Granger Causal Chain Discovery for Sepsis-Associated Derangements via Continuous-Time Hawkes Processes

Song Wei
Georgia Institute of Technology
Atlanta, GA, USA
song.wei@gatech.edu

Yao Xie
Georgia Institute of Technology
Atlanta, GA, USA
yao.xie@isye.gatech.edu

Christopher S. Josef
Emory University
Atlanta, GA, USA
cjosef@emory.edu

Rishikesan Kamaleswaran
Emory University
Atlanta, GA, USA
rkamales@dbmi.emory.edu

ABSTRACT

Modern health care systems are conducting continuous, automated surveillance of the electronic medical record (EMR) to identify adverse events with increasing frequency; however, many events such as sepsis do not have elucidated prodromes (i.e., event chains) that can be used to identify and intercept the adverse event early in its course. Clinically relevant and interpretable results require a framework that can (i) infer temporal interactions across multiple patient features found in EMR data (e.g., Labs, vital signs, etc.) and (ii) identify patterns that precede and are specific to an impending adverse event (e.g., sepsis). In this work, we propose a linear multivariate Hawkes process model, coupled with ReLU link function, to recover a Granger Causal (GC) graph with both exciting and inhibiting effects. We develop a scalable two-phase gradient-based method to obtain a maximum surrogate-likelihood estimator, which is shown to be effective via extensive numerical simulation. Our method is subsequently extended to a data set of patients admitted to Grady hospital system in Atlanta, GA, USA, where the estimated GC graph identifies several highly interpretable GC chains that precede sepsis. The code is available at https://github.com/SongWei-GT/two-phase-MHP.

CCS CONCEPTS

• **Mathematics of computing** → **Time series analysis**; • **Applied computing** → **Health care information systems**.

KEYWORDS

Continuous-time event data, Electronic medical record, Gradient-based approach, Granger Causality, Multivariate Hawkes process

ACM Reference Format:
Song Wei, Yao Xie, Christopher S. Josef, and Rishikesan Kamaleswaran. 2023. Granger Causal Chain Discovery for Sepsis-Associated Derangements via Continuous-Time Hawkes Processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3580305.3599369

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599369

1 INTRODUCTION

Continuous, automated surveillance systems that use machine learning models to identify adverse patient events are being incorporated into healthcare environments with increasing frequency. One of the most notable adverse events is sepsis, a life-threatening medical condition contributing to one in five deaths globally [46] and stands as one of the most important cases for automated in-hospital surveillance. Sepsis is formally defined as life-threatening organ dysfunction caused by a dysregulated host response to infection [41]. Delays in recognizing sepsis and initiating appropriate treatment can adversely impact patient outcomes. In a recent study of adult sepsis patients, each hour of delayed treatment was associated with higher risk-adjusted in-hospital mortality (odds ratio, 1.04 per hour) [38]. It logically follows that early recognition of the physiologic aberrations preceding sepsis would afford clinicians more time to intervene and may contribute to improving outcomes and reducing costs. Many machine learning methods have been developed to predict the onset of sepsis, utilizing data from the electronic medical record (EMR) [13, 34, 39]. While many approaches can be designed to provide an alert preceding an event, most are not designed to discover and report the causal chains that preceded an adverse event. Developing and reporting a causal chain of events not only serves as a foundation for prognosticating adverse event occurrence, but more importantly it reveals the pathways of deterioration which may afford clinicians the additional context to corroborate or modify existing treatment modalities in a way that is superior to a simple alarm.

Recently, Hawkes processes [19–21], which model self- and mutual- exciting patterns among continuous-time events, have drawn much attention in the field of health analytics [2, 8, 31, 37, 45]. The linear multivariate Hawkes process (MHP) seems highly relevant to our problem since (i) the support of the excitation matrix enjoys a natural interpretation as a Granger Causal (GC) graph [47], (ii) given its interpretation as a clustering process [21], we can infer the commonly observed chain pattern that precedes sepsis from the estimated GC graph, and (iii) with proper domain expertise, simple methods, such as (generalized) linear model, are proven effective in outputting highly explainable results [8, 45].

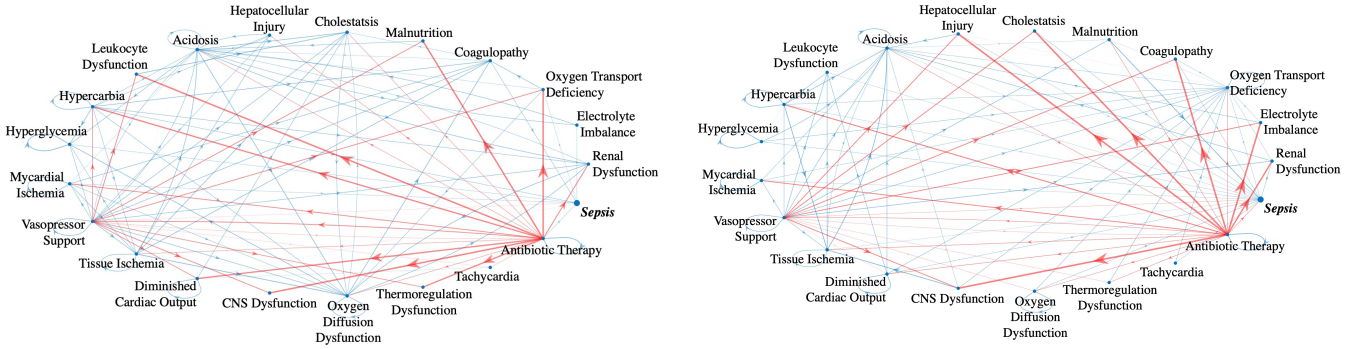

Figure 1: GC graphs over SADs for Sepsis-3 cohort (left) and full patient cohort (right). The width of the directed edge is proportional to the exciting (blue) or inhibiting (red) effect magnitude. We can observe that our proposed method can output highly interpretable GC graphs; for example, the observation that Antibiotic Therapy inhibits most of the SADs agrees with the well-known physiologic relationship.

However, there are two major challenges preventing us from applying naive linear MHP to recover the GC graph. First and foremost, linear MHP itself fails to model inhibiting effects (e.g., proper medication will inhibit the occurrence of a certain disease), since "negative triggering effects" could lead to a negative conditional intensity and thus intractable likelihood. Second, the well-established expectation–maximization (EM) stochastic declustering algorithm [14, 51] sufferers from scalability issue and cannot be applied to EMR data with thousands of patients' trajectories. Recently, Bonnet et al. [3, 4] proposed a linear MHP coupled with ReLU link function $g(x) = x^+ := \max\{0, x\}$ to handle the potential inhibiting effects. To evaluate and maximize the likelihood, they calculated the "re-start time", at which the conditional intensity becomes nonzero. However, such a calculation has quadratic complexity, making it unscalable. Scalable methods to infer GC graph with both exciting and inhibiting effects for linear MHP are still largely missing.

In this paper, we adopt the ReLU link function in linear MHP to recover a Granger Causal graph with both exciting and inhibiting effects. We propose a maximum surrogate-likelihood formulation to tackle the scalability issue caused by the re-start time calculation [4]. Furthermore, we develop a two-phase gradient-based method to solve the optimization problem, and we observe improved empirical performance through extensive numerical simulation. Most importantly, our method can output graphs (i.e., Figure 1) that afford clinicians a simple mechanism for interpreting both promoting and inhibitory causal relationship amongst the data — Networks are exceptionally important for syndromic (i.e. a constellation of different physiologic derangements can be manifested) conditions like sepsis. These graphs can be used to differentiate cohorts and to identify important, intra-cohort relationships. For clinicians the utility of these graphs is two-fold: they can be used to (i) quantify a patient's risk of developing subsequent physiologic derangements in the future and (ii) discover new relationships. The estimated GC graphs here are highly interpretable and can be used to create or augment surveillance systems for high-risk patients. Here, we demonstrate the effectiveness of our approach in learning a Granger Causal graph for Sepsis Associated Derangements (SADs), but it can be generalized to other applications with similar requirements.

Related work. Granger Causality is well-studied in time series literature via the vector autoregressive (VAR) model; see Shojaie and Fox [40] for a recent survey. VAR models and MHP models share many similarities and some have recently recognized that the self- and mutual-excitation matrix in the Hawkes process model can be interpreted as Granger Causal graph in a similar way. The study of GC under the context of MHP can be traced back to Kim et al. [26]. Recent development includes leveraging the alternating direction method of multipliers to infer the low-rank structure in mutual excitation matrix [50], applying EM algorithm with various constraints [7, 22, 47] and using powerful neural networks [48] to infer the GC graph.

