INFERRING CAUSAL RELATIONS BETWEEN TEMPORAL EVENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Due to the popularity of event-based data, causal inference from event datasets has attracted increasing interest. However, inferring causalities from observational event sequences is challenging because of the heterogeneous and irregular nature of event-based data. Existing work on causal inference for temporal events disregards the event durations, and is thus unable to capture their impact on the causal relations. In the present paper, we overcome this limitation by proposing a new modeling approach for temporal events that captures event durations. Based on this new temporal model, we propose a set of novel Duration-based Event Causality (DEC) scores, including the *Duration-based Necessity and Sufficiency Trade-off* score, and the *Duration-based Conditional Intensity Rates* scores that utilizes event durations when inferring causal relations between temporal events. We show that the proposed scores follow the causality hypothesis testing framework. We conduct a comprehensive experimental evaluation using both synthetic datasets, and two real-world event datasets in the medical and environmental domains to evaluate our proposed scores, and compare them against the closest baseline. The experimental results show that our proposed scores outperforms the baseline with a large margin using the popular evaluation metric Hits@K.

1 INTRODUCTION

Discovering causal relationships in observational data is a fundamentally important yet very challenging task. Often, researchers are interested in understanding what causes a phenomenon, or how the occurrence of one event can affect others. Such understandings are useful in many domains, from economics to public health, social and natural sciences. For example, in the medical domain, understanding what causes a certain reaction in a patient given a medical treatment can help doctors create or change a treatment plan for better outcomes. In stock trading, knowing causal factors behind the increase or decrease in stock prices can improve trading strategies. Research in causality has been mainly driven by two tasks: causal discovery and causal inference Nogueira et al. (2022). Starting with observational data, causal discovery tries to predict causal relations across variables directly from the data, without assuming any relationships among the involved variables. In comparison, causal inference assumes an existing relationship among tested variables, and tries to quantify that specific relationship to assess how one variable impacts the others given the available data. Causal discovery and inference have been applied to different types of data, from text documents to understand causal implications between phrases Luo et al. (2016) to temporal data to infer causal relations in time series Huang & Kleinberg (2015) and event sequences Bhattacharjya et al. (2021). Specially, event sequences, a type of temporal data which consist of series of events occurring over a timeline, have become increasingly popular and can be collected from a broad range of sources such as computer logs, financial transactions, electronic health records, etc. Unlike time series data which consist of observations sampled from real-valued variables at regular times (e.g., stock prices), event sequences represent events that are typically generated at irregular and asynchronous time intervals (e.g., medical records, meteorological phenomena). In an event sequence, the timestamp is associated with each generated event to denote its occurrence time, from which the duration of an event can be deduced. Causal relations inferred from such temporal event sequences can provide important insights into how the events progress, and how they impact and depend on each other so that precise event forecasting or intervention can be made. For instance, knowing that a heavy rain event lasting for more than two days will likely cause flooding will enable better preventive action and

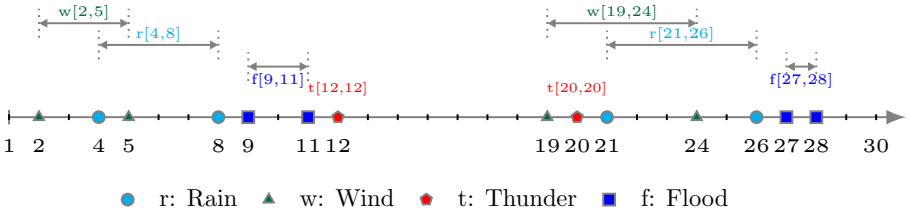


Figure 1: Temporal events over the time horizon $\mathcal{T}$

preparation; predicting that certain drugs, e.g, benzodiazepine, can incur addiction or side-effects after a certain use duration will enable doctors to make more informed choices about medication.

Inferring causal relationships in temporal data is commonly based on the theory of Granger causality Granger (1969), where causal relations are defined based on two conditions: the temporal ordering of events, and the conditional dependency between causes and effects. There has been extensive research focusing on learning Granger causality in temporal data, e.g., based on graphical models Lozano et al. (2009), Peters et al. (2013) and Hawkes processes Achab et al. (2017). Although these techniques have demonstrated their reliability in identifying causal relations in time series data, they still have limitations when applied to event datasets. As many of these techniques rely on assumptions about the data distributions and regular arrival of data samples, they fail to capture the irregular and asynchronous nature of event datasets. The most recent and notable existing work for causal inference in event-based data is that of Bhattacharjya et al. (2021) which investigates causal associations in structured event datasets. The authors propose a set of scores that can infer causal associations between sequential events. Although shown to work well on event datasets, the proposed scores in Bhattacharjya et al. (2021) are limited to only sequential events, which assumes events occur in sequential order as a series of points in time. This assumption prevents the proposed model in Bhattacharjya et al. (2021) from considering event durations and their roles in causality, and thus, fail to capture causal relations such as between rain and flooding events mentioned above.

In this work, we overcome this limitation by proposing a new way to model event datasets that is able to capture and utilize event durations when assessing causality. This improves the model in Bhattacharjya et al. (2021) by providing a more accurate causal inference approach for real-world data, especially in cases where causal relations strongly depend on whether a cause lasts long enough for an effect to occur, as argued by Hicks et al. (1980).

Contributions. This paper presents our Duration-based Event Causality (DEC) model with several key contributions. (1) First, we propose a new modeling approach for event datasets where events have detailed temporal information, including occurrence time and duration. Using this model, we can describe causal relations between events where both cause and effect span a finite period of time. (2) Building on this temporal event model, we propose a set of novel Duration-based Event Causality (DEC) scores, including the *Duration-based Necessity and Sufficiency Trade-off* score, and the *Duration-based Conditional Intensity Rates* scores to infer causalities between pairs of temporal events, taking into consideration event durations when assessing their causal relations. (3) We prove that the proposed duration-based scores follow the causality hypothesis testing framework. Finally (4), we perform an extensive experimental evaluation on synthetic and real-world datasets which show that our proposed causality model outperforms the state-of-the-art baseline in Bhattacharjya et al. (2021) with a large margin, using the evaluation metric Hits@K. For reproducibility purpose, we make code and data available at: `https://github.com/causalityinf/causality`.

2 PRELIMINARIES

Discrete time horizon: To represent the occurrence of cause and effect through time, we use a discrete time horizon $\mathcal{T}$ where time is discretized into equal time steps, each representing a point in time. The time steps in $\mathcal{T}$ adhere to an increasing order, which starts at the first time step $t_1 = 1$, and ends at the last time step $t_N = T$, where $T = |\mathcal{T}|$ is the total time steps over $\mathcal{T}$.

Fig. 1 shows an example of the discrete time horizon $\mathcal{T}$ that has $T = 30$ time steps in total. Over $\mathcal{T}$, multiple events occur at different time steps. For example, the Wind event occurs at time step 2 and ends at 5, while the Rain event occurs at time step 4 and ends at 8.

Temporal event: A *temporal event* E is a tuple $E = (\omega, t)$ where ω is the event label, and $t = \{[t_{s_i}, t_{e_i}]\}$ is the set of time intervals during which E occurs. In each time interval $t_i = [t_{s_i}, t_{e_i}]$, t_{s_i} is the start time step, and t_{e_i} is the end time step of that interval.

