Uncovering Causal Relation Shifts in Event Sequences under Out-of-Domain Interventions

Kazi Tasnim Zinat[1][0000−0001−7914−5955], Yun Zhou[2], Xiang Lyu[2], Yawei Wang[2], Zhicheng Liu[1][0000−0002−1015−2759], and Panpan Xu[2]

[1] University of Maryland, College Park MD 20740, USA {kzintas}@umd.edu
[2] Amazon Web Services {yunzzhou,xupanpan}@amazon.com

Abstract. Inferring causal relationships between event pairs in a temporal sequence is applicable in many domains such as healthcare, manufacturing, and transportation. Most existing work on causal inference primarily focuses on event types within the designated domain, without considering the impact of exogenous out-of-domain interventions. In real-world settings, these out-of-domain interventions can significantly alter causal dynamics. To address this gap, we propose a new causal framework to define average treatment effect (ATE), beyond independent and identically distributed (i.i.d.) data in classic Rubin's causal framework, to capture the causal relation shift between events of temporal process under out-of-domain intervention. We design an unbiased ATE estimator, and devise a Transformer-based neural network model to handle both long-range temporal dependencies and local patterns while integrating out-of-domain intervention information into process modeling. Extensive experiments on both simulated and real-world datasets demonstrate that our method outperforms baselines in ATE estimation and goodness-of-fit under out-of-domain-augmented point processes. The supplementary materials are available here.

Keywords: Sequential Data · Causal Estimation · Transformer

1 Introduction

Multivariate event sequences are prevalent across diverse domains, capturing time-stamped data from shared environments or subjects, such as Electronic Health Records [21] and industrial maintenance records [26]. Real-world data generation is often implicitly conditioned on unobserved out-of-domain variables, leading to generalization failures due to distribution shifts [11, 25].

Causal relation analysis within event sequences aims to quantify how changes in one event influence another [7, 16]. The Rubin Causal Model [9] characterizes causal relations through the average treatment effect (ATE), representing the average difference between treatment & control potential outcomes. However, existing work focuses on causal relations within designated domains, overlooking the influence of known exogenous out-of-domain interventions that can induce causal relationship shifts. For instance, while meal ingestion increases blood glucose, insulin injection (an out-of-domain intervention) can counteract this effect [3].

To address this gap, we propose a novel approach to detect causal relation shifts in temporal processes under out-of-domain interventions. Our framework extends ATE to account for unique intervention scenarios, enabling comparison across such contexts. To our knowledge, this is the first work to explicitly model out-of-domain intervention impact on causal relations in temporal event data.

Our main contributions are multi-fold. First, we develop a theoretical framework of ATE that moves beyond i.i.d. assumptions to explicitly condition on intervention states and temporal dependencies. Second, we develop a propensity score-based treatment effect estimator to mitigate confounder bias. We justify its estimation consistency under our new ATE framework. Third, we propose a new Transformer architecture that captures intervention-induced temporal pattern changes. The architecture consists of several new network components: 1) a novel out-of-domain intervention embedding mechanism that enables direct influence modeling on event sequences; b) a weighted combination module that adaptively balances intervention and event embeddings; 3) a hybrid Transformer-CNN structure to simultaneously capture global dependencies and local temporal patterns specific to intervention effects, 4) a multi-objective loss function that jointly optimizes for intensity estimation and event type prediction. Fourth, we provide comprehensive experimental validation on simulated and real-world datasets, demonstrating improved performance in both causal effect estimation and temporal process representation compared to existing methods.

2 Related Work

Point processes [4] model labeled event sequences, with Hawkes processes [8] capturing excitation/inhibition effects of past events on current occurrences. Neural approaches include RNNs [6, 22] for handling time and labels simultaneously, and Transformers [23, 27] for long-range dependencies [18].