Even outside the context of Granger Causality, the Hawkes process itself has drawn much attention recently — there have been many (semi-)parametric Hawkes process models by considering different types of triggering kernel function, such as probability weighted kernel estimation with adaptive bandwidth [51], probability weighted histogram estimation [29] and with inhomogeneous spatial background rate [14] and so on. In addition, there are also many non-parametric methods, e.g., the Neural Hawkes process [30] and the Transformer Hawkes process [52].

Despite those advancements in semi- and non-parametric Hawkes process models, Choi et al. [8], Wei et al. [45] showed that simple linear models can output meaningful results in practice. However, the state-of-the-art method is the stochastic declustering algorithm, which is based on the EM algorithm and is thus highly unscalable. This scalability issue makes it a less desirable option when we handle EMR data. Recently, there are attempts to explore the powerful yet simple gradient-based method to infer the problem parameters; notable contributions include Cartea et al. [6], Wang et al. [44]. In particular, we want to mention that using the ReLU link to allow potential inhibiting effect in linear MHP was recently proposed by Bonnet et al. [3, 4] and relatively novel in literature — there have not been many methods tailored to this particular parameterization, and thus we only numerically compare our method with this re-start time method as well as some naive gradient-based methods.

Another closely related topic is causal discovery, which has drawn much attention in the past few decades. The state-of-the-art constraint-based algorithms include PC and Fast Causal Inference (FCI) [42]. Both algorithms can output the underlying true graph structure in the large sample limit. However, PC cannot deal with unobserved confounding whereas FCI is capable of dealing with

confounders. However, since those algorithms rely on conditional independence tests to eliminate edges from the complete graph, they are not scalable when the number of nodes becomes large. Existing work to handle this includes a fast and memory-efficient PC algorithm using the parallel computing technique [27]. Moreover, there is a continuous optimization-based approach to infer the underlying directed acyclic graph (DAG) structure, e.g., Zheng et al. [49], which alleviates the aforementioned scalability issue. In addition, for time series data, existing causal discovery algorithms need to adapt to the potential temporal dependence. The most well-known method would be using AR time series to infer the Granger Causality, and later on, Xu et al. [47] extend GC to the context of the point process. However, the GC framework typically relies on the "no unobserved confounding" assumption. Examples to handle this include the FCI algorithm for time series to handle confounders [11]. It remains an open problem how to apply PC and FCI to point process data. For a complete survey on recent developments in causal inference, we refer readers to Glymour et al. [15].

2 BACKGROUND

2.1 Multivariate Hawkes Process

Consider d types events modeled by a counting process $N = (N^1, \ldots, N^d)$, where each process $N^i = \{N_t^i : t \in [0, T]\}$ itself is a counting measure on time horizon T and records the number of type-i events before time t. Such a process is called a linear MHP if the conditional intensity of i-th process ($i = 1, \cdots, d$) is defined as:

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{d} \int_0^t \varphi_{i,j}(s)\, dN_{t-s}^j,$$

where μ_i is the exogenous background intensity for type-i event and independent of the history, and kernel function $\varphi_{i,j}(\cdot)$ captures the impact from historical type-j event to subsequent type-i event.

Here, we adopt a very common and popular exponential kernel function $\varphi_{i,j}(t) = \alpha_{i,j} \exp\{-\beta t\}$. The parameter $\alpha_{i,j}$ represents the magnitude of the impact from type-j event to type-i event and β characterizes the rate of decay of that impact. Most importantly, unlike the classic model, we consider both exciting and inhibiting effects by allowing negative magnitude parameters $\alpha_{i,j}$'s. However, this could lead to negative intensity, which contradicts the understanding of conditional intensity as the instantaneous probability of event occurrence. Following Bonnet et al. [3], we apply the *ReLU link function* $(\cdot)^+ = \max\{0, \cdot\}$ to the linear conditional intensity to fix this issue and get

$$\lambda_i(t) = \left(\mu_i + \sum_{j=1}^{d} \int_0^t \alpha_{i,j} e^{-\beta s} dN_{t-s}^j\right)^+. \tag{1}$$

We denote the background intensity vector as $\mu = (\mu_1, \ldots, \mu_d)^T$ and the self and mutual excitation/inhibition matrix as $A = (\alpha_{i,j}) \in \mathbb{R}^{d \times d}$. We will show the support of matrix A can be interpreted as a Granger Causal graph.

2.2 Granger Causality

In the seminal paper, Eichler et al. [9] showed that the Granger Causal structure of the MHP is fully encoded in matrix A:

Proposition 2.1 (Eichler et al. [9]). *Let $N = (N^1, \ldots, N^d)$ be a d-dimensional multivariate Hawkes process with conditional intensity defined in* (1)*, then N^j does NOT Granger-cause N^i if and only if $\alpha_{i,j} = 0$.*

We need to remark that inferring Granger Causality needs "all the information in the universe" [16–18]. In the graph induced by the matrix $A = (\alpha_{i,j})$, the absence of an edge means Granger non-causality whereas only when there is no unobserved confounding can the presence of an edge in A imply Granger causality. Here, we assume there is *no unobserved confounding* and we will take this matrix A as the Granger Causal graph.

3 ESTIMATION

Consider the following continuous-time event data over a time horizon $T > 0$:

$$(u_1, t_1), \ldots, (u_N, t_N),$$

where $0 \le t_1 < \cdots < t_N \le T$ denote the exact occurrence times of the events and $u_n \in \{1, \ldots, d\}$ represents the type of the n-th event. The conditional intensity function of type-i event at time $0 \le t \le T$ is as follows:

$$\lambda_i(t) = \left(\mu_i + \sum_{j: t_j < t} \alpha_{i,u_j} e^{-\beta(t - t_j)}\right)^+.$$

Typically, we use the Maximum likelihood estimation (MLE) to learn model parameters, where the *true log likelihood* is:

$$\ell(\mu, A; \beta) = \sum_{i=1}^{d} \left(\int_0^T \log \lambda_i(t) dN_t^i - \int_0^T \lambda_i(t) dt\right). \tag{2}$$

3.1 Existing method

In (2), the first term reduces to a summation over the log-intensities on event occurrence times $\sum_{n=1}^{N} \log \lambda_{u_n}(t_n)$, which will be well-defined since the conditional intensity at the event occurrence time will be positive. To be precise, the feasible region is

$$\Theta = \{(\mu, A) : \tilde{\lambda}_{u_n}(t_n) > 0,\ n = 1, \ldots, N\}, \tag{3}$$

where the *surrogate conditional intensity* is defined as:

$$\tilde{\lambda}_i(t) = \mu_i + \sum_{j: t_j < t} \alpha_{i,u_j} e^{-\beta(t - t_j)}. \tag{4}$$

After each event occurrence, due to the potential inhibiting effect, there could be an event with negative surrogate intensity; the ReLU link enforces such negative value to be zero and ensures that $\lambda_i(t) = (\tilde{\lambda}_i(t))^+$ is still a valid intensity. Nevertheless, it still takes some time for the process to "re-start", and we will call the time when the surrogate intensity increases to zero again as the "re-start time"; see a graphical illustration in Figure 9 in the appendix. To be precise, after the occurrence of n-th event (u_n, t_n), the n-th re-start time for i-th process is as follows [3, 4]:

$$T_{(n,u_n)}^{(i)} = \min\left\{t_{n+1},\ \arg\min_{t:\ t > t_n} \tilde{\lambda}_i(t) \ge 0\right\}. \tag{5}$$

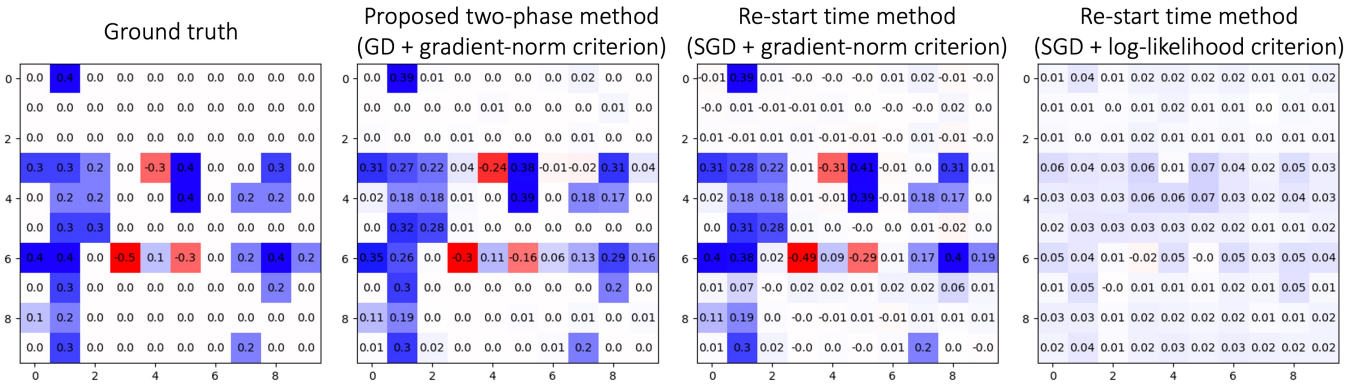

Figure 2: Comparison of the estimated adjacency matrices for a $d = 10$ illustrative example; the corresponding method is specified on the top of each panel. We can observe that relaxing the feasibility constraint and using the gradient-norm as the stopping criterion can improve the estimation accuracy compared with the conventional likelihood criterion; further quantitative comparison can be found in Table 4.