Instance of a temporal event: Let $E = (\omega, t)$ be a temporal event, and $t_i = [t_{s_i}, t_{e_i}] \in t$ be a single time interval in the set t. The tuple $e = (\omega, t_i)$ is called an *instance* of the event E, representing a single occurrence of E during $[t_{s_i}, t_{e_i}]$.

Duration of an event instance: Given an event instance $e = (\omega, [t_{s_i}, t_{e_i}])$, the duration of e is: $d_e = t_{e_i} - t_{s_i}$, denoting the number of time steps that e lasts.

Fig. 1 illustrates four different temporal events: **r** (Rain), **w** (Wind), **t** (Thunder), and **f** (Flood), and their arrangement over the time horizon $\mathcal{T}$. Event Wind has two instances: (w, [2, 5]), (w, [19, 24]), Rain has two instances: (r, [4, 8]) and (r, [21, 26]), Thunder has two instances: (t, [12, 12]) and (t, [20, 20]), and finally Flood has two instances: (f, [9, 11]) and (f, [27, 28]). The first instance of Wind (w, [2, 5]) has a duration $d_w = 3$, indicating that it lasted 3 time steps.

Temporal order: Let $e_A = (\omega_A, [t_{s_i}, t_{e_i}])$ and $e_B = (\omega_B, [t_{s_j}, t_{e_j}])$ be event instances of two temporal events A and B, respectively. We say that e_A and e_B adhere to the *temporal order* if $t_{s_i} \leq t_{s_j}$ and e_A is ordered before e_B in the time horizon $\mathcal{T}$.

For example, in Fig. 1, two event instances Wind = (w, [2, 5]) and Rain = (r, [4, 8]) adhere to the temporal order since $t_{s_w} = 2 \leq t_{s_r} = 4$, and thus, w is ordered before r in $\mathcal{T}$.

Temporal event dataset: A temporal event dataset $\mathcal{D}$ is an ordered list of temporal events where every pair of instances adhere to the temporal order over $\mathcal{T}$.

An example of a temporal event dataset can be seen in Fig. 1. Here, the time horizon $\mathcal{T}$ spans from 1 to 30, and we have a set of eight event instances of four temporal events where every instance pair adheres to the temporal order.

Time window: Let $\mathcal{D}$ be a temporal event dataset over a time horizon $\mathcal{T}$. A window $w = [t_l, t_r]$ where t_l is the left time step marking the start of w, and t_r is the right time step marking the end of w, is a time window in $\mathcal{T}$ if $[t_l, t_r] \subseteq \mathcal{T}$ and w contains all event instances from t_l to t_r in $\mathcal{D}$. The window w has size n where $n = t_r - t_l$.

Preceding time window: Let $\mathcal{D}$ be a temporal event dataset over $\mathcal{T}$, and $e_i = (\omega, [t_{s_i}, t_{e_i}])$ be an instance of event E. A time window $w = [t_l, t_r]$ of size n is called a *preceding time window* of E w.r.t e_i if $t_r + 1 = t_{s_i}$, i.e., w precedes the start of e_i.

Succeeding time window: Let $\mathcal{D}$ be a temporal event dataset over $\mathcal{T}$, and $e_i = (\omega, [t_{s_i}, t_{e_i}])$ be an instance of event E. A time window $w = [t_l, t_r]$ of size n is called a *succeeding time window* of E w.r.t e_i if $t_{e_i} + 1 = t_l$, i.e., w follows the end of e_i.

In Fig. 1, consider the window $w_1 = [2, 8]$ of size $n = 6$. Then, w_1 contains two instances: Wind = (w, [2, 5]) and Rain = (r, [4, 8]). Moreover, w_1 is the *preceding window* of the Flood event w.r.t the instance Flood = (f, [9, 11]). Instead, another time window $w_2 = [12, 18]$ of size $n = 6$ is the *succeeding window* of Flood = (f, [9, 11]).

3 Duration-Based Causality Scores

In this section, we propose a set of duration-based causality scores to assert causal relations between pairs of temporal events. These scores utilize event durations to measure their impact on the strength of causalities. The causality hypothesis testing framework from which the proposed scores rely on, together with the necessity and sufficiency conditions and the temporal classification of causal relations are discussed in the Appendix A.1.

3.1 Duration-Based Necessity-Sufficiency Score

When asserting causal relations between pairwise temporal events, we utilize a window-based approach which is built on the assumption that causal effect occurs only within a limited time duration. Below, we propose a novel adaptation of Bhattacharjya et al. (2021) where event durations are explicitly considered.

Notation convention. In the following sections, we use uppercase letters to denote events, such as X, Y, Z, and lowercase letters to denote event instances such as x, y, z.

Necessity causality of a pair of temporal events: Let w be a time window of size n, and (Y, X) be a pair of temporal events with the causal relation Y is the cause of X. The window-based conditional probability of observing X given Y, which we call the *necessity causality*, is computed as

$$p^w(X|Y) = \frac{p^w(Y \leftarrow X)}{p(Y)} \cdot \frac{D^w(Y)}{D(Y)} \tag{1}$$

In equation 1, $p^w(Y \leftarrow X)$ represents the probability of X occurring if Y has occurred in X's preceding time window, and $p(Y)$ is the probability of Y occurring over a time horizon $\mathcal{T}$. These two terms are respectively computed as

$$p^w(Y \leftarrow X) = \frac{N^w(Y \leftarrow X)}{T}; \quad p(Y) = \frac{N(Y)}{T} \tag{2}$$

where $N^w(Y \leftarrow X)$ is the number of preceding time windows w of X that contains Y, $N(Y)$ is the number of Y instances occurring over $\mathcal{T}$, and T is the total time steps in $\mathcal{T}$.

The two terms $D^w(Y)$ and $D(Y)$ in equation 1 are the total duration of Y in the preceding window w of X, and the total duration of Y over $\mathcal{T}$. They are respectively computed as

$$D^w(Y) = \sum_{y \in N^w(Y \leftarrow X)} d_y; \quad D(Y) = \sum_{y \in \mathcal{T}} d_y \tag{3}$$

In equation 1, $p^w(X|Y)$ represents the window-based conditional probability of X (the effect) given Y (the cause). The first term, $\frac{p^w(Y \leftarrow X)}{p(Y)}$ compares the frequency of cause Y and effect X co-occurring with the condition of Y preceding X, to the frequency of Y alone. The second term, $\frac{D^w(Y)}{D(Y)}$ compares the duration of cause Y when it co-occurs with effect X, to its total duration. If the causal relation between Y and X hold only when cause Y lasts long enough, it will be captured by this duration ratio since it strengthens the necessity causality.

Sufficiency causality of pairwise temporal events: The window-based conditional probability of cause Y, given that effect X has been observed is computed as

$$p^w(Y|X) = \frac{p^w(Y \rightarrow X)}{p(X)} \cdot \frac{D^w(X)}{D(X)} \tag{4}$$

where $p^w(Y \rightarrow X)$, $p(X)$, $D^w(X)$ and $D(X)$ are:

$$p^w(Y \rightarrow X) = \frac{N^w(Y \rightarrow X)}{T}; \quad p^w(X) = \frac{N(X)}{T}; \quad D^w(X) = \sum_{x \in N^w(Y \rightarrow X)} d_x; \quad D(X) = \sum_{x \in T} d_x \tag{5}$$

In equation 4, $p^w(Y \rightarrow X)$ is the probability of cause Y such that effect X occurs in Y's succeeding window, $p(X)$ is the probability of X over $\mathcal{T}$, $D^w(X)$ is the duration of X in Y's succeeding window, and $D(X)$ is the duration of X over $\mathcal{T}$.

