Adverse Drug Reaction Discovery from Electronic Health Records with Deep Neural Networks

Wei Zhang
zhangwei@cs.wisc.edu
Computer Sciences Department
University of Wisconsin–Madison

Zhaobin Kuang
kuangz@stanford.edu
Computer Science Department
Stanford University

Peggy Peissig
Peissig.Peggy@marshfieldresearch.org
Biomedical Informatics Research Center
Marshfield Clinic Research Institute

David Page
david.page@duke.edu
Department of Biostatistics and Bioinformatics
Duke University

ABSTRACT

Adverse drug reactions (ADRs) are detrimental and unexpected clinical incidents caused by drug intake. The increasing availability of massive quantities of longitudinal event data such as electronic health records (EHRs) has redefined ADR discovery as a big data analytics problem, where data-hungry deep neural networks are especially suitable because of the abundance of the data. To this end, we introduce neural self-controlled case series (NSCCS), a deep learning framework for ADR discovery from EHRs. NSCCS rigorously follows a self-controlled case series design to adjust implicitly and efficiently for individual heterogeneity. In this way, NSCCS is robust to time-invariant confounding issues and thus more capable of identifying associations that reflect the underlying mechanism between various types of drugs and adverse conditions. We apply NSCCS to a large-scale, real-world EHR dataset and empirically demonstrate its superior performance with comprehensive experiments on a benchmark ADR discovery task.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Computing methodologies** → **Neural networks**.

KEYWORDS

Adverse Drug Reaction Discovery, Electronic Health Records, Deep Neural Networks, Self-Controlled Case Series

ACM Reference Format:
Wei Zhang, Zhaobin Kuang, Peggy Peissig, and David Page. 2020. Adverse Drug Reaction Discovery from Electronic Health Records with Deep Neural Networks. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '20), April 2–4, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3368555.3384459

ACM CHIL '20, April 2–4, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7046-2/20/04…$15.00
https://doi.org/10.1145/3368555.3384459

1 INTRODUCTION

Adverse drug reactions (ADRs) are detrimental and unexpected clinical incidents caused by taking medication. Because the precise pharmacological mechanisms for many ADRs are still largely unknown, post-marketing surveillance of drug safety is indispensable. The use of longitudinal event data (LED), such as electronic health records (EHRs), is widely regarded as one of the most promising paths to achieve proactive pharmacovigilance [12].

In EHRs, health information from millions of patients is collected over time including drug prescription records and condition diagnosis records. In principle, by applying computational methods to analyze the co-occurrences between various drug prescriptions and condition diagnoses in EHRs, we can infer drugs that potentially cause conditions as ADRs. Of course, major challenges remain, including the following: (a) EHR data are purely observational, and hence a good predictor of an event is not necessarily causal but may be the result of confounding; (b) confounding variables are often unmeasured, or "hidden"; and (c) ADRs in many cases can be caused by drug interactions or other complex nonlinear effects [22, 36].

A major change in recent years is that researchers and practitioners now have access to many massive EHR databases, which has redefined ADR discovery as a big data analytics problem, as demonstrated in modern drug safety surveillance systems [11, 28, 32]. Meanwhile, the recent success of deep neural networks (DNNs) in various domains has demonstrated their power in capturing intricate patterns and providing insights in the big data setting. We believe that ADR discovery from EHRs can also possibly benefit from DNN-driven algorithms. However, DNN-driven algorithms for identifying ADRs from LED, to the best of our knowledge, are still non-existent. Two obstacles must be overcome in order to facilitate the introduction of DNNs to the task of ADR discovery.

First, ADR discovery seeks to identify *associations that reflect the underlying causation* between various types of drugs and conditions, in the hope of detecting potential ADRs. The recently proposed DNNs, while making progress in modeling LED [7, 23, 38] or sequential data in general [3, 35], focus on prediction tasks, in particular, regarding when and which event type the next event will be. A DNN-based algorithm focusing on estimating causally credible associations among various event types, to our knowledge, has not been invented.

The more fundamental obstacle is *individual heterogeneity*, due to the fact that LED like EHRs are collected from subjects with potentially diverse health profiles. For example, when finding drugs that could potentially cause myocardial infarction (MI, or "heart attack") as an ADR, distinct patients could have drastically different risks of MI, due to their intrinsic differences in genetics and epigenetics, socioeconomic status, and diet or other unobserved environmental exposures. Such differences can confound the actual signals generated by drugs, yet might be unobserved and even completely unconsidered by investigators. Furthermore, since the occurrences of ADRs are usually rare, signals resulting from drug intakes are usually weak. This fact further elevates the importance of ascertaining the contributions accurately from individual heterogeneity vs. actual ADR signals. Recently proposed DNNs fail to take into consideration the heterogeneity among different individuals, and effectively assume that data are generated from homogeneous subjects. Such an assumption is obviously too restrictive and may introduce spurious relationships due to the modeling flexibility of DNNs.