Then, for $(\mu, A) \in \Theta$, we can re-write (2) into:

$$\ell(\mu, A; \beta) = \sum_{n=1}^{N} \log \tilde{\lambda}_{u_n}(t_n) - \sum_{i=1}^{d} \left(\int_0^{t_1} + \sum_{n=1}^{N-1} \int_{T^{(i)}_{(n,u_n)}}^{t_{n+1}} + \int_{T^{(i)}_{(N,u_N)}}^{T} \right) \tilde{\lambda}_i(t) dt. \tag{6}$$

Now, we remove the non-differentiable ReLU link in the log likelihood and the objective becomes differentiable. Despite its complicated form, the log likelihood objective can be calculated in closed-from due to the analytical expression of the re-start times [4]. Thus, we can leverage the powerful stochastic gradient descent (SGD) method to numerically solve the MLE.

3.2 Proposed gradient-based method

3.2.1 Empirical challenge. The difficulty of applying gradient descent (GD) comes from the optimization landscape — the log likelihood can become *intractable*, i.e., the GD iterate could go outside the feasible region Θ, especially when the it is close to the empirical optimizer as the empirical optimizer often lies on the edge of the feasible region (see Figure 3 for illustration), and the log likelihood will no longer be well-defined, rendering us unable to accurately track or maximize the likelihood to learn the problem parameters. This suggests that naively applying GD will result in a highly *unstable* procedure (as verified by Figure 7 in the appendix). Moreover, searching for the empirical optimizer within the feasible region based on the log likelihood criterion may not be the best option — indeed, our empirical findings from Figure 8 in the appendix show that, even when the log likelihood becomes intractable, the estimation error continues to decrease when using the matrix Frobenius norm (F-norm) of the gradient with respect to (w.r.t.) adjacency matrix as the stopping criterion (referred to as the *gradient-norm criterion* below), suggesting that we could relax the *feasibility constraint* $(\mu, A) \in \Theta$ and use gradient-norm criterion instead of the log likelihood one. To support this claim, we use SGD to solve for MLE within the feasible region (3) and report the estimated adjacency matrix in the last panel in Figure 2. In comparison, we relax the feasibility constraint and use the gradient-norm criterion. We report the resulting estimated A in the third panel of Figure 2, and we can see the estimation is more accurate when we use gradient-norm criterion compared with the conventional log likelihood criterion.

Table 1: Complexity analysis of different estimation methods; d denotes the dimensionality and N is the number of events. Since there is no adaptation of EM algorithm [47] to handle the instability issue as illustrated in Figure 3, the gradient evaluation of EM is left blank.

	EM	Re-start time	Proposed
Number of parameters	$O(N^2 + d^2)$	$O(dN + d^2)$	$O(d^2)$
Gradient evaluation	–	$O(dN^2)$	$O(N^2 + dN)$

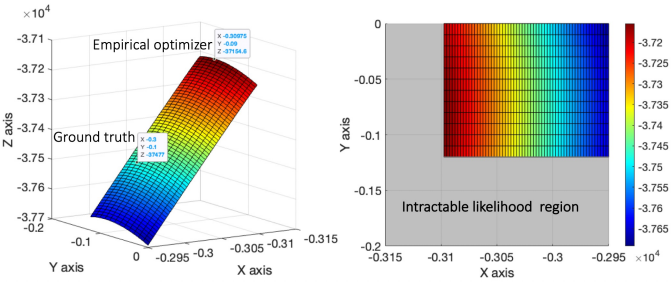

Figure 3: Optimization landscape for a $d = 3$ example; the X, Y and Z axes correspond to α_{13}, α_{21} and the log likelihood, respectively. "No Z value for pair (X, Y)", which is the grey region in the right panel, means the log likelihood becomes intractable for the corresponding $(\alpha_{13}, \alpha_{21})$ pair. We can see the empirical optimizer lies on the border of the intractable likelihood region. Complete details of this illustrative example can be found in Appendix A.1.

3.2.2 A maximum surrogate likelihood formulation. Another practical issue comes from the re-start time (5), which needs to be re-calculated after each iteration, making it highly non-scalable; see Table 1 for the complexity analysis. To alleviate this scalability issue caused by the re-start time calculation while harvesting the

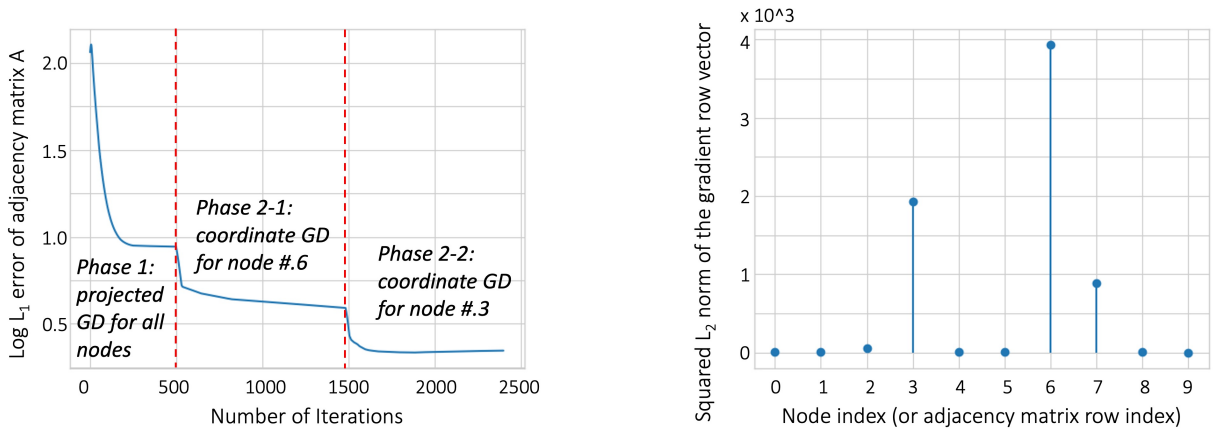

Figure 4: Illustration of the proposed two-phase method. After Phase 1, we select nodes # 3 and # 6 based on gradient-norm (see the right panel; the percentage threshold $p = 0.85$ in Algorithm 2) and proceed to phase 2, where we perform GD without projection for selected nodes (see the left panel for illustration and evidence of convergence).

empirical good performance of (6), we propose to maximize the following *surrogate log likelihood*:

$$\tilde{\ell}(\mu, A; \beta) = \sum_{n=1}^{N} \log \tilde{\lambda}_{u_n}(t_n) - \sum_{i=1}^{d} \int_0^T \tilde{\lambda}_i(t)dt \tag{7}$$
$$= \sum_{i=1}^{d} \sum_{j=1}^{N-1} \frac{\alpha_{i,u_j}}{\beta} \left(e^{-\beta(t_N - t_j)} - 1\right) - T \sum_{i=1}^{d} \mu_i$$
$$+ \sum_{n=1}^{N} \log \left(\mu_{u_n} + \sum_{j<n} \alpha_{u_n,u_j} e^{-\beta(t_n - t_j)}\right),$$

which serves as a computationally stable and efficient approximation to the true log likelihood (6). The above surrogate, which differs from true log likelihood in the integration region in the second term, is a computationally friendly and differentiable approximation to the true log likelihood, which can be understood as either (i) replacing the true conditional intensity with its differentiable surrogate (4) or (ii) ignoring the re-start time calculation and integrating the surrogate intensity on $[0, T]$. We will maximize this surrogate log likelihood to estimate the problem parameters, i.e.,

$$\hat{\mu}, \hat{A} = \text{argmin}_{(\mu,A)\in\Theta} - \tilde{\ell}(\mu, A; \beta). \tag{8}$$