The probability $p^w(Y|X)$ represents the conditional probability of Y (the cause) given X (the effect). The second term $\frac{D^w(X)}{D(X)}$ in equation 4 compares how long X lasts when Y occurs in its preceding window, to the total duration of X over $\mathcal{T}$. A causal relation between Y and X that depends on the durations of cause and effect will be captured by this term, e.g., when an effect X lasts longer in the presence of cause Y.

Duration-based Necessity-Sufficiency Trade-off Score: Using the proposed necessity and sufficiency causalities in equation 1 and equation 4, we compute the *Duration-based Necessity-Sufficiency Trade-off Score* (DNST), adapted from NST score in Bhattacharjya et al. (2021), to consider event durations when measuring their causal association as

$$DNST(Y, X) = \left[\frac{p^w(X|Y)}{p(X)^{-\alpha}}\right]^{\delta} \left[\frac{p^w(Y|X)}{p(Y)^{-\alpha}}\right]^{1-\delta} = \left[\frac{p^w(Y \leftarrow X)}{p(X)^{-\alpha}p(Y)} \cdot \frac{D^w(Y)}{D(Y)}\right]^{\delta} \left[\frac{p^w(Y \rightarrow X)}{p(Y)^{-\alpha}p(X)} \cdot \frac{D^w(X)}{D(X)}\right]^{1-\delta} \tag{6}$$

The $DNST$ score in equation 6 requires two additional parameters, a penalization parameter $\alpha \geq 0$ to penalize frequent events, and a trade-off parameter $\delta \in [0, 1]$ to weigh the importance between necessity and sufficiency causalities. As frequent events have high probabilities and thus, can create bias by making the causal association highly likely, a penalty is applied to reduce their impact on the

causal association. The higher the α, the larger the penalty. Note that we use $-\alpha$ in the denominator since $p(X)$ and $p(Y)$ are ≤ 1.

Using equation 6, the higher the $DNST$ score, the more confidence that pair of events exhibit causal association. Furthermore, based on the trade-off parameter δ, we have the option to weigh more necessity or more sufficiency: a high δ value will weigh more on the necessity, and vice versa.

Theorem 1. *Let w be a time window of size n, and (Y, X) be a pair of temporal events with the causal relation Y is the cause of X. Then $DNST(Y, X) = 0$ indicates independence, and $DNST(Y, X) > 0$ indicates dependence of the pair (Y, X).*

The proof is provided in Appendix A.2. From Theorem 1, the occurrence of effect X depends on the occurrence of cause Y if $DNST(Y, X) > 0$. Hence, $DNST$ follows the hypothesis testing framework (equation 18, Appendix A.1), with the pair (Y, X) being tested for the causal relation.

3.2 Duration-Based Conditional Intensity Scores

The DNST score introduced in Section 3.1 relies on $p(X)$ and $p(Y)$ which measure the probabilities of X and Y over $\mathcal{T}$. This is a limitation since $p(X)$ and $p(Y)$ can be computed only if $\mathcal{T}$ is finite, and that X and Y have to occur at least once throughout the time horizon $\mathcal{T}$. In the setting where time is measured continuously and infinite, DNST score becomes impractical. Therefore, causality scores for infinite time domains are needed. For this reason, we introduce the duration-based conditional intensity rate scores that rely on event durations, thus free themselves from such limitations.

Duration-based Conditional Intensity Rate: When modeling an event dataset as marked point processes, Didelez (2008) introduce the conditional intensity function $\lambda_X(t \mid \mathcal{H}) \geq 0$ to measure the rate at which event X occurs at time t, given the available history $\mathcal{H}$. Bhattacharjya et al. (2021) propose conditional intensity rates that treat each event occurrence as a point in time. In our adaptation with respect to Didelez (2008) and Bhattacharjya et al. (2021), we formulate the duration-based conditional intensity rate using event durations as follows.

Let (Y, X) be a pair of temporal events with the causal relation that Y is the cause of X. To measure the rate at which X occurs w.r.t Y, we define two conditional intensity rates: $\lambda_{X|Y}$, representing the rate at which effect X occurs given the presence of cause Y, called the *positive rate*, and $\lambda_{X|\overline{Y}}$ measuring the rate at which X occurs without Y occurring, and is called the *negative rate*.

Positive rate: the *positive rate* $\lambda_{X|Y}$ is computed as

$$\lambda_{X|Y} = \frac{D^w(Y)}{D(Y)} \tag{7}$$

where $w \in N^w(Y \leftarrow X)$ is the preceding window of X that contains Y, $D^w(Y)$ is the duration of Y in w, and $D(Y)$ is the duration of Y in $\mathcal{T}$. In equation 7, $\lambda_{X|Y}$ is computed as the ratio between the duration of Y occurring in the preceding window of X, and the total duration of Y over $\mathcal{T}$. This implies that the rate of effect X occurring w.r.t Y is entirely dependent on the presence and the duration of its cause Y. The longer the cause, the higher the occurring rate of the effect. The terms $D^w(Y)$ and $D(Y)$ are respectively computed as

$$D^w(Y) = \sum_{w \in N^w(Y \leftarrow X)} \int_{t=t_l}^{t_r} d_Y(t)dt; \quad D(Y) = \int_{t=1}^{T} d_Y(t)dt \tag{8}$$

where t_l, t_r are the left time step and the right time step of the preceding time window w of X, respectively, and $d_Y(t)$ is the indication function of event Y at time t:

$$d_Y(t) = \begin{cases} 1, & \text{if Y occurs at time t} \\ 0, & \text{otherwise} \end{cases}$$

The integration in equation 8 integrates the occurrence of Y over $\mathcal{T}$, i.e., entailing its duration.

Negative rate: the *negative rate* $\lambda_{X|\overline{Y}}$ is computed as

$$\lambda_{X|\overline{Y}} = \frac{D^{\overline{w}}(X)}{T - D^w(Y)} \tag{9}$$

where $w \in N^w(Y \leftarrow X)$ is the preceding window of X that contains Y, while $\overline{w} \in N^{\overline{w}}(\overline{Y} \leftarrow X)$ is the preceding window of X that does not contain Y. The term $D^w(Y)$ is computed as in equation 8, while $D^{\overline{w}}(X)$ is computed as

$$D^{\overline{w}}(X) = \sum_{\overline{w} \in N^{\overline{w}}(\overline{Y} \leftarrow X)} \int_{t=t_l}^{t_r} d_X(t) dt; \tag{10}$$

where $d_X(t)$ is the indication function of event X at time t:

$$d_X(t) = \begin{cases} 1, & \text{if X occurs at time t} \\ 0, & \text{otherwise} \end{cases}$$

In equation 9, $\lambda_{X|\overline{Y}}$ is computed as the ratio between the duration of X during which Y has not occurred in its preceding window, and the duration in $\mathcal{T}$ during which Y does not have an impact on X. We note that T in equation 9 is the number of time steps in $\mathcal{T}$, however, its meaning is different from T used in DNST. More specifically, since the duration-based conditional intensity rates rely on the available history $\mathcal{H}$, the value of T in equation 9 only reflects the time duration in $\mathcal{H}$. In contrast, T in DNST refers to the entire time dimension from which the probabilities of events are computed.

