying Cox model, it lacks the capacity to predict expected survival times on held-out data, being primarily used to report hazard ratios at different follow-up times.

D Causal inference experiment

We used an 80/20 training/validation split to tune DynST and the Cox model as outcome regression models, selecting for hyperparameters that produced the best validation MAE. We used the same hyperparameter spaces as in the prediction experiment. To fit the IPW estimator, we used the `sklearn` implementation of a cross-validated logistic regression with L2 penalty, fitted over 5 random folds. (Note that the standard deviation for the IPW estimator in Table 2 is zero; this is because a single model was validated over five random splits of the data.) This logistic model estimated propensity score as a function of three binary variables, indicating the presence of the diagnoses relevant to both treatment assignment and outcome (namely, hypertension, coronary atherosclerosis, and atrial

[1] `https://lifelines.readthedocs.io/en/latest/`

Table 3: True ATEs versus unadjusted differences in means for RMST at varying cutoffs τ

τ	Average Treatment Effect	Unadjusted Difference
8	0.265	−0.237
12	0.572	−0.539
16	0.946	−0.933

fibrillation). Counterintuitively, an IPW estimator benefits from an underlying propensity model that uses a “minimal” set of confounding features that makes conservative estimates of treatment assignment (i.e., not too close to zero or one) [22]. Using additional features to estimate propensity score did not improve ATE estimation.

We designed the treatment and outcome simulations so that the data exhibited a strong confounding structure: while the true effect of treatment is to decrease the hazards at all time steps, treatment assignment is strongly correlated with covariates that predict shorter survival times. Table 3 compares ATE (the true effect) and the unadjusted difference in means (the apparent effect) on RMST for three cutoff times. At each cutoff, treatment causes an increase in average survival time but correlates with a decrease in average survival time.