3.2.3 A two-phase gradient descent algorithm. Since the objective function (8) is convex w.r.t. (μ, A) [1], projected gradient descent (PGD) is a tempting choice, which enjoys a strong convergence guarantee. However, despite the above simple closed-form expression, the projection back to Θ to maintain feasibility is computationally intense, making PGD again unscalable. Fortunately, the gradient field of this surrogate remains well-defined even outside the feasible region Θ, making the vanilla GD possible. However, vanilla GD (without projection) will suffer from divergence issues, as the iterate can easily go outside the feasible region Θ (see Figure 7 in the appendix for empirical evidence). Thus, we need to gradually decay the learning rate during the learning process. Since the (surrogate) log likelihood is also intractable, it cannot be used to fulfill this purpose. To tackle those difficulties, we propose a two-phase GD-based method coupled with a learning rate decaying scheme based on the gradient-norm; this method is illustrated in Figure 4 and one can see its good performance in the second panel in Figure 2. Next, we will briefly introduce this algorithm.

Phase 1: Projected Gradient Descent. In the first phase, we constrain all parameters to be non-negative and perform projected GD with fixed step length. We denote $\hat{\mu}_t$ and $\hat{A}_t$ to be the iterates at t-th step, and the update rule is as follows:

$$\hat{\mu}_t \leftarrow \hat{\mu}_{t-1} + \gamma \nabla_\mu \tilde{\ell} / \|\nabla_\mu \tilde{\ell}\|_2, \quad \hat{A}_t \leftarrow \hat{A}_{t-1} + \gamma \nabla_A \tilde{\ell} / \|\nabla_A \tilde{\ell}\|_F,$$

where γ is the learning rate, $\|\cdot\|_2$ and $\|\cdot\|_F$ represent vector L_2 norm and matrix Frobenius norm, respectively, and the gradient fields are defined as:

$$\nabla_\mu \tilde{\ell} = \nabla_\mu \tilde{\ell}(\hat{\mu}_{t-1}, \hat{A}_{t-1}; \beta), \quad \nabla_A \tilde{\ell} = \nabla_A \tilde{\ell}(\hat{\mu}_{t-1}, \hat{A}_{t-1}; \beta).$$

The parameter β is assumed to be known; in practice, we will perform a grid search to select the best β. To make sure we do not get negative intensity, we perform the following projection:

$$\hat{\mu}_t \leftarrow \text{argmin}_{\mu\in\mathbb{R}_+} \|\hat{\mu}_t - \mu\|_2, \quad \hat{A}_t \leftarrow \text{argmin}_{A\in\mathbb{R}_+^{d\times d}} \|\hat{A}_t - A\|_F,$$

where $\mathbb{R}_+ = [0, \infty)$. This projection can be easily achieved by setting all negative entries to zeros; complete details of the PGD can be found in Algorithm 1 in Appendix A.2.

This warm-up phase guides us to a neighborhood around the global optimizer while ensuring the stability/convergence of the algorithm. Moreover, it reduces the computation cost by finding a small batch of coordinates for further optimization in phase 2; see the illustration in Figure 4 and the description of the phase 2 algorithm below.

Phase 2: Batch Coordinate Gradient Descent. In the second phase, we consider those variables/nodes whose corresponding rows (in A) could have negative values. We identify those nodes by the L_2 norm of the gradient (w.r.t. A) row vector — large gradient-norm indicates that the convergence of the corresponding row is not achieved yet after phase 1; see the right panel in Figure 4 for a graphical illustration and complete details on how to identify those rows in Algorithm 2 in Appendix A.2. Next, we need to keep performing GD without the constraint/projection for those selected rows in A to estimate those negative entries (and PGD for the corresponding background intensities). Despite the intractable

Table 2: Performance of proposed method when β is assumed to be known. We observe that all error metrics are decreasing with either an increasing number of sequences or time horizon, which numerically verifies the consistency of our method.

Varying sequence time horizon T (sequence number fixed to be 1).

	$d = 5$					$d = 10$				
T	500	2000	5000	10000	20000	500	2000	5000	10000	20000
μ err.*	7.41 (3.42)	5.34 (2.93)	4.25 (2.76)	3.81 (2.8)	3.69 (2.72)	17.96 (5.98)	11.26 (4.47)	8.65 (4.01)	8.78 (3.8)	7.58 (3.75)
A err.	8.94 (5.96)	2.26 (2.81)	1.01 (0.82)	0.75 (0.43)	0.57 (0.23)	20.95 (12.55)	4.93 (2.55)	2.6 (1.04)	1.87 (0.61)	1.52 (0.48)
A HD	0.24 (0.13)	0.08 (0.091)	0.04 (0.076)	0.04 (0.059)	0.0 (0.056)	0.245 (0.075)	0.07 (0.063)	0.03 (0.035)	0.015 (0.025)	0.01 (0.018)
A SHD	6.0 (3.43)	2.0 (2.28)	1.0 (1.9)	1.0 (1.48)	0.0 (1.41)	24.5 (7.72)	7.0 (6.34)	3.0 (3.57)	1.5 (2.55)	1.0 (1.86)

Varying sequence number (time horizon T fixed to be 500).

	$d = 5$					$d = 10$				
Seq. Num.	1	10	20	50	100	1	10	20	50	100
μ err.*	6.25 (3.29)	3.91 (2.81)	3.86 (2.72)	3.41 (2.66)	2.91 (2.50)	17.96 (5.98)	8.61 (4.01)	8.67 (3.78)	7.54 (3.74)	6.9 (3.49)
A err.	9.42 (5.50)	1.19 (1.22)	0.86 (1.00)	0.6 (0.91)	0.54 (0.91)	20.96 (12.56)	2.62 (1.04)	1.84 (0.61)	1.4 (0.46)	1.51 (0.47)
A HD	0.26 (0.120)	0.06 (0.075)	0.04 (0.061)	0.04 (0.045)	0.0 (0.050)	0.245 (0.075)	0.03 (0.036)	0.015 (0.026)	0.01 (0.018)	0.01 (0.017)
A SHD	7.0 (3.14)	1.5 (1.89)	1.0 (1.53)	1.0 (1.14)	0.0 (1.27)	24.5 (7.72)	3.0 (3.66)	1.5 (2.71)	1.0 (1.85)	1.0 (1.76)

$\star$ the value times 10^{-2} is the actual μ estimation error; we omit $\times 10^{-2}$ in the value due to space consideration.

Table 3: Performance of proposed method when β is unknown. The last row corresponds to selected β based on end-of-phase 1 log likelihood, where we can observe its performance (italic) is almost the same with the best achievable performance (bold).