Duration-based Conditional Intensity Rate Causality Scores (DCIR): using the above *positive rate* and *negative rate*, we adapt the conditional intensity rate scores in Bhattacharjya et al. (2021) to compute two duration-based conditional intensity rate causality scores as

$$DCIR_P = \frac{\lambda^w_{X|Y}}{\lambda_X}; \quad DCIR_N = \frac{\lambda^w_{X|Y}}{\lambda^w_{X|\overline{Y}}} \tag{11}$$

where λ_X is the rate of X occurring throughout the history $\mathcal{H}$, without considering any causes, and is computed as: $\lambda_X = \frac{D(X)}{T}$. The subscripts P and N in $DCIR_P$ and $DCIR_N$ stand for *positive* and *negative* score. The positive causality score $DCIR_P$ compares the rate of effect X occurring when cause Y is present, to the rate of X throughout $\mathcal{H}$. Instead, the negative causality score $DCIR_N$ compares the rate of effect X occurring when cause Y is present, to the rate of X when Y is not present. The higher the value of $DCIR_P$ (or $DCIR_N$), the more likely that Y causes X.

Theorem 2. *Let $DCIR_{P/N}$ be either $DCIR_P$ or $DCIR_N$. Let w be a time window of size n, and (Y, X) be a pair of temporal events with the causal relation Y is the cause of X. Then $DCIR_{P/N} = 1$ indicates independence, and $DCIR_{P/N} \neq 1$ indicates dependence of the pair (Y, X).*

We provide the proof in Appendix A.2. Theorem 2 says that effect X depends on cause Y if $DCIR_{P/N} \neq 1$. Hence, $DCIR_{P/N}$ follows the hypothesis testing framework (equation 18, Appendix A.1), with the pair (Y, X) being tested for the causal relation.

Multi-cause Duration-based Conditional Intensity Rate: When computing the $DCIR_P$ and $DCIR_N$ scores for the causal pair (Y, X), we assume that event X has only one cause Y. This assumption is not realistic, since in practice, it is more common that an effect X has multiple causes. To generalize this causal scenario, we extend $DCIR_P$ and $DCIR_N$ to consider multiple causes for an effect X as follows.

Consider an event pair (Y, X) with the causal relation Y is the cause of X. Further, let $\mathcal{Z}$ be the set of all possible causes of X differing from Y. We call $\mathcal{Z}$ the parent set of X. The *positive rate* of effect X in the presence of cause Y and other causes $Z \in \mathbf{Z}$ is computed as

$$\lambda_{X|Y,Z} = \frac{D^w(Y)}{D(Y)} \tag{12}$$

where $w \in N^w(Y, Z \leftarrow X)$ is the preceding window of X that contains both Y and Z, $D^w(Y)$ is the duration of Y in w, and $D(Y)$ is the duration of Y throughout $\mathcal{H}$.

The *negative rate* of effect X occurring together with Z but without Y, is computed as

$$\lambda_{X|\overline{Y},Z} = \frac{D^{\overline{w}}(X)}{T - D^w(Y)} \tag{13}$$

where $\overline{w} \in N^{\overline{w}}(\overline{Y}, Z \leftarrow X)$ is the preceding window of X that contains Z but does not contain Y, $D^{\overline{w}}(X)$ is the duration of X in $\overline{w}$, while $T - D^w(Y)$ is the duration throughout $\mathcal{H}$ during which Y does not have an impact on X.

If X's rate depends on whether or not cause Y and any of other parents $Z \in \mathcal{Z}$ have occurred in the preceding window w, then there are $2^{|Z|}$ conditional intensity rates. To take into account all the different parents of X, we use an aggregation function g where g = {min, max, average} over all possible combinations of Y and Z, and compute the causality score as

$$DCIR_M(Y, X) = g\left(\frac{\lambda^w_{X|Y,Z}}{\lambda^w_{X|\overline{Y},Z}}\right) \tag{14}$$

Table 1: Dataset summary statistics

Datasets	#Samples	#Features	#Events	Sampling
Syn.(× 3)	72217	2	676	5-minute
Air (× 12)	35065	17	42	hourly
Diab. (× 70)	29183	4	25	≈6-hour

Table 2: Parameters

Params	Value
α	$[0, 0.5, 1, 2, 4, 5]$
δ	$[0, 0.25, 0.5, 0.75, 1]$
Window	$[1, ..., 144]$ (Syn.), $[1, ..., 24]$ (Air), $[1, ...4]$ (Diab.)

Theorem 3. *Let w be a time window of size n, (Y, X) be a pair of temporal events with the causal relation Y is the cause of X, and $\mathcal{Z}$ be the set of all possible causes of X differing from Y. Then for $g \in \{\min, \max, avg\}$, $DCIR_M = 1$ indicates conditional independence, and $DCIR_M \neq 1$ indicates conditional dependence of the pair (Y, X).*

From Theorem 3, the conditional independence between effect X and cause Y held if $DCIR_M = 1$, and the conditional dependence between X and Y held if $DCIR_M \neq 1$. Hence, $DCIR_M$ follows the hypothesis testing framework (equation 18, Appendix A.1).

Finding the parent set $\mathcal{Z}$ of X: An event Z that occurs earlier than X and affects the probability of X occurring is called a parent of X. We use the Proximal Graphical Event Model (PGEM) proposed by Bhattacharjya et al. (2018) to discover the parent set $\mathcal{Z}$ of X. Given an event X, PGEM finds the parent set $\mathcal{Z}$ such that $\mathcal{Z}$ maximizes the Bayesian Information Criterion (BIC) score. The BIC score indicates the optimal structural dependencies between X and its parents. We provide the pseudo code to compute Duration-based Causality Scores in Alg. 1, Appendix A.3.

Complexity: The time complexity of the DEC model is $O(E^2 \times T)$, where E is the number of temporal events and $T = |\mathcal{T}|$ is the number of time steps in $\mathcal{T}$.

4 Experimental Evaluation

4.1 Experimental Setup

Evaluation metric: We use *Hits@K*, a popular metric that has been widely adopted in other work such as Bhattacharjya et al. (2021), Luo et al. (2016) to assert causal relations. *Hits@K* counts the number of causal pairs that matches the ground truth.

$$Hits@K = |\ \#Hits\ | \text{ in top-K} \tag{15}$$

where K is the number of top-K desired results, identified by selecting the K event pairs that have highest scores, and #Hits is the number of truly causal pairs, based on the ground truth.

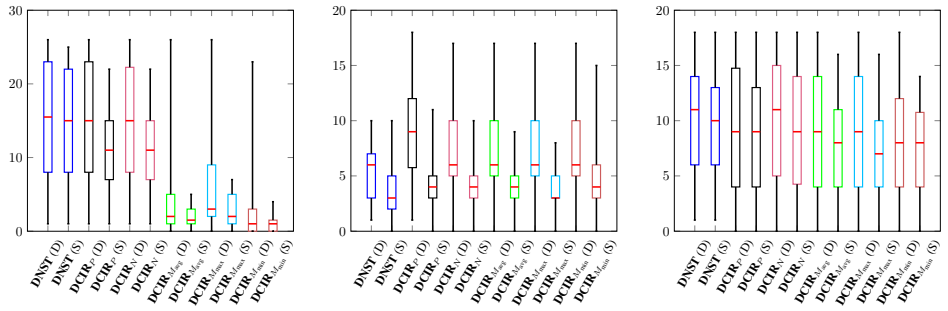
Baselines: To the best of our knowledge, the causality model proposed by Bhattacharjya et al. (2021) which adopts the sequential event modeling is the closest to our work. We name this approach Sequential Event Causality (SEC), and use SEC as the baseline to compare against our Duration-based Event Causality (DEC) model.