Causal analysis compares potential treatment and control outcomes using the Rubin Causal Model [17], with ATE as the primary metric. Propensity scores mitigate confounding bias in observational studies [9], while deep learning advances causal estimation [15, 19]. Temporal causal inference extends these concepts through Granger causality [24] and point process-based causal modeling [7, 16]. A detailed discussion with other methods is provided in the supplementary materials

3 Out-of-Domain Intervention Augmented Causal Inference

3.1 Notation

Let us consider a set of n individual sequences $\{\mathbf{s}_1, , \ldots, \mathbf{s}_n\}$. In a generic event sequence scenario, for a single sequence $\mathbf{s}_k$, the observation we collect for the i-th event can be expressed as a tuple $(e_{k,i}, t_{k,i})$. Here, $t_{k,i}$ represents the timestamp of the i-th event in sequence $\mathbf{s}_k$, and $e_{k,i}$ denotes the corresponding event type

which belongs to a set of events $\mathbb{E}$. The count of observed events of sequence $\mathbf{s}_k$ is denoted by L_k. The time duration of each sequence is T. We categorize the events based on their types: cause events c, outcome events o, out-of-domain interventions v, and other measured events $\mathbf{x}$.

3.2 Temporal Point Process

We focus on temporal point processes, where both event type and timestamp are observed. Hawkes processes, known for their self-excitation property, are commonly used to model real-world event sequences [14, 23]. Specifically, for an n-dimensional Hawkes process, the conditional intensity function (CIF) of outcome event o_i at time t is expressed as:

$$\lambda(t) = \mu(t) + \sum_{k=1}^{n} \sum_{t_{k,i}<t} \phi(t - t_{k,i}), \tag{1}$$

where μ denotes the baseline intensity function, ϕ is the excitation function capturing the influence of past events on outcome event. Note that $\lambda(t)$ is not observable. In this work, we estimate it from sequence data.

3.3 Out-of-Domain Intervention Augmented Causal Framework

We extend the Rubin causal framework to incorporate out-of-domain interventions, leveraging two key concepts.

First, we adopt the notion of process independence from graphical models [5], to establish direct causal relationships in multivariate point processes. For event variables (x, y, z), x is process independent of y given z if CIF of x is not functionally dependent on the history of y given the history of z. Consequently, a set of events $\mathbb{X}$ is considered a direct cause of event y if y is process independent of all other events given $\mathbb{X}$. We exclude the trivial case where y is constant over $\mathbb{X}$.

Second, we introduce the concept of proximal history, a time-based simplification widely used in point process literature [2]. Proximal history posits that cause events only within a recent time window influences the outcome event, allowing earlier history to be disregarded.

Combining these notions, we establish a causal framework and define ATE under out-of-domain interventions, drawing inspiration from [7] while adapting the theoretical assumptions for unbiasedness to accommodate out-of-domain interventions. We further incorporate these interventions into the propensity score-adjusted estimator, modifying the score definition.

Definition 1. *For an event tuple (c, o, v), binary cause variable c_t^w at time t indicates whether cause event c has occurred at least once in the time window $[t-w, t)$. Similarly, binary out-of-domain intervention variable v_t^w indicates whether out-of-domain intervention v occurred within the same window. The potential outcome variable $\lambda^{(c_t^w, v_t^w)}(t)$ denotes the CIF of outcome event o at time t, given the values of c_t^w and v_t^w. The binary vector $\mathbf{x}_t^w$ captures the occurrence of all*

other observed events in the time window $[t-w, t)$*. The ATE of cause* c_t^w *on outcome* $\lambda(t)$ *under out-of-domain intervention* v_t^w *is defined as:*

$$\tau(v_t^w) = \mathbb{E}\left[\frac{1}{T}\int_0^T \{\lambda^{(1,v_t^w)}(t) - \lambda^{(0,v_t^w)}(t)\}dt\right].$$

Unlike standard causal settings, $\lambda(t)$ *is latent and estimated from event type and timestamp observations.*

Cause and out-of-domain interventions may exhibit multiple occurrences prior to outcome within the time window. To focus on detecting causal relation shifts under interventions, we simplify the framework to study the treatment effect of occurrence, rather than count, as an initial step. Our real-world datasets validate this binary simplification. We defer continuous and ordinal cause and intervention modeling to future work. While not explicitly defined, process independence is crucial for linking the new ATE definition to causal relation shifts.

Theorem 1. *If event c is the direct cause of event o when out-of-domain intervention variable* $v_t^w = 1$*, and not if* $v_t^w = 0$*, then* $\tau(1) \neq 0$ *and* $\tau(0) = 0$*.*

Theorem 1 demonstrates that our new ATE effectively characterizes causal relations under out-of-domain interventions; the proof is provided in Appendix A of the supplementary materials.

3.4 Treatment Effect Estimation

In point process causal studies, cause event occurrences are typically not randomized, leading to potential imbalances between treatment groups. To address this, we propose a propensity score, adapted for our out-of-domain intervention augmented ATE framework, building on the established use of propensity scores for confounder adjustment in classical observational studies.