	$d = 5$				$d = 10$				$d = 20$			
β	μ err.*	A err.	A HD	A SHD	μ err.*	A err.	A HD	A SHD	μ err.*	A err.	A HD	A SHD
0.4	5.01 (3.16)	1.51 (0.85)	0.06 (0.092)	1.5 (2.3)	8.53 (3.81)	4.39 (0.63)	0.03 (0.052)	3.0 (5.27)	14.97 (5.13)	13.46 (2.14)	0.047 (0.044)	19.0 (17.82)
0.5	5.78 (3.3)	1.26 (0.86)	0.04 (0.08)	1.0 (2.01)	10.57 (4.39)	3.49 (0.6)	0.02 (0.033)	2.0 (3.39)	20.51 (6.16)	10.59 (2.27)	**0.043 (0.047)**	**17.5 (18.88)**
0.6	5.39 (3.24)	1.05 (0.88)	0.02 (0.071)	0.5 (1.79)	10.04 (4.36)	2.58 (0.64)	0.02 (0.023)	2.0 (2.38)	21.18 (6.41)	8.55 (2.35)	0.045 (0.046)	18.0 (18.67)
0.7	5.2 (3.12)	0.86 (0.88)	0.0 (0.065)	0.0 (1.64)	8.94 (4.05)	1.86 (0.61)	0.01 (0.024)	1.0 (2.47)	19.35 (6.09)	6.54 (2.43)	0.048 (0.048)	19.5 (19.24)
0.8	4.67 (3.02)	0.74 (0.89)	0.0 (0.039)	0.0 (0.99)	7.54 (3.74)	**1.4 (0.46)**	**0.01 (0.018)**	**1.0 (1.85)**	17.11 (5.65)	**4.98 (2.5)**	0.06 (0.053)	24.0 (21.52)
0.9	4.51 (2.94)	**0.66 (0.9)**	0.0 (0.036)	0.0 (0.91)	**6.79 (3.53)**	1.52 (0.41)	0.01 (0.02)	1.0 (2.08)	**16.34 (5.21)**	5.11 (2.13)	0.07 (0.055)	28.0 (22.32)
1	**4.46 (2.87)**	0.79 (0.92)	0.0 (0.034)	0.0 (0.86)	7.16 (3.47)	1.84 (0.39)	0.01 (0.022)	1.0 (2.28)	18.0 (5.29)	5.93 (1.82)	0.088 (0.055)	35.5 (22.29)
1.1	4.56 (2.79)	1.03 (0.91)	0.0 (0.035)	0.0 (0.87)	7.73 (3.47)	2.3 (0.34)	0.02 (0.026)	2.0 (2.65)	19.95 (5.55)	6.83 (1.64)	0.103 (0.056)	41.5 (22.52)
1.2	4.75 (2.74)	1.2 (1.22)	**0.0 (0.032)**	**0.0 (0.85)**	8.4 (3.46)	2.72 (0.38)	0.03 (0.033)	3.0 (3.36)	22.84 (6.08)	7.8 (1.53)	0.121 (0.052)	48.5 (21.1)
–	*4.57 (2.96)*	*0.74 (0.89)*	*0.0 (0.036)*	*0.0 (0.9)*	*7.04 (3.55)*	*1.62 (0.41)*	*0.01 (0.021)*	*1.0 (2.17)*	*16.7 (5.34)*	*5.06 (2.19)*	*0.07 (0.055)*	*28.0 (22.02)*

$\star$ the value times 10^{-2} is the actual μ estimation error; we omit $\times 10^{-2}$ in the value due to space consideration.

log likelihood in this phase, we develop a learning rate decaying scheme based on the gradient F-norm to guarantee convergence empirically. Complete details of the PGD algorithm can be found in Algorithm 3 in Appendix A.2.

Recently, Juditsky et al. [23], Juditsky and Nemirovski [24] showed that a projected GD along some (strong) monotone vector field can be interpreted as a solution to a stochastic variation inequality (VI) and enjoys both signal estimation guarantee and convergence guarantee. However, since we do not constraint the iterate within Θ in phase 2, the vector fields $\nabla_\mu \tilde{\ell}$ and $\nabla_A \tilde{\ell}$ are no longer monotone. Hence, we could only use numerical evidence to show the effectiveness of our method. Nevertheless, this vector field view under the VI framework might give us a chance to theoretically explain our heuristic's empirical success.

4 EXPERIMENTS

4.1 Numerical simulation

In this subsection, we will show the good performance of our proposed two-phase method. We report (i) L_1 norm of β estimation error (β err.), (ii) L_1 norm of μ estimation error (μ err.), (iii) L_1 norm of A estimation error (A err.), (iv) Hamming Distance (A HD) and (v) Structural Hamming Distance (A SHD) between ground truth and estimated adjacency matrix A as our evaluation metrics. All experiments in this subsection are carried out for randomly generated problem parameters and repeated 100 times; here we report the mean and standard deviation of those metrics. One can see Appendix B.1 for further details.

4.1.1 Experiment 1. We begin with a simple setting where we know the ground truth β. We want to numerically verify the consistency with respect to the time horizon T and the total number of sequences. To be precise, we generate (1) one single sequence on time horizon $T \in \{500, 2000, 5000, 10000, 20000\}$ and (2) multiple sequences (total sequence number chosen from $\{1, 10, 20, 50, 100\}$) on time horizon $T = 500$ and learn the parameter via our proposed two-phase method. We report the results for $d = 5, 10$ cases in Table 2, from which we can see that, with longer sequences (or more sequences), all those errors decrease monotonically. To further validate our findings, we also perform the experiment for $d = 20$ case; the results can be found in Table 9 in Appendix B.2, from which we can still see the decaying error pattern as observed in the above $d = 5, 10$ cases. Therefore, we numerically verify the consistency of our proposed method.

4.1.2 Experiment 2. Next, we consider a more general scenario where we do not know the true β (ground truth is 0.8) — we treat it as a hyperparameter and perform a grid search over $\beta \in \{0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2\}$. We propose to use the *end-of-phase 1 likelihood* as the goodness-of-fit (GoF) criterion to select

hyperparameter β. In comparison, we also consider the end-of-phase 1 gradient-norm as the GoF criterion, but it does not perform as well as the end-of-phase 1 likelihood criterion; one can see the corresponding results in Table 10 in Appendix B.2. For each grid value, we randomly generate synthetic data (50 sequences with $T = 500$) and fit our model. We repeat this procedure independently 100 times, and at each trial, we select β with the largest end-of-phase 1 likelihood. We report the results in Table 3, where we can observe that the grid search approach coupled with end-of-phase 1 likelihood GoF criterion achieves almost the same performance with the best achievable performance (oftentimes it is better than true β's performance). This shows the effectiveness of our approach in practice.

4.1.3 Experiment 3. Lastly, we compare our proposed method with two benchmark methods for $d = 20$ setting. Here, we consider vanilla gradient descent (vanilla GD) and the re-start time method [3, 4] coupled with SGD (gradient-norm as the stopping criterion); further details of the benchmarks can be found in Appendix B.3.1. We report the results in Table 4. As we can see, for both cases, our proposed method outperforms benchmarks in terms of most evaluation metrics; in particular, our method does the best in terms of adjacency matrix recovery. Although GD with proper stopping criterion achieves slightly better L_1 estimation error in case 1, its pattern recovery of A is much worse than our method (i.e., larger HD and SHD), making it unable to return a reliable GC graph in practice.

Table 4: Comparison with benchmarks. The best results are highlighted. We can observe our method performs the best in terms of the adjacency matrix recovery.

Case 1: single sequence with time horizon $T = 10000$.

Method	Two-phase method	Vanilla GD	Early stopped GD	Re-start
β err.	.312 (.112)	.393 (.035)	**.264** (.137)	.837 (.246)
μ err.	**.0386** (.0252)	.0413 (.0317)	.0398 (.0281)	.239 (.102)
A err.	1.726 (0.785)	23.58 (7.93)	**1.494** (0.731)	8.828 (1.213)
A HD	**.0304** (.0416)	.1336 (0.118)	.0936 (.0926)	.3576 (.0459)
A SHD	**.76** (1.04)	3.37 (2.96)	2.34 (2.32)	8.98 (1.19)

Case 2: multiple (100) sequences with time horizon $T = 500$.

Method	Two-phase method	Vanilla GD	Early stopped GD	Re-start + SGD
β err.	.264 (.151)	.367 (.074)	**.254** (.136)	.295 (.128)
μ err.	.0489 (.0269)	.0367 (.0228)	.0435 (.0272)	**.0198** (.0136)
A err.	**0.983** (0.301)	12.13 (2.419)	1.759 (0.480)	1.067 (0.401)
A HD	**.0236** (0.0376)	.0748 (0.0680)	.0808 (0.0642)	.044 (0.0639)
A SHD	**.59** (.94)	1.88 (1.72)	2.02 (1.61)	1.1 (1.60)

Additionally, as shown in Table 1, our proposed approach is scalable, which is another major advantage compared with the re-start time approach. Here, we demonstrate this benefit by performing a run time analysis. Due to space consideration, the results are deferred to Table 11 in Appendix B.3.2.

4.2 Real data example

We created a retrospective cohort of patients utilizing in-hospital data derived from Grady hospital system in Atalanta, GA, an academic level 1 trauma center, spanning 2018-2019. This data was collected and analyzed in accordance with Emory Institutional Review Board (IRB) approved protocol #STUDY00000302. Patients were included in the Sepsis-3 cohort if they met Sepsis-3 criteria while in the hospital and were admitted for $\geq$ 24 hours. Patients were included in the Non-Septic cohort if they had a Sequential Organ Failure Assessment (SOFA) score $\geq$2. A total of 37 patient features comprised of laboratory results (Labs) and observations (vital signs) were examined for this work. Treatments were limited to two classes of medication: antimicrobial therapy (e.g., antibiotics) and vasopressor therapy. We report the median and interquartile range (IQR) in Table 5 and defer the cohort construction details to Appendix C.1.