Datasets: We generate three synthetic datasets with known ground truths, and use two real-world datasets, Air Quality and Diabetes, in our evaluation. Table 1 summarizes the dataset statistics, with data provided in the github link. Due to space limitations, we provide the detailed descriptions and the ground truth of each dataset in Appendix A.4. Below, we discuss the obtained experimental results.

4.2 Synthetic Datasets

We compute the $Hits@K$ with $K = [1, ..., 30]$ for the synthetic datasets across all window sizes and hyper-parameters values as in Table 2, using the DEC and SEC scores. We then use a boxplot to visualize the $Hits@K$ distributions, and compare the performance of two models. Fig. 2a shows the performance of the DEC and SEC scores on the synthetic data. The DEC scores are appended with letter (D), and the SEC scores with letter (S).

From Fig. 2a, our DEC model obtains higher $Hits@K$ than SEC over all scores on the synthetic datasets. Specifically, across all window sizes, DEC achieves 56.2% higher $Hits@K$ than SEC on average in all scores. We also report in Table 3 the average $Hits@K$ across window sizes for each DEC and SEC score. Due to space limitations, we only include the results of $K = [10, 15, 25]$. From



(a) Synthetic (b) Air Quality (c) Diabetes

Figure 2: Hits@K distribution comparison between DEC (D) and SEC (S) scores

Table 3: $Hits@K$ comparison per score

Score	Dataset																	
	Synthetic						Air Quality						Diabetes					
	K=10		K=15		K=25		K=10		K=15		K=25		K=10		K=15		K=25	
	DEC	SEC	DEC	SEC	DEC	SEC	DEC	SEC	DEC	SEC	DEC	SEC	DEC	SEC	DEC	SEC	DEC	SEC
DNST	10	9.93	15	14.87	25	24.7	4.1	2.9	5.1	3.57	8.03	5.27	7.5	7.4	10.6	10	14.8	14.2
DCIR$_P$	9.9	7.73	14.87	10.57	24.6	17.13	6.87	3.43	9.87	3.97	12.73	5.53	5.2	5.2	8.8	8.8	15.4	14.4
DCIR$_N$	9.9	7.8	14.87	10.9	24.47	17.1	5.47	3.3	6.6	4.1	10.77	5.37	7.2	5.8	11	9.2	16.2	15
DCIR$_{M_{avg}}$	1.77	0.1	2.9	0.1	4.83	0.17	5.43	3.27	6.23	3.47	10.67	6.33	5.8	5.6	9.6	7.6	14.8	11.6
DCIR$_{M_{max}}$	2.5	0.17	4.43	0.43	8.13	0.53	5.47	3.17	6.5	3.33	10.83	5.23	5.6	4.8	9.2	6.6	14.2	11.6
DCIR$_{M_{min}}$	0.8	0.03	1.33	0.03	2.1	0.03	5.53	3.37	6.13	4.03	10.73	6.8	5.6	5.4	8.6	7.4	12.8	11.4

Table 4: Window size evaluation

Dataset	Best window size	Hits@K					
		DNST	**DCIR**$_P$	**DCIR**$_N$	**DCIR**$_{M_{avg}}$	**DCIR**$_{M_{max}}$	**DCIR**$_{M_{min}}$
Synthetic	2	26	26	26	26	23	26
Air Quality	1	10	18	17	17	17	17
Diabetes	1	18	18	18	18	18	18

Table 3, our DEC scores improve the $Hits@K$ by: 1% ($DNST$), 27% ($DCIR_P$), 26% ($DCIR_N$), 95.8% ($DCIR_{M_{avg}}$), 92.3% ($DCIR_{M_{max}}$), and 97.3% ($DCIR_{M_{min}}$) compared to SEC.

When comparing the DEC scores against each other as in Fig. 2a, we observe that the single-cause scores $DNST$, $DCIR_P$, and $DCIR_N$ perform better than the multi-cause score $DCIR_M$ (avg, max, and min) on the tested datasets. This is because there are more single-cause causal pairs than multi-cause causal pairs in the DAGs of the synthetic datasets. When analyzing the performance of DEC w.r.t the window sizes, we see that DEC achieves the highest $Hits@K$ with window size $w = 2$, as reported in Table 4. This is due to the defined DAGs, where the parent and child pairs are often placed one or two time steps apart, meaning that windows of size two can capture them.

4.3 Air Quality Dataset

We visualize the $Hits@K$ distributions, with $K = [1, ..., 30]$, obtained from the Air Quality dataset in Fig. 2b. It is seen that, our DEC model obtains significantly higher $Hits@K$ than SEC for all scores in this dataset. On average across all window sizes, DEC scores provide 41.3% higher $Hits@K$ than SEC scores. Table 3 reports the average $Hits@K$ ($K = [10, 15, 25]$) of each score in DEC and SEC models for the Air Quality dataset. Compared to SEC, our DEC scores improve the $Hits@K$ by: 31.2% ($DNST$), 55.4% ($DCIR_P$), 42.5% ($DCIR_N$), 41.6% ($DCIR_{M_{avg}}$), 47.5% ($DCIR_{M_{max}}$), and 36.7% ($DCIR_{M_{min}}$).

Comparing DEC scores against each other, we observe that the conditional intensity rate scores $DCIR_P$, $DCIR_N$, and $DCIR_M$ perform better than $DNST$. This implies that the conditional intensity scores are more robust than $DNST$ on the tested datasets. Among the conditional intensity scores, $DCIR_P$ has the best performance, while $DCIR_N$ performance is close to $DCIR_M$. This suggests two things: (1) a cause Y has strong impact on an effect X in this dataset (shown by $DCIR_P$ performance), and (2) multiple causes do not impact strongly on the effect as compared to a single cause (shown by $DCIR_M$ performance). When analyzing the performance of DEC w.r.t the window sizes, we see that our DEC model achieves highest $Hits@K$ with window size $w = 1$

in the Air Quality datasets, as reported in Table 4. This implies that the impact of weather conditions on air quality can clearly be seen within 1-hour window.

4.4 DIABETES DATASET

Fig. 2c visualizes the $Hits@K$ ($K = [1, ..., 30]$) distributions obtained from the Diabetes dataset, showing that our DEC model outperforms SEC. Specifically, our DEC scores achieve 10.3% higher $Hits@K$ than SEC scores on average. Table 3 also reports the average $Hits@K$ ($K = [10, 15, 25]$) of each DEC and SEC score on this dataset. Compared to SEC, our DEC scores improve the $Hits@K$ by: 3.7% ($DNST$), 2.1% ($DCIR_P$), 14.4% ($DCIR_N$), 15.3% ($DCIR_{M_{avg}}$), 20.3% ($DCIR_{M_{max}}$), and 9.5% ($DCIR_{M_{min}}$). Comparing the DEC scores against each other (Fig. 2c), we observe that $DNST$ has similar performance with $DCIR_P$, $DCIR_N$, and $DCIR_M$. Furthermore, as $DCIR_M$ performance is similar to, or slightly worse than the rest, we conclude that in this dataset, multiple causes also do not have strong impact on the effect as compared to a single cause. When analyzing DEC performance w.r.t the window sizes, we also see that the window size of one ($w = 1$) provides the best results, as reported in Table 4.