Definition 2. *The propensity score at time t for out-of-domain intervention variable* v_t^w *is*

$$e_t^w(v) = \mathbb{P}\{c_t^w = 1, v_t^w = v | \mathbf{x}_t^w\}, v \in \{0, 1\}.$$

Our novel propensity score, $e_t^w(v)$, incorporates both cause and out-of-domain intervention variables, conditioned on the proximal history of all other observed events. By treating binary out-of-domain interventions as alternative treatment assignments, rather than effect heterogeneity factors, we simplify the theoretical framework for estimation consistency.

We demonstrate that, with minor modifications to standard observational study assumptions to accommodate point process data and out-of-domain interventions, $e_t^w(v)$ effectively mitigates confounding bias, yielding an unbiased effect estimator.

Assumption 1 (SUTVA): For each assignment pair (c_t^w, v_t^w) and any t, there is only a single version of population outcome $\lambda^{(c_t^w, v_t^w)}(t)$, and the time window receives the assignment will not affect the outcome of other time windows.

The assumption is fundamental in causal inference. It ensures we can leverage

observations under treatment assignment to infer potential outcomes.

Assumption 2 (unconfoundedness): For any t, we have

$$\left\{\lambda^{(0,0)}(t), \lambda^{(0,1)}(t), \lambda^{(1,0)}(t), \lambda^{(1,1)}(t)\right\} \perp\!\!\!\perp \left(c_t^w, v_t^w\right) | \mathbf{x}_t^w.$$

This assumption guarantees measured covariates sufficiently balance treatment groups and adjust for confounding bias across different out-of-domain interventions. While untestable, it is commonly postulated in observational studies.

Assumption 3 (overlap): There exists a constant ϵ such that $\epsilon < e_t^w(v) < 1 - \epsilon, \forall t, \forall v \in \{0, 1\}$.

This assumption ensures sufficient observations across treatment groups under varying out-of-domain interventions for accurate estimation, and is verifiable in practice. Our real-world datasets satisfy this condition.

To simplify theoretical justification, we treat out-of-domain interventions as a second binary treatment variable, consistent with our framework's focus on the occurrence of cause and intervention events. This approach avoids the need for complex assumptions regarding potential outcomes and the impact of out-of-domain interventions on cause events, which would be overly intricate for our current framework and real-data applications.

We leverage propensity score $e_t^w(v)$ to construct weight

$$\alpha_t^w(v) = \frac{\mathbb{1}\{c_t^w=1, v_t^w=v\}}{e_t^w(v)} - \frac{\mathbb{1}\{c_t^w=0, v_t^w=v\}}{1-e_t^w(v)}.$$

We weight the outcome $\lambda(t)$ to derive the inverse probability weighting (IPW) estimator of ATE,

$$\hat{\tau}(v) = \mathbb{E}\left[\frac{1}{T}\int_0^T \alpha_t^w(v)\lambda(t)dt\right]. \tag{2}$$

We show that the IPW estimator is unbiased.

Theorem 2. *Under Assumption 1-3, we have* $\mathbb{E}[\hat{\tau}(v)] = \tau(v), \forall v \in \{0, 1\}$. *The propensity score can be estimated by event duration ratio,*

$$\hat{e}_t^w(v) = \frac{\sum_{k=1}^n \sum_{i=1}^{L_k} \int_{t_{i-1}}^{t_i} \mathbb{1}\{c_t^w=1, v_i^w=v, \mathbf{x}_t^w\} dt}{\sum_{k=1}^n \sum_{i=1}^{L_k} \int_{t_{i-1}}^{t_i} \mathbb{1}\{\mathbf{x}_t^w\} dt}. \tag{3}$$

We obtain the effect estimate by plugging $\hat{e}_t^w(v)$ into the weight $\alpha_t^w(v)$ in eq. (2). The outcome $\lambda(t)$ is estimated using a Transformer-based model on sequence data; see supplementary materials for the proof.

4 Transformer-based Process Model

We propose a neural network model for learning temporal point processes and estimating the conditional intensity function (CIF) $\lambda(t)$ of outcome events under both cause and out-of-domain interventions. A key innovation is the integration of out-of-domain interventions into the input representation.