Table 5: Summary statistics of patients' demographics.

	Sepsis-3 patients		Non-sepatic patients	
year	2018 (n = 409) ⋆	2019 (n = 454)	2018 (n = 960)	2019 (n = 1169)
Age (median and IQR)	58 (38 - 68)	59 (46 - 68)	56 (38 - 67)	55 (37 - 66)
Female (percentage)	30.1 %	36.6 %	37.1 %	35.8 %
SOFA score (mean)	3.32	3.14	2.18	2.28
Traj. len. (median and IQR)	25 (25 - 25)	25 (25 - 25)	17 (13 - 22)	17 (13 - 22)

⋆ n represents the total number of patients in the corresponding cohort.

4.2.1 Sepsis-Associated Derangements. Integrating high dimensional information (via, e.g., clustering) is essential in causal discovery and explainable machine learning [36]; examples include Braman et al. [5], Uleman et al. [43], Wei et al. [45]. While the Sepsis-3 definition provides the explicit features necessary for identifying the presence of sepsis, there is no consensus as to which features are best for prognosticating the disease. To reduce the complexity of our computations, expert opinion was utilized to identify common and clinically relevant Sepsis-Associated Derangements (SADs) that could be detected using structured EMR data. A total of 18 SADs and 2 relevant treatments shown in Table 6 were identified using 37 patient features and treatments gathered from the medical record. A SAD was considered present if the patient's features were outside of normal limits. Details on how SADs were constructed based on vital signs and Labs can be found in Table 12 in Appendix C.1.

Table 6: Measurements to construct sepsis-associated events.

Sepsis-Associated Derangement

Full name	Abbreviation	Measurement name
Renal Dysfunction	RenDys	creatinine, blood_urea_nitrogen_(bun)
Electrolyte Imbalance	LyteImbal	calcium, chloride, magnesium, potassium, phosphorus
Oxygen Transport Deficiency	O2TxpDef	hemoglobin
Coagulopathy	Coag	partial_prothrombin_time_(ptt), fibrinogen, platelets, d_dimer, thrombin_time, prothrombin_time_(pt), inr
Malnutrition	MalNut	transferrin, prealbumin, albumin
Cholestatsis	Chole	bilirubin_direct, bilirubin_total
Hepatocellular Injury	HepatoDys	aspartate_aminotransferase_(ast), ammonia, alanine_aminotransferase_(alt)
Acidosis	Acidosis	base_excess, ph
Leukocyte Dysfunction	LeukDys	white_blood_cell_count
Hypercarbia	HypCarb	partial_pressure_of_carbon_dioxide_(paco2), end_tidal_co2
Hyperglycemia	HypGly	glucose
Mycardial Ischemia	MyoIsch	troponin
Tissue Ischemia	TissueIsch	base_excess, lactic_acid
Diminished Cardiac Output	DCO	best_map
CNS Dysfunction	CNSDys	gcs_total_score
Oxygen Diffusion Dysfunction	O2DiffDys	spo2, fio2
Thermoregulation Dysfunction	ThermoDys	temperature
Tachycardia	Tachy	pulse

Other Sepsis-Associated Events

Full name	Abbreviation	Measurement name
Vasopressor Support	VasoSprt	norepinephrine_dose_weight, epinephrine_dose_weight, dobutamine_dose_weight, dopamine_dose_weight, phenylephrine_dose_weight, vasopressin_dose_weight
Antibiotic Therapy	ABX	–
Sepsis	SEP3	–

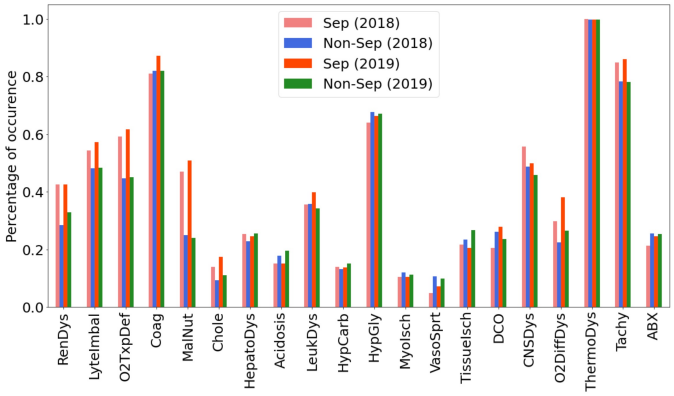

Figure 5: Percentage of SAD's occurrence. We can observe that most SADs' occurrences are more frequent in Sepsis-3 cohort, which can justify our approach to construct SADs.

Sepsis often shares symptoms with other disease processes making discrimination challenging. To evaluate the appropriateness of the constructed SADs, the percentage of SAD occurrence (within the selected time window) was calculated for both Sepsis and Non-Septic patients and can be seen for both years in Figure 5. It is expected that SADs would be present in both cohorts; however, the Sepsis-3 cohort demonstrated patterns showing a closer relationship with the SADs than the Non-Septic cohort.

4.2.2 Recovering GC graphs. To study the temporal interactions between SADs (and other SAEs), we fit two GC graphs — one graph is on the Sepsis-3 cohort and the other is on the full patient cohort (i.e., the Sepsis-3 and Non-Septic cohorts combined). We report the results in Figure 6 and defer the training details to Appendix C.2. Both graphs demonstrate examples of clinically reasonable interactions between individual SADs (i.e., Oxygen Diffusion Dysfunction promotes Renal Dysfunction in the Septic cohort) and between SADs and Sepsis (i.e., Diminished Cardiac Output promotes Sepsis in both graphs). Interestingly the graph examining only the Sepsis-3 cohort identified more interactions between SADs than the one for the full patient cohort whereas the graph for the full patient cohort presented a higher number of strong relationships between SADs and sepsis suggesting that a time-dependent, causal relationship exists between individual SADs and sepsis. A key finding across both graphs was the inhibitory effect of antibiotics on most SADs, which is consistent with the known ability of antibiotics to reduce in-hospital mortality in sepsis patients [38] presumably through preventing organ dysfunction like those identified via SADs.

While most of the relationships identified in these graphs are expected or feasible, vasopressors appear to unexpectedly inhibit both sepsis and the administration of antibiotics. In the year 2018, among 409 (960) selected septic (non-septic) patients, there were 15 (96) receive vasopressor support and 84 (231) received antibiotics during the window, and only 3 (38) received both vasopressors and antibiotics. This low number of vasopressor patients in the Sepsis-3 cohort is not unexpected as the time window for analysis is 24 hours prior to meeting the Sepsis-3 definition when most patients are not severely ill (see Appendix C.1 for more details). Additionally, antibiotics are dosed at scheduled intervals (e.g., once every six hours) whereas vasopressors are administered in a continuous fashion. These two attributes of the data set create a number of instances where vasopressors are administered without a formal antibiotic administration event in the following hour (though the patient may be on antibiotics). Additionally, each patient in the Sepsis-3 cohort is right censored after sepsis which means there is only one hour where the sepsis label is positive. Taken together these attributes of the data set likely explain why this unexpected relationship is seen.

4.2.3 Identifying GC chains. The estimated GC graphs help reduce the problem of enumerating combinatorially many possible chains to find the chains that only exist in the Sepsis-3 graph. However, even for a 2-by-2 sub-adjacency matrix, there could be multiple potential chain interpretations. We validate whether or not the chain structure reflects a unique pattern in the Sepsis-3 cohort by performing Fisher's exact test and reporting the p-value. Here, we only focus on the "++" and "+++" exciting effects when forming all possible chains. This method allows chains to be ranked in order of significance, affording those with domain expertise an efficient mechanism to inspect results. We report the top GC chains which are unique in the Sepsis-3 cohort for years 2018 (in-sample test) and 2019 (out-of-sample test) in Table 7. More details on those chains (including how to perform the test) and more identified chains can be found in Tables 14, 15 and 16 in Appendix C.3.

Table 7: Granger Causal chains which are significantly unique in Sepsis-3 cohort in both years 2018 and 2019.