5 RELATED WORK

Causal inference for event dataset: Most of the work on causal inference in temporal data has been on time series Runge et al. (2019), Moraffah et al. (2021). However, causal models applied on time series often have limitations for event datasets, as they cannot capture the irregular and asynchronous nature of events. Existing work that considers pairwise causal associations between events has been conducted in the field of natural language processing and computational linguistics to discover implicit causal connections between text phrases Luo et al. (2016). For example, Luo et al. (2016) try to find syntax structure such as 'if A then B', together with statistical co-occurrence-based scores for discovering cause effect pairs in text. Generally, inferring causality between events is often based on the fundamental principle that *causes* change the probabilities of their *effects*. For a pair of events (y, x), y could potentially be a cause of effect x if x happens more frequently when y happens, compared to when x happens alone, i.e. $p(x \mid y) > p(x)$. Bhattacharjya et al. (2020) rely on this principle to propose a set of causality scores to infer causal associations between pairwise events. Their model works on temporal event datasets where events are modeled as points in time with sequential order. However, the model in Bhattacharjya et al. (2020) does not consider event durations, and thus, cannot capture causal relations where the causality strongly depends on how long the cause has lasted. In this present paper, we extend Bhattacharjya et al. (2020) by considering event durations in our causality model. Specifically, we adopt contemporaneous causality where the occurrences of cause and effect are modeled with time intervals, and use this new model to propose a set of duration-based causality scores to infer causal relations between events. The experimental results show that our duration-based model outperforms the sequential model in Bhattacharjya et al. (2020) using the popular evaluation metric Hits@K.

6 CONCLUSION AND FUTURE WORK

In the present paper, we propose the Duration-based Event Causality (DEC) model to infer causal relations between temporal events. Specifically, we propose a new approach to model event datasets using event time intervals, from which event durations are deduced. Based on this new event model, we propose a set of novel duration-based causality scores, including *Duration-based Necessity Sufficiency Trade-off* (DNST) and *Duration-based Conditional Intensity Rate* (DCIR$_P$, DCIR$_N$, DCIR$_M$), to infer causal relations from pairs of events in event datasets. The proposed scores utilize event durations to capture the impact of temporal duration on event causal relations. We prove that the proposed duration-based scores are sound and follow the hypothesis testing framework. Finally, we conduct an extensive experimental evaluation using synthetic and real-world datasets from the medical and environmental domains, showing that our DEC model significantly outperforms the baseline at a large margin using the popular evaluation metric Hits@K. For future work, we will extend the DEC model to quantify and estimate the treatment effect in domains such as healthcare.

REFERENCES

Massil Achab, Emmanuel Bacry, Stéphane Gaıffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate hawkes integrated cumulants. In *International Conference on Machine Learning*, pp. 1–10. PMLR, 2017.

Debarun Bhattacharjya, Dharmashankar Subramanian, and Tian Gao. Proximal graphical event models. *Advances in Neural Information Processing Systems*, 31:8136–8145, 2018.

Debarun Bhattacharjya, Karthikeyan Shanmugam, Tian Gao, Nicholas Mattei, Kush Varshney, and Dharmashankar Subramanian. Event-driven continuous time bayesian networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3259–3266, 2020.

Debarun Bhattacharjya, Tian Gao, Nicholas Mattei, and Dharmashankar Subramanian. Cause-effect association between event pairs in event datasets. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 1202–1208, 2021.

Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264, 2008.

Clive Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2(1):329–352, 1980. URL `https://EconPapers.repec.org/RePEc:eee:dyncon:v:2:y:1980:i:1:p:329-352`.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

John Hicks et al. *Causality in economics*. Australian National University Press, 1980.

Yuxiao Huang and Samantha Kleinberg. Fast and accurate causal inference from time series data. In *The twenty-eighth international flairs conference*, 2015.

Michael Kahn. Diabetes data set. URL `https://archive.ics.uci.edu/ml/datasets/diabetes`.

Yansui Liu, Yang Zhou, and Jiaxin Lu. Exploring the relationship between air pollution and meteorological conditions in china under environmental governance. *Scientific reports*, 10(1):1–11, 2020.

Aurelie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. Grouped graphical granger modeling methods for temporal causal modeling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 577–586, 2009.

Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. Commonsense causal reasoning between short texts. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2016.

Raha Moraffah, Paras Sheth, Mansooreh Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, pp. 1–45, 2021.

Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pp. e1449, 2022.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in Neural Information Processing Systems*, 26, 2013.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019.

Norbert Wiener. The theory of prediction. *Modern mathematics for engineers*, 1956.

Ting You, Renguang Wu, Gang Huang, and Guangzhou Fan. Erratum to: Regional meteorological patterns for heavy pollution events in beijing. *Journal of Meteorological Research*, 32:516–516, 2018.

S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S.X. Chen. (2017) Cautionary Tales on Air-Quality Improvement in Beijing. *Proceedings of the Royal Society A*, 473(2205):20170457, 2017.

A Appendix

A.1 Background

We discuss in this section the causality hypothesis testing framework, the necessity and sufficiency conditions, and the temporal classification of causal relations.

A.1.1 Causal Inference

Granger Causality: Causal inference is a broad subject that has not yet reached the consensus of its definite definition. The earliest concept of causality was suggested by Granger (1969), building on a notion of Wiener (1956). The idea behind Granger causality is that a stochastic process Y causes another stochastic process X if Y contains some unique information about X which is not available in X's past as well as all other information in the universe. To assert this condition, Granger causality compares the ability to predict X using all the information in the universe, denoted as U, to the ability of predicting X using all information in U except for Y, denoted as $U \setminus Y$. If discarding Y reduces the predictive power regarding X, then Y contains some unique information of X, and we thus say that Y *Granger-causes* X. Granger causality is based on two axioms when applying to data that have the time dimension such as time series Granger (1980):

Axiom A: The past and present may affect the future, but the future cannot affect the past.

Axiom B: In order for Y to be a cause of X, Y must have some unique information about X.

Based on these, *Granger causality* is formally defined as:

$$Y_t \textit{ is said to be a cause of } X_{t+1} \textit{ iff: } P\left(X_{t+1} \in \mathcal{A} | \Omega_t\right) \neq P\left(X_{t+1} \in \mathcal{A} | \Omega_t - Y_t\right) \tag{16}$$

where $\mathcal{A}$ is a non-empty set of random variables, Y_t denotes the occurrence of Y at time t, X_{t+1} denotes the occurrence of X at time $t+1$, Ω_t denotes the information available in the universe at t.

Hypothesis Testing: Another popular approach to model causality is based on the conditional independence test proposed by Judea Pearl Pearl et al. (2000). In this statistical model, Pearl measures the conditional independence as

$$P(x \mid y, z) = P(x \mid z) \textit{ whenever } P(y, z) > 0 \tag{17}$$

where $x \in X$, $y \in Y$, $z \in Z$ are the elements of the sets X, Y, Z, respectively. If Eq. equation 17 hold, then the sets X and Y are said to be conditionally independent given Z. This indicates that learning the value of Y does not provide additional information about X, once we know Z. Metaphorically, Z "screens off" X from Y Pearl et al. (2000).

Adopting the conditional independence test paradigm of Judea Pearl, Bhattacharjya et al. (2021) propose a general hypothesis testing framework to assert causal association between pairwise events as

$$\begin{aligned} H_0 &: P(X \mid Y, \mathbf{Z}) = P(X \mid \mathbf{Z}) \\ H_1 &: P(X \mid Y, \mathbf{Z}) > P(X \mid \mathbf{Z}) \end{aligned} \tag{18}$$

where X, Y, Z are events occurring in an event dataset. The pair (Y, X) is the event pair to be tested for the causal association, with the condition that Y has occurred before X, and Z is another random variable indicating whether the events in the set Z have occurred or not in the dataset.