Our neural network model, illustrated in fig. 1, adopts a hybrid Transformer-CNN architecture to capture long-range temporal dependencies and local patterns shaped by intervention dynamics. A self-attention block [20] models event

interactions via attention weights, with multi-head attention enhancing expressiveness by capturing diverse relational patterns. To incorporate temporal context, we embed relative time differences $t_i - t_{i-1}$ using temporal positional encodings [27]. This architecture is used to model the logarithm of the outcome CIF, $\lambda(t)$, which is central to ATE estimation.

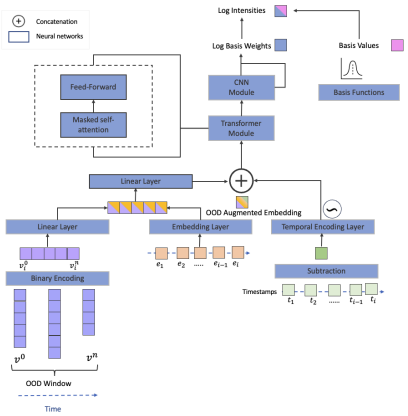

Fig. 1: Our neural network architecture for learning temporal point processes with out-of-domain interventions. It combines out-of-domain intervention-augmented (V_i) event embeddings (e_i) with positional encoding (Δt), uses Transformer and CNN modules for pattern extraction, and estimates the CIF $\lambda(t)$ for event prediction

Our model input comprises three encoded components: 1) binary indicators of out-of-domain interventions v occurring within the time window; 2) dense embeddings of event types (cause, outcome, and other observed events); and 3) relative event times, $t_i - t_{i-1}$, encoded using trigonometric positional encodings to capture temporal positions. We combine intervention and event type embeddings via a weighted sum to form a unified, intervention-aware event representation, which is then augmented with temporal positional encodings. The model outputs the conditional intensity function (CIF) for each event and timestamp, expressed as a sum over basis functions.

We optimize our model using three criteria combined into a single loss function. raining is performed on batches of sequences $B = \{s_1, s_2, ..., s_b\}$ where b is the batch size. For each batch, we compute the average loss across all sequences.

The first criterion is negative log likelihood (NLL) of event observations within the sequences.

$$L_{NLL} = \frac{1}{b} \sum_{k=1}^{b} \left[- \sum_{i=1}^{L_k} \log \lambda_{e_{k,i}}(t_{k,i}) + \sum_{e \in \text{outcome}} \int_0^T \lambda_e(t) dt \right], \quad (4)$$

where $\lambda_{e_{k,i}}(t_{k,i})$ is the conditional intensity function for event $e_{k,i}$ at timestamp $t_{k,i}$, and L_k is the length of sequence k. The first term represents the log-likelihood of observed events, while the second accounts for the probability of no events occurring in the remaining time intervals.

The second criterion is a prediction loss based on cross-entropy that measures

how well the model predicts event types.

$$\mathcal{L}_{CE} = -\frac{1}{b}\sum_{k=1}^{b}\sum_{i=1}^{L_k}\sum_{e\in\mathbb{E}} y_{k,i,e}\log(p_{k,i,e}), \tag{5}$$

where $\mathbb{E}$ is the set of event classes, $y_{k,i,e}$ is the true binary indicator (0 or 1) of whether event e is the correct classification for i-th event in $\mathbf{s}_k$, $p_{k,i,e}$ is the predicted probability that i-th event belongs to class e.

The third criterion is L2 regularization term that penalizes large basis weights to prevent overfitting.

$$L_{reg} = \frac{1}{b}\sum_{k=1}^{b}\sum_{i=1}^{L_k}\sum_{l=1}^{\mathcal{B}}\exp(w_{l,k,i})^2, \tag{6}$$

where $w_{l,k,i}$ represents the log basis weights at time $t_{k,i}$ for basis function l, $\mathcal{B}$ denotes the total number of basis functions. The regularization prevents the basis weights from growing too large, which could lead to numerical instabilities.

Our final batch loss function combines the three criteria, (4), (5) and (6),

$$L_{batch} = L_{NLL} + \alpha L_{CE} + \beta L_{reg}. \tag{7}$$

The hyperparameters α and β control the relative importance of the cross-entropy loss and regularization term, respectively. We minimize this batch loss using the Adam optimizer and evaluate convergence on a held-out validation set. Model selection is performed via 5-fold cross-validation. Additional hyperparameter and implementation details are provided in supplementary materials.