Chain:	TissueIsch →	O2DiffDys	
p-value:	0.004 (2018)	0.092 (2019)	
Chain:	O2DiffDys →	RenDys →	O2DiffDys
p-value:	0.107 (2018)	0.004 (2019)	
Chain:	VasoSprt →	TissueIsch →	HepatoDys
p-value:	0.052 (2018)	0.088 (2019)	
Chain:	LyteImbal →	Acidosis →	O2DiffDys
p-value:	0.009 (2018)	0.088 (2019)	
Chain:	Acidosis →	O2DiffDys →	HypGly
p-value:	0.039 (2018)	0.063 (2019)	

In Table 7, the chains possess a statistically strong relationship with patients in the Sepsis-3 cohort and correlate with clinical patterns that are often seen in sepsis. For example, Oxygen Diffusion Dysfunction (i.e., low oxygen saturation in the blood) is found to promote Renal Dysfunction and subsequent Oxygen Diffusion Dysfunction. Though not reflected in this table, septic patients could experience multiple chains simultaneously in addition to experiencing other discrete SADs simultaneous to events in a chain. This method to select and rank chains affords clinicians the ability to efficiently discover or follow those temporal patterns that differentiate septic patients from those experiencing organ injury caused by other diseases.

4.2.4 Evaluating the quantitative performance. Due to the lack of time granularity of the time series data and the overly simple parametric form of the MHP model, we do not build a sequential prediction model to validate our method's usefulness. Instead, we quantitatively validate the usefulness of the identified GC chains by using them to construct features and apply a more sophisticated (but less interpretable) machine learning method to perform sequential prediction tasks. Here, we choose XGBoost due to its

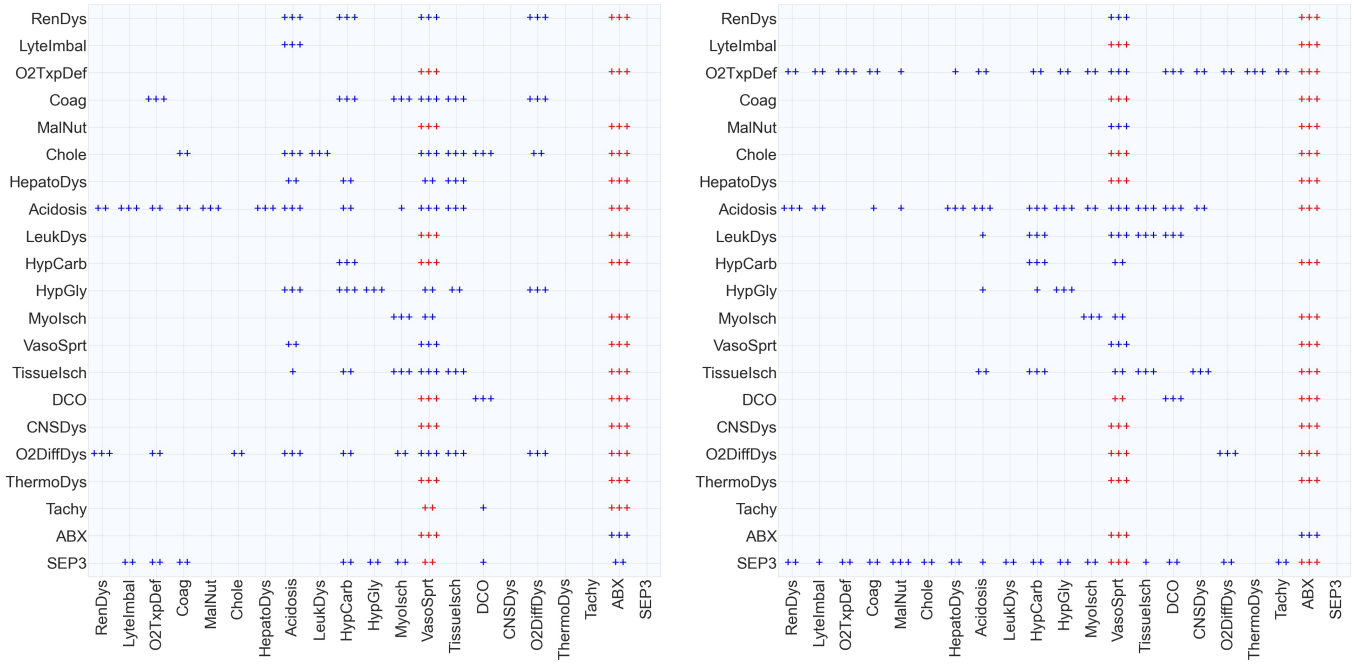

Figure 6: Adjacency matrices of the Granger Causal graphs for Sepsis-3 (left) and full (right) patient cohorts in Figure 1. "+", "++" and "+++" correspond to the (absolute) value in $(0, .0005)$, $[.0005, .001)$ and $[.001, \infty)$, respectively, where the original values in the adjacency matrices are reported in Figure 11 in the Appendix. Nodes (i.e., SADs or SAEs) named along the X-axis can have either an inhibitory (red) or promoting effect (blue) on the nodes named on the Y-axis.

Table 8: Sepsis event prediction: the proposed method that incorporates the identified GC chains as input features achieves better performance than the benchmark method.

	In-sample (year 2018)		Out-of-sample (year 2019)	
	Benchmark	Proposed	Benchmark	Proposed
Accuracy	0.7183	0.7862	0.7214	0.7789
Sensitivity	0.7258	0.7983	0.7300	0.7930

good performance (compared with other choices such as neural networks) in the sepsis prediction challenge [35].

In the benchmark XGBoost method, we use the mean values of the past 12 hours' SADs as input features. In contrast, we additionally include binary variables indicating whether or not there exist chain patterns as shown in Table 7 in the past 12 hours as the input features, to see whether or not this can improve the prediction accuracy and sensitivity. We use 5-fold cross validation (for grid search of hyperparameters in XGBoost) and train the model using 2018 data. We test the performance on 2019 data. The results are reported in Table 8, where we can observe improvements in both accuracy and sensitivity when predicting sepsis using our identified GC chains as input features. This suggests the usefulness of the identified GC chains; however, building a powerful prediction model with such GC chains is still on-going work.

5 CONCLUSION

To conclude this paper, we briefly summarize the contribution and limitations of current work. We defer an extended discussion to Appendix D. Our proposed method for Granger Causal chain discovery provides a novel and scalable approach to leverage clinical expertise to elucidate patterns of interest amongst large amounts of related EMR data. Though we do not build or validate a clinical alarm, this is a very useful and logical extension of this work. Additionally, knowledge from the GC chains could be used to estimate the risk of a future SAD (e.g., Renal Dysfunction) which might prompt a clinician to alter treatment (e.g., modify IV fluids therapy). A limitation of this work stems from the grouped nature of many lab results and vital sign measurements. It is not uncommon for multiple patient features to be recorded in the EMR with identical timestamps which means that multiple SADs can occur simultaneously. This presents challenges to our point process model which can not capture relationships between simultaneously occurring SADs. This could be remedied by incorporating second or third-order interaction effect in ANOVA into the work to evaluate the effect of combined SADS on future patient states. Another limitation of the method arises from the way treatments are administered. Some treatments (i.e., antibiotics) are dosed on an interval whereas others (i.e., vasopressors) are dosed continuously. This results in a higher number of "vasopressor" events than antibiotic events for certain patients and can lead to the false conclusion that vasopressors are inhibiting antibiotics which is not an expected finding. Possible solutions include representing antibiotics as a continuous medication similar to vasopressors so that the continuous effects of antibiotics are appreciated by the model.

ACKNOWLEDGMENT

The work of S. Wei and Y. Xie is partially funded by an NSF CAREER CCF-1650913, and NSF DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, DMS-1830210, and an Emory Hospital grant.

REFERENCES

[1] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. 2015. Hawkes processes in finance. *Market Microstructure and Liquidity* 1, 01 (2015), 1550005.

[2] Yujia Bao, Zhaobin Kuang, Peggy Peissig, David Page, and Rebecca Willett. 2017. Hawkes process modeling of adverse drug reactions with longitudinal observational data. In *Machine learning for healthcare conference*. PMLR, 177–190.

[3] Anna Bonnet, Miguel Martinez Herrera, and Maxime Sangnier. 2021. Maximum Likelihood Estimation for Hawkes Processes with self-excitation or inhibition. *Statistics & Probability Letters* 179 (2021), 109214.

[4] Anna Bonnet, Miguel Martinez Herrera, and Maxime Sangnier. 2022. Inference of multivariate exponential Hawkesprocesses with inhibition and application toneuronal activity. *arXiv preprint arXiv:2205.04107* (2022).