The null hypothesis H_0 asserts whether $P(X \mid Y, \mathbf{Z})$ and $P(X \mid \mathbf{Z})$ are from the same distribution, which indicates that the presence of Y has no impact on X, conditioned on Z, and hence cannot be a cause of X. The alternative hypothesis H_1 indicates that given both Y and Z, X has higher probability than when only Z is present, implying the impact of Y on X.

In this paper, we extend the hypothesis testing framework of Bhattacharjya et al. to assert the causal association between pairwise temporal events, taking into consideration the impact of temporal duration on the strength of the causal association.

A.1.2 Necessity and Sufficiency Causalities

When asserting causality, *necessity* and *sufficiency* are two distinct cases of a causal association. Consider a causal pair (Y, X) with Y being the cause of X. Granger (1969) formally defines the two cases as:

Necessity: if cause Y occurs, then the probability of effect X occurring increases.

Sufficiency: if effect X is observed, then cause Y likely has occurred.

The necessity causality encoded by the pair (Y, X) represents the case where the presence of the cause Y is the condition for the effect X to take place. On the other hand, the sufficiency causality encoded by (Y, X) demonstrates the case where the presence of the effect X provides the evidence for the presence of the cause Y.

By distinguishing the two cases of *necessity* and *sufficiency*, we can assert the roles of cause and effect in a causal association separately. Intuitively, the stronger the necessity causality is, the larger the probability $p(X \mid Y)$ should be; and the stronger the sufficiency causality is, the larger the probability $p(Y \mid X)$ should be. For example, the causal pair (rainfall, flooding) with rainfall being the cause of flooding encodes more necessity causality than sufficiency causality, since in most situations the effect flooding cannot happen if rainfall did not happen as its cause. Similarly, the causal pair (storm, thunder) encodes more sufficiency causality.

In this paper, we adopt the notions of *necessity* and *sufficiency* causalities to assert the roles of cause and effect in a causal association between two temporal events.

A.1.3 Temporal Classification of Causal Relations

When studying causality in economics, British economist John Hicks proposed a taxonomy of causal relationships based on temporal representation of events Hicks et al. (1980). Specifically, based on the time dimension of the cause and the temporal relationship between cause and effect, Hicks defined three causal relations: static, contemporaneous, and sequential which correspond to three perspectives on time: eternity, a period of time, and a point in time.

Static causality describes causal relations in which both cause and effect are eternal, resulting in causal relations that are permanent through time. Examples of static causality are astronomical phenomena, which from the human perspective appear indefinite.

Contemporaneous causality describes causal relations where both cause and effect span a finite period of time, resulting in causal relations that are hold only through some time periods. An example of contemporaneous causality is the causal pair (rainfall, flooding) where rainfall and flooding are measured in time periods, and last within their respective time intervals.

Finally, *sequential causality* is defined when cause and effect are measured as points in time, and thus, cause has to precede effect for the causality to be hold.

In Bhattacharjya et al. (2021), Bhattacharjya et al. adopt the sequential causality to represent events as points in time, and assert their causal relations accordingly. In this paper, we instead adopt the contemporaneous causality to represent cause and effect over time intervals. This event model enables us to study the impact of temporal duration on the strength of causal relations between temporal events.

A.2 Detail proofs of Theorems

We provide in this section the detail proofs of all theorems.

Theorem 1. *Let w be a time window of size n, and (Y, X) be a pair of temporal events with the causal relation Y is the cause of X. Then $DNST(Y, X) = 0$ indicates independence, and $DNST(Y, X) > 0$ indicates dependence of the pair (Y, X).*

Proof. If $DNST(Y, X) = 0$, it follows from Eq. equation 6 that $p^w(Y \leftarrow X) = 0$, or $D^w(Y) = 0$, or $p^w(Y \rightarrow X) = 0$, or $D^w(X) = 0$. In the first two cases, there are no preceding time windows w of X that contain Y; while in the last two cases, there are no succeeding time windows w of Y that contain X. This indicates the independence of the pair (Y, X). Similarly, if $DNST(Y, X) > 0$, all quantities $p^w(X|Y)$, $D^w(Y)$, $p^w(Y \leftarrow X)$ and $D^w(X)$ are positive, indicating the dependence of the pair (Y, X).

□

Theorem 2. *Let $DCIR_{P/N}$ be either $DCIR_P$ or $DCIR_N$. Let w be a time window of size n, and (Y, X) be a pair of temporal events with the causal relation Y is the cause of X. Then $DCIR_{P/N} = 1$ indicates independence, and $DCIR_{P/N} \neq 1$ indicates dependence of the pair (Y, X).*

Proof. First, we provide the proof for the positive score $DCIR_P$. Let us assume $p(X \mid Y, Z) = p^{dt}(X \mid Y) = \lambda_{X|Y}dt$, and $p(X \mid Z) = p^{dt}(X) = \lambda_X dt$, where Z is a random variable. In this setting, if $p(X \mid Y, Z) = p(X \mid Z)$, then $\lambda_{X|Y}dt = \lambda_X dt$, implying that $\frac{\lambda_{X|Y}}{\lambda_X} = 1$. Hence, $DCIR_P = 1$ indicates independence. Else, if $p(X \mid Y, Z) \neq p(X \mid Z)$, then $\frac{\lambda_{X|Y}}{\lambda_X} \neq 1$. Hence, $DCIR_P \neq 1$ indicates dependence.

The proof for the negative score is similar by considering the relation between two conditional probabilities $p(X \mid Y, Z)$ and $p(X \mid \overline{Y}, Z)$. □

Theorem 3. *Let w be a time window of size n, (Y, X) be a pair of temporal events with the causal relation Y is the cause of X, and $\mathcal{Z}$ be the set of all possible causes of X differing from Y. Then for $g \in \{\min, \max, avg\}$, $DCIR_M = 1$ indicates conditional independence, and $DCIR_M \neq 1$ indicates conditional dependence of the pair (Y, X).*

Proof. For any $Z \in \mathcal{Z}$, assume $p(X \mid Y, Z) = p^{dt}(X \mid Y, Z) = \lambda_{X|Y,Z}dt$, and $p(X \mid Z) = p^{dt}(X \mid Z) = \lambda_{X|Z}dt$. In this setting, if $p(X \mid Y, Z) = p(X \mid Z)$, then $p(X \mid Y, Z) = p(X \mid Z) = p(X \mid \overline{Y}, Z)$, i.e. $\lambda_{X|Y,Z}dt = \lambda_{X|Z}dt = \lambda_{X|\overline{Y},Z}dt$, implying $\frac{\lambda_{X|Y,Z}}{\lambda_{X|\overline{Y},Z}} = 1$. It follows that $g\left(\frac{\lambda_{X|Y,Z}}{\lambda_{X|\overline{Y},Z}}\right) = 1$ for $g \in \{\min, \max, \text{avg}\}$. Hence $DCIR_M = 1$ indicates conditional independence. Else, if $p(X \mid Y, Z) \neq p(X \mid Z)$, then $\frac{\lambda_{X|Y,Z}}{\lambda_{X|\overline{Y},Z}} \neq 1$. It follows that $g\left(\frac{\lambda_{X|Y,Z}}{\lambda_{X|\overline{Y},Z}}\right) \neq 1$. Hence, $DCIR_M \neq 1$ indicates conditional dependence. □

A.3 Pseudo code to compute Duration-based Causality Scores

We provide the pseudo code to compute Duration-based Causality Scores in Algorithm 1. First, we obtain the list of windows of size n (line 1). Next, we apply PGEM to find the parent set $\mathcal{Z}$ of X (line 2). We iterate through the window list to calculate window-based statistics (lines 3-6). The event statistics are computed in line 7. Finally, we compute the DEC scores in lines 8-11. We note that for each event pair (Y, X), the window-based statistics, e.g., $D^w(X), D^w(Y)$, and the event statistics, e.g., $P(X)$, $P(Y)$ can be computed in a one-pass scan of the event dataset $\mathcal{D}$.