5 Numerical Study

We evaluated our method on both simulated and real-world data, comparing it against the CAUSE model [24] for ATE estimation and out-of-domain intervention-augmented process learning. CAUSE, which combines point processes and attribution methods for Granger causality inference, estimates the CIF without accounting for out-of-domain interventions. A comparative discussion with other models is provided in the supplementary materials.

5.1 Simulated Data

We generated synthetic event sequences with injected out-of-domain interventions, modeled as random events influencing specific cause–outcome pairs. Each intervention was assigned a random occurrence probability and time window, with dynamic injection based on these probabilities. Additional simulation details are provided in the supplementary materials.

We simulated three intervention impact types, increasing in complexity and coverage of out-of-domain intervention impact:

1. **No out-of-domain intervention (No OOD)**: we do not impose out-of-domain intervention on process, which also can serve as sanity check.
2. **Baseline out-of-domain intervention (Baseline**): we modify the baseline intensity $\mu(t)$ of an outcome event CIF $\lambda(t)$ under out-of-domain intervention shown in Eq. 1. We implement 30 out-of-domain interventions per sequence.
3. **All out-of-domain intervention (All impact)**: we alter baseline intensity $\mu(t)$, the influence of cause and other observed events through $\phi(\cdot)$ on outcome

event CIF $\lambda(t)$ under out-of-domain intervention. We implement 30 out-of-domain interventions per sequence.

We generate $1,000$ sequences for each intervention impact type, with sequence lengths drawn from a Poisson distribution (mean = 500). Each sequence includes 30 distinct event types: 20 used as causes and covariates, and the remaining 10 are designated as outcomes.

We compare our method with CAUSE in ATE estimation and process fitting in simulated data. The performance measures we adopt for ATE estimation are

Bias: the average of absolute difference between ATE estimate and true ATE across repetitions.

Variance: the variance of ATE estimate across repetitions.

Mean squared error (MSE): the average of squared difference between ATE estimate and true ATE across repetitions.

Lower values indicate better ATE estimation. As shown in table 1, our method significantly reduces bias across all scenarios, with slight increase in variance. This trade-off is beneficial, as it results in lower MSE in most cases.

Table 1: Simulated data ATE estimation performance comparison between our method and CAUSE for Baseline and All impact interventions. Lower values are better.

Out-of-Domain Intervention	Intervention Status	Bias		Variance		MSE	
		Ours	Cause	Ours	Cause	Ours	Cause
Baseline	0	**0.0155**	0.0522	0.0141	**0.00001**	**0.003**	0.0046
	1	**0.1561**	0.262	**0.0002**	**0.0002**	**0.027**	0.0708
All Impact	0	**0.0209**	0.0494	0.0027	**0.00001**	**0.0038**	0.0042
	1	**0.1615**	0.2192	0.0008	**0.0002**	**0.0283**	0.0498

We use the following performance measures for process fitting,

Negative log likelihood (NLL): the NLL of fitted process model.

Root mean squared error (RMSE): the square root of the average of squared outcome event occurrence time prediction error.

Mean absolute error (MAE): the average of absolute outcome event prediction error.

Averaged across repetitions, table 2 shows that our model consistently outperforms CAUSE across all scenarios by achieving lower NLL, RMSE, and MAE; demonstrating superior modeling of event distributions and predictive accuracy, even under varying intervention complexities.

5.2 Real-World Data

We evaluated our method on two real-world datasets:

Predictive Maintenance[1]: This dataset contains hourly sensor readings (voltage, rotation, pressure, vibration) from 100 machines. Proactive maintenance events (scheduled component replacements) were modeled as out-of-domain interventions, while reactive maintenance events (failures requiring unscheduled

Table 2: Simulated data process fitting performance comparison (NLL, RMSE, MAE) between our method and CAUSE with varying Out-of-Domain interventions. Lower values are better.

Out-of-Domain Intervention	NLL		RMSE		MAE	
	Ours	Cause	Ours	Cause	Ours	Cause
No OOD	**1009.06**	2456.53	**1.78**	3.89	**0.89**	2.60
Baseline	**852.48**	2438.36	**2.1**	3.84	**1.04**	2.56
All Impact	**1299.61**	2443.81	**2.18**	3.86	**1.17**	2.58

replacements) served as outcomes. Continuous sensor readings were discretized into 625 bins based on mean deviation, with each bin treated as a potential cause event and others as time-varying covariates.