[5] Nathaniel Braman, Jacob WH Gordon, Emery T Goossens, Caleb Willis, Martin C Stumpe, and Jagadish Venkataraman. 2021. Deep orthogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 667–677.

[6] Álvaro Cartea, Samuel N Cohen, and Saad Labyad. 2021. Gradient-based estimation of linear Hawkes processes with general kernels. *arXiv preprint arXiv:2111.10637* (2021).

[7] Wei Chen, Jibin Chen, Ruichu Cai, Yuequn Liu, and Zhifeng Hao. 2022. Learning granger causality for non-stationary Hawkes processes. *Neurocomputing* 468 (2022), 22–32.

[8] Edward Choi, Nan Du, Robert Chen, Le Song, and Jimeng Sun. 2015. Constructing disease network and temporal progression model via context-sensitive hawkes process. In *2015 IEEE International Conference on Data Mining*. IEEE, 721–726.

[9] Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. 2017. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis* 38, 2 (2017), 225–242.

[10] Michael Eichler and Vanessa Didelez. 2010. On Granger causality and the effect of interventions in time series. *Lifetime data analysis* 16, 1 (2010), 3–32.

[11] Doris Entner and Patrik O Hoyer. 2010. On causal discovery from time series data using FCI. *Probabilistic graphical models* (2010), 121–128.

[12] Zhuangyan Fang, Shengyu Zhu, Jiji Zhang, Yue Liu, Zhitang Chen, and Yangbo He. 2020. Low Rank Directed Acyclic Graphs and Causal Structure Learning. *arXiv preprint arXiv:2006.05691* (2020).

[13] Lucas M Fleuren, Thomas LT Klausch, Charlotte L Zwager, Linda J Schoonmade, Tingjie Guo, Luca F Roggeveen, Eleonora L Swart, Armand RJ Girbes, Patrick Thoral, Ari Ercole, et al. 2020. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive care medicine* 46, 3 (2020), 383–400.

[14] Eric Warren Fox, Frederic Paik Schoenberg, and Joshua Seth Gordon. 2016. Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *The Annals of Applied Statistics* 10, 3 (2016), 1725–1756.

[15] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics* 10 (2019), 524.

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

[17] Clive WJ Granger. 1980. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control* 2 (1980), 329–352.

[18] Clive WJ Granger. 1988. Some recent development in a concept of causality. *Journal of econometrics* 39, 1-2 (1988), 199–211.

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

[20] Alan G Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58, 1 (1971), 83–90.

[21] Alan G Hawkes and David Oakes. 1974. A cluster process representation of a self-exciting process. *Journal of Applied Probability* 11, 3 (1974), 493–503.

[22] Tsuyoshi Ide, Georgios Kollias, Dzung Phan, and Naoki Abe. 2021. Cardinality-Regularized Hawkes-Granger Model. *Advances in Neural Information Processing Systems* 34 (2021).

[23] Anatoli Juditsky, Arkadi Nemirovski, Liyan Xie, and Yao Xie. 2020. Convex Parameter Recovery for Interacting Marked Processes. *IEEE Journal on Selected Areas in Information Theory* (2020).

[24] Anatoli B Juditsky and AS Nemirovski. 2019. Signal Recovery by Stochastic Optimization. *Automation and Remote Control* 80, 10 (2019), 1878–1893.

[25] Diviyan Kalainathan and Olivier Goudet. 2019. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278* (2019).

[26] Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N Brown. 2011. A Granger causality measure for point process models of ensemble neural spiking activity. *PLoS computational biology* 7, 3 (2011), e1001110.

[27] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. 2016. A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *IEEE/ACM transactions on computational biology and bioinformatics* 16, 5 (2016), 1483–1495.

[28] PA W Lewis and Gerald S Shedler. 1979. Simulation of nonhomogeneous Poisson processes by thinning. *Naval research logistics quarterly* 26, 3 (1979), 403–413.

[29] David Marsan and Olivier Lengline. 2008. Extending earthquakes' reach through cascading. *Science* 319, 5866 (2008), 1076–1079.

[30] Hongyuan Mei and Jason M Eisner. 2017. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*. 6754–6764.

[31] Sebastian Meyer and Leonhard Held. 2014. Power-law models for infectious disease spread. *The Annals of Applied Statistics* 8, 3 (2014), 1612–1639.

[32] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. 2020. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems* 33 (2020), 17943–17954.

[33] Kimia Noorbakhsh and Manuel Gomez Rodriguez. 2021. Counterfactual Temporal Point Processes. *arXiv preprint arXiv:2111.07603* (2021).

[34] Matthew A Reyna, Chris Josef, Salman Seyedi, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D Clifford. 2019. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. In *2019 Computing in Cardiology (CinC)*. IEEE, Page–1.

[35] M. A. Reyna, C. S. Josef, R. Jeter, S. P. Shashikumar, M. B. Westover, S. Nemati, G. D. Clifford, and A. Sharma. 2020. Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019. *Crit Care Med* 48, 2 (02 2020), 210–217.

[36] Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. 2022. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science* 9, 8 (2022), 220638.

[37] Frederic Paik Schoenberg, Marc Hoffmann, and Ryan J Harrigan. 2019. A recursive point process model for infectious diseases. *Annals of the Institute of Statistical Mathematics* 71, 5 (2019), 1271–1287.

[38] Christopher W. Seymour, Foster Gesten, Hallie C. Prescott, Marcus E. Friedrich, Theodore J. Iwashyna, Gary S. Phillips, Stanley Lemeshow, Tiffany Osborn, Kathleen M. Terry, and Mitchell M. Levy. 2017. Time to Treatment and Mortality during Mandated Emergency Care for Sepsis. *N. Engl. J. Med.* 376, 23 (2017), 2235–2244. https://doi.org/10.1056/NEJMoa1703058

[39] Supreeth P. Shashikumar, Christopher S. Josef, Ashish Sharma, and Shamim Nemati. 2021. DeepAISE – An interpretable and recurrent neural survival model for early prediction of sepsis. *Artificial Intelligence in Medicine* 113 (March 2021), 102036. https://doi.org/10.1016/j.artmed.2021.102036

[40] Ali Shojaie and Emily B Fox. 2022. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application* 9 (2022), 289–319.

[41] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. 2016. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jama* 315, 8 (2016), 801–810.

[42] Pater Spirtes, Clark Glymour, Richard Scheines, Stuart Kauffman, Valerio Aimale, and Frank Wimberly. 2000. Constructing Bayesian network models of gene expression networks from microarray data. (2000).

[43] Jeroen F Uleman, René JF Melis, Rick Quax, Eddy A van der Zee, Dick Thijssen, Martin Dresler, Ondine van de Rest, Isabelle F van der Velpen, Hieab HH Adams, Ben Schmand, et al. 2021. Mapping the multicausality of Alzheimer's disease through group model building. *GeroScience* 43, 2 (2021), 829–843.

[44] Haoyun Wang, Liyan Xie, Alex Cuozzo, Simon Mak, and Yao Xie. 2020. Uncertainty Quantification for Inferring Hawkes Networks. *arXiv preprint arXiv:2006.07506* (2020).

[45] Song Wei, Yao Xie, Christopher S Josef, and Rishikesan Kamaleswaran. 2021. Causal Graph Recovery for Sepsis-Associated Derangements via Interpretable Hawkes Networks. In *International Conference on Machine Learning (Workshop on Interpretable Machine Learning in Healthcare (IMLH))*.

[46] WHO. 2020. Global report on the epidemiology and burden of sepsis: current evidence, identifying gaps and future directions. (2020).

[47] Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. 2016. Learning granger causality for hawkes processes. In *International Conference on Machine Learning*. PMLR, 1717–1726.

[48] Wei Zhang, Thomas Panum, Somesh Jha, Prasad Chalasani, and David Page. 2020. Cause: Learning granger causality from event sequences using attribution methods. In *International Conference on Machine Learning*. PMLR, 11235–11245.

[49] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems* 31 (2018).

[50] Ke Zhou, Hongyuan Zha, and Le Song. 2013. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*. PMLR, 641–649.

[51] Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. 2002. Stochastic declustering of space-time earthquake occurrences. *J. Amer. Statist. Assoc.* 97, 458 (2002), 369–380.

[52] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. 2020. Transformer hawkes process. In *International conference on machine learning*. PMLR, 11692–11702.

APPENDIX

Appendix, including step-by-step details of the proposed method, experimental settings and additional results on both simulated and real data, as well as an extended discussion, can be found in the full paper available at https://arxiv.org/abs/2209.04480.