Algorithm 1 Duration-based Causality Scores

Input: $\mathcal{D}$: temporal event dataset
(Y, X): pair of temporal events
Params: α: penalization parameter,
δ: trade-off parameter,
n: window size
Output: $<,,,>$: tuple of four duration-based causality scores for (Y, X)

$W \leftarrow$ getWindows(n, $\mathcal{D}$) ▷ get a list of windows of size n
$\mathcal{Z} \leftarrow$ PGEM($X, \mathcal{D}$) $\setminus \{Y\}$ ▷ find parents of X
for $w \in W$ **do**
  $p^w(Y \leftarrow X)$, $p^w(Y \rightarrow X)$, $D^w(Y)$, $D^w(X)$, $D^{\overline{w}}(X)$
    $\leftarrow$ calcSingleCauseScoreStatistics((Y, X), n, $\mathcal{D}$)
  $p_z^w(Y \leftarrow X)$, $p_z^w(Y \rightarrow X)$, $D_z^w(Y)$, $D_z^w(X)$, $D_z^{\overline{w}}(X)$
    $\leftarrow$ calcMultiCauseScoreStatistics((Y, X), $\mathcal{Z}$, n, $\mathcal{D}$)
end for
$D(Y)$, $p(Y)$, $D(X)$, $p(X)$, $\lambda_X \leftarrow$ calcEventStatistics((Y, X), $\mathcal{D}$)
DNST$\leftarrow$ calcDNST(α, δ, $p^w(Y \leftarrow X)$, $p^w(Y \rightarrow X)$,
  $D^w(Y)$, $D^w(X)$, $p(X)$, $p(Y)$) ▷ Eq. equation 6
DCIR$_P \leftarrow$ calcDCIR$_P(D^w(Y), D(Y), \lambda_X)$ ▷ Eq. equation 11
DCIR$_N \leftarrow$ calcDCIR$_N(D^w(Y), D(Y), D^w(Y), D^{\overline{w}}(X))$ ▷ Eq. equation 11
DCIR$_M \leftarrow$ calcDCIR$_M(D_z^w(Y), D(Y), D_z^w(Y), D_z^{\overline{w}}(X))$ ▷ Eq. equation 14
return <DNST, DCIR$_P$, DCIR$_N$, DCIR$_M$>

A.4 Dataset Description and Ground Truth

A.4.1 Synthetic Dataset

Data generation: We generate three synthetic datasets, each has 26 known causal event pairs. To define the ground truths, we use a Directed Acyclic Graph (DAG) to encode the causal relations between events. In a DAG $G(V, E)$, V is the set of nodes representing event labels, and E is the set of weighted edges connecting two nodes. A truly causal pair between two events is encoded as a parent-child relation, represented by an edge where the weight is the conditional probability $p(child \mid parent)$, indicating how likely a child is generated given the parent. Fig. 3 shows an example of a DAG representing the ground truths. Given a DAG, our synthetic data generator iterates through a predefined number of samples to generate the event datasets. During the process, each time a parent occurs, we will generate a child if its accumulated conditional probability from previous iterations reaches a predefined threshold θ. Other events that are not included in the DAG will be generated randomly.

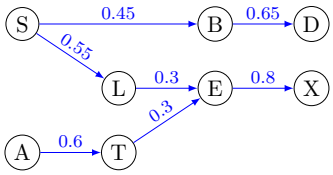


Figure 3: Example of ground truths defined in a DAG

A.4.2 Air Quality

Dataset description: The Beijing Multi-Site Air-Quality Dataset Zhang et al. (2017) collects air pollution data from 12 different sites in Beijing, China, including air quality indicators such as Particulate Matter, PM10 and PM2.5 concentration, and meteorological data such as temperature, humidity, air pressure. Each dataset consists of 17 features, and ≈ 35000 records.

Ground truth. We base the ground truth on expert opinions, in particular, the work of You et al. (2018) and Liu et al. (2020) which study the impact of meteorological events on air quality indicators in Dongsi, China, the area covered by the dataset we used. The quantitative impact is presented in

Table 5, showing how each meteorological variable increases or decreases the concentration of air quality indicators.

To identify the ground truth, we follow the procedure described in Liu et al. (2020) as follows. For each (air quality, weather) event pair, we compute the predicted value of air quality using the weather coefficients in Table 5. If the predicted value is within the event range, the causal pair is deemed true. For example, using the average PM10 intensity, the low wind speed value and the wind coefficient, we compute a predicted PM10 value. If this value belongs to high PM10 concentration range, then the pair (High PM10, Low Wind Speed) is a truly causal pair. We obtain 18 truly causal pairs from this dataset.

Table 5: Quantifiable data for Air Quality dataset

Variable	$\mathbf{PM_{2.5}}$	$\mathbf{PM_{10}}$	**CO**	$\mathbf{NO_2}$	$\mathbf{O_3}$	$\mathbf{SO_2}$
WS	–0.155	–0.220	–0.002	–0.122	0.365	–0.116
Pre	–0.083	–0.097	0.000	–0.049	–0.089	0.042
AP	0.053	0.139	–0.001	0.138	–0.289	–0.056
Temp	–1.429	–1.486	–0.023	–0.519	2.626	–0.849
RH	0.087	–0.587	0.003	–0.001	–0.813	–0.181

**Notes: WS: Wind speed, Pre: Precipitation, AP: Atmosphere pressure, Temp: Temperature, RH: Relatively humidity

A.4.3 DIABETES

Dataset description: Finally, we evaluate our scores on a clinical dataset which describes the basic physiology and patho-physiology of diabetes mellitus and its treatment Kahn. The dataset contains data of 70 diabetic patients with measurements done at recurring daily events like meals, insulin injection doses, etc.

Ground truth. This dataset is provided with a description document, written by medical experts to provide the ground truth for treatment of diabetic patients. We rely on this expert knowledge, reported in Table 6 to evaluate our scores.

Table 6: Expert knowledge for Diabetes dataset

Treatment & Activity	**Blood glucose (BG)**
Regular Insulin	O (15-45 M), P (1-3 H), D (4-6 H)
NPH Insulin	O (1-3 H), P (4-6 H), D (10-14 H)
Ultralente	O (2-5 H), P (NA), D (24-30 H)
Moderate exercise	Reduce BG
Strenuous exercise/ mild dehydrate	Transient increase in BG
Large meal	High BG
Missing or smaller meals	Low BG
Hypoglycemic symptoms	Low BG

**Notes: O: Onset effect, P: Time of peak action, D: Effective duration, M: Minute, H: Hour, NA: not much of a peak