Diabetes [10]: This dataset includes lab events from 70 patients, such as insulin injections, glucose levels, meal consumption, physical activity, and hypoglycemia symptoms. We focused on glucose as the outcome and selected cause and OOD intervention events based on established medical knowledge. For example, in analyzing insulin-mediated effects of meals on glucose, we treated meal as the cause, insulin as the out-of-domain intervention, and activity and hypoglycemia symptoms as covariates. A baseline experiment without interventions was also conducted to assess model robustness.

Since ground truth ATEs and prior benchmarks are unavailable for the predictive maintenance dataset, we evaluated process fitting performance. As shown in table 3, our model significantly outperformed CAUSE in future event prediction:achieving an 88% reduction in RMSE and a 90% reduction in MAE, while maintaining comparable distribution modeling (only 4% higher NLL). These improvements highlight the value of explicitly modeling out-of-domain interventions in capturing complex real-world dynamics.

While ground truth ATEs for the diabetes dataset are unavailable, we validated the detected causal relation shifts against established medical literature. As shown in table 4, the shifts identified by our method align with known findings. For example, our model captured the insulin-mediated effect where glucose levels decrease despite increased meal intake, consistent with results reported in [3]. Additionally, table 5 demonstrates that our method outperforms CAUSE in process fitting across all intervention types, underscoring the benefits of incorporating out-of-domain interventions. The model also maintained strong perfor-

Table 3: Predictive maintenance data process fitting performance (NLL, RMSE, MAE) comparison between our method and CAUSE, with proactive maintenance as out-of-domain intervention. Lower values are better.

NLL		RMSE		MAE	
Ours	Cause	Ours	Cause	Ours	Cause
300.66	**288.84**	**419.38**	3501.14	**59.88**	628.37

Table 4: ATE-estimated causal relations in diabetes dataset under Out-of-domain interventions, validated by medical literature.

Cause	Outcome	Out-of-Domain Intervention	Covariate	Our Conclusion & Literature Evidence
More than usual meal ingestion	Blood glucose decrease	Insulin	Activity, hypoglycemia symptopm	✓ Insulin injection, even when more than usual meal is ingested, may help to decrease blood glucose [3]
Hypoglycemia symptom	Blood glucose decrease	NPH insulin	Meal, activity	✓ Insulin injection, when hypoglycemic symptom is observed, may cause blood glucose to decrease further [13]
Insulin	Blood glucose decrease	Typical or more than usual activity	Meal, hypoglycemia symptom	✓ More activity increases the effect of insulin, boosting blood glucose decrease [12]

Table 5: Process fitting performance comparison between our method and CAUSE on the diabetes dataset, with activity and insulin as out-of-domain interventions. Sanity check: no intervention. Lower NLL, RMSE, and MAE indicate better performance.

Out-of-Domain Intervention	NLL		RMSE		MAE	
	Ours	Cause	Ours	Cause	Ours	Cause
Sanity Check	**613.21**	619.74	**15644.33**	15718.8	**2104.87**	2239.53
Activity	**563.47**	1345.61	**15729.25**	15770.73	**2108.1**	2111.2
Insulin	**181.57**	190.11	**342.74**	360.78	**55.81**	61.69

mance in the baseline setting without interventions, confirming its robustness.

6 Discussion

While our method offers significant advantages in modeling causal relation shifts under out-of-domain interventions, two key limitations merit consideration. First, our framework simplifies interventions to binary occurrence indicators, potentially limiting expressiveness for complex interventions with varying intensity or duration. Second, if interventions open backdoor causal paths between cause and outcome events, estimation accuracy may be compromised, requiring careful domain knowledge for intervention selection. Scalability concerns are discussed separately in the supplementary materials. Future work should extend the framework to continuous intervention variables for more fine-grained modeling and improved estimation precision.

7 Conclusion

We propose a novel causal framework leveraging out-of-domain interventions to analyze causal relation shifts in temporal process data. Our Transformer-based Hawkes process integrates intervention effects with an unbiased treatment effect estimator for accurate causal quantification. Extensive simulations and real-world validation, corroborated by literature, demonstrated superior performance in both process fitting and ATE estimation, enabling practical extraction of actionable causal knowledge for informed decision-making.

Bibliography

[1] Azure: AI Predictive Maintenance. https://github.com/Azure/AI-PredictiveMaintenance (2018)

[2] Bhattacharjya, D., Subramanian, D., Gao, T.: Proximal graphical event models. null (2018). https://doi.org/null

[3] Campbell, M.D., Walker, M., Trenell, M.I., Luzio, S., Dunseath, G., Tuner, D., Bracken, R.M., Bain, S.C., Russell, M., Stevenson, E.J., et al.: Metabolic implications when employing heavy pre-and post-exercise rapid-acting insulin reductions to prevent hypoglycaemia in type 1 diabetes patients: a randomised clinical trial. PLoS One **9**(5), e97143 (2014)

[4] Daley, D.J., Vere-Jones, D., et al.: An introduction to the theory of point processes: volume I: elementary theory and methods. Springer (2003)

[5] Didelez, V.: Graphical models for marked point processes based on local independence. null (2005). https://doi.org/null

[6] Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., Song, L.: Recurrent marked temporal point processes: Embedding event history to vector. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1555–1564 (2016)

[7] Gao, T., Subramanian, D., Bhattacharjya, D., Shou, X., Mattei, N., Bennett, K.P.: Causal inference for event pairs in multivariate point processes. Advances in Neural Information Processing Systems **34**, 17311–17324 (2021)

[8] Hawkes, A.G.: Point spectra of some mutually exciting point processes. Journal of the Royal Statistical Society Series B: Statistical Methodology **33**(3), 438–443 (1971)

[9] Imbens, G.W., Rubin, D.B.: Causal inference in statistics, social, and biomedical sciences. Cambridge University Press (2015)

[10] Kahn, M.: Diabetes. UCI Machine Learning Repository (1994), DOI: https://doi.org/10.24432/C5T59G

[11] Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: International conference on machine learning. pp. 5637–5664. PMLR (2021)

[12] Lawrence, R.: The effect of exercise on insulin action in diabetes. British medical journal **1**(3406), 648 (1926)

[13] Lee, S.J.: So much insulin, so much hypoglycemia. JAMA internal medicine **174**(5), 686–688 (2014)

[14] Mei, H., Eisner, J.M.: The neural hawkes process: A neurally self-modulating multivariate point process. Advances in neural information processing systems **30** (2017)

[15] Melnychuk, V., Frauen, D., Feuerriegel, S.: Causal transformer for estimating counterfactual outcomes. In: International Conference on Machine Learning. pp. 15293–15329. PMLR (2022)

[16] Noorbakhsh, K., Rodriguez, M.: Counterfactual temporal point processes. Advances in Neural Information Processing Systems **35**, 24810–24823 (2022)
[17] Schulam, P., Saria, S.: Reliable decision support using counterfactual models. Advances in neural information processing systems **30** (2017)
[18] Shchur, O., Türkmen, A.C., Januschowski, T., Günnemann, S.: Neural temporal point processes: A review (2021)
[19] Shi, C., Blei, D., Veitch, V.: Adapting neural networks for the estimation of treatment effects. Advances in neural information processing systems **32** (2019)
[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
[21] Xiao, C., Choi, E., Sun, J.: Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association **25**(10), 1419–1428 (2018)
[22] Xiao, S., Yan, J., Yang, X., Zha, H., Chu, S.: Modeling the intensity function of point process via recurrent neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017)
[23] Zhang, L.n., Liu, J.w., Song, Z.y., Zuo, X.: Temporal attention augmented transformer hawkes process. Neural Computing and Applications pp. 1–15 (2022)
[24] Zhang, W., Panum, T., Jha, S., Chalasani, P., Page, D.: Cause: Learning granger causality from event sequences using attribution methods. In: International Conference on Machine Learning. pp. 11235–11245. PMLR (2020)
[25] Zheng, Y., Chen, G., Huang, M.: Out-of-domain detection for natural language understanding in dialog systems. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 1198–1209 (2020)
[26] Zhou, Y., Wang, Y., Song, H., Meharizghi, T., Jazaery, M.A., Xu, P., Colin-Magana, D., Abeyakoon, A.: Deep sequence modeling for event log-based predictive maintenance. In: KDD 2023 International Workshop on Mining and Learning from Time Series (MileTS) (2023)
[27] Zuo, S., Jiang, H., Li, Z., Zhao, T., Zha, H.: Transformer hawkes process. In: International conference on machine learning. pp. 11692–11702. PMLR (2020)