In this paper, we offer solutions to the aforementioned obstacles and propose neural self-controlled case series (NSCCS), a DNN-based model that rigorously follows a self-controlled case series (SCCS) design. NSCCS can adjust implicitly and efficiently for individual heterogeneity, meanwhile enjoys the representation power of DNNs to capture probable drug interactions or other complex nonlinear effects. In this way, our work is a pilot effort to endow DNNs with the ability to automatically control for time-invariant confounders that are prevalent in EHRs, regardless of whether these confounders are observed or unobserved [9]. We apply NSCCS to a large-scalereal-world EHR and demonstrate its superior performance with comprehensive experiments on a benchmark ADR discovery task.

2 BACKGROUND

In this section, we illustrate how EHRs can be viewed as longitudinal event data and introduce some necessary notation. We then provide the likelihood of the occurrences of the outcome (e.g., a potential ADR) for a particular individual using the conditional intensity function.

2.1 Electronic Health Records as Longitudinal Event Data

Figure 1a illustrates an EHR from two patients. As shown, EHRs are a example of time-stamped, multi-type event data collected from a large group of heterogeneous individuals, where the event type represents the drug prescription or the condition diagnosis associated to the event. Formally, suppose there are P individuals (patients) and S event types in total. We use $[n]$ as shorthand for the set $\{1, 2, \ldots, n\}$ for any positive integer n. Without loss of generality, suppose that we are interested in modeling the occurrences of the first O event types in $[S]$, where $O \leq S$; we call these first O event types of interest *outcomes*. In ADR discovery from EHRs, outcomes are the adverse conditions, and we are interested in predicting adverse condition occurrences based on their previous occurrences as well as drug prescription history for each patient.

For any individual $p \in [P]$, we denote the whole event sequence collected from that individual as $\mathcal{H}_p \triangleq \{(t_{pj}, s_{pj}) \mid j \in [n_p]\}$, where t_{pj} and s_{pj} are the time-stamp and the type of the j-th event, respectively, n_p is the total number of events, and "$\triangleq$" represents "defined as." We also assume that $\mathcal{H}_p$ are independently generated for distinct $p \in [P]$. We also assume that that these events are restricted to a fixed observation time interval $[l_p, u_p]$ and are ordered by their timestamps, i.e., $l_p < t_{p1} < t_{p2} < \cdots < t_{pn_p} < u_p$. We further define the history of the p-th individual prior to time t as $\mathcal{H}_p(t) \triangleq \{(t_{pj}, s_{pj}) \mid t_{pj} < t\}$, and use $\mathcal{O}_p \triangleq \{(t_{pj}, s_{pj}) \mid (t_{pj}, s_{pj}) \in \mathcal{H}_p, s_{pj} \in [O]\}$ to represent the set of all occurrences of the outcomes for the p-th individual. With the notation introduced above, we will present the likelihood of observing $\mathcal{O}_p$ in the data in the next section.

2.2 Outcome Likelihood

To describe the likelihood of outcome event sequence $\mathcal{O}_p$ given the whole event sequence $\mathcal{H}_p$ of individual p, point process theory [4] introduces the concept of *conditional intensity function* $\lambda_{po}\left(t; \mathcal{H}_p(t)\right)$ to characterize the occurrence of outcomes. Formally, let dt be an infinitesimal time interval; the probability of the occurrence of the outcome o during the time span $[t, t + dt)$ is equal to $\lambda_{po}(t; \mathcal{H}_p(t))dt$. To simplify the notation, we use $\lambda^*_{po}(t)$ as shorthand for $\lambda_{po}(t; \mathcal{H}_p(t))$, and the use of asterisk in $\lambda^*_{po}(t)$, by convention, is to emphasize that the intensity in question is conditioned upon the history prior to time t. The log-likelihood of observing $\mathcal{O}_p$ for the p-th individual is:

$$\log \mathrm{P}(\mathcal{O}_p) = \sum_{o=1}^{O} \sum_{j=1}^{n_p} \log \lambda^*_{po}\left(t_{pj}\right) \cdot \mathbb{I}(s_{pj} = o) - \int_{l_p}^{u_p} \lambda^*_{po}\left(t\right) dt, \quad (1)$$

where $\mathbb{I}(\cdot)$ is the indicator function. The derivation of (1) is provided in the Appendix. Note that the actual computation of (1) is highly dependent on the choice of $\lambda^*_{po}(t)$, and that $\lambda^*_{po}(t)$ is indexed by $p \in [P]$, which suggests that the choice of $\lambda^*_{po}(t)$ is specific to each individual. Recall that by definition $\lambda^*_{po}(t)$ is dependent on $\mathcal{H}_p(t)$. Therefore, modeling both individual heterogeneity and the influences of the occurrences of various types of events on the occurrences of the outcome can be achieved by choosing appropriate $\lambda^*_{po}(t)$. In the next section, we will present our the choice of $\lambda^*_{po}(t)$ adhere to an SCCS design and the solution to the corresponding optimization problem of (1).

3 NEURAL SELF-CONTROLLED CASE SERIES

We present the neural self-controlled case series model in this section. We start by introducing conditional intensity functions that incorporate individual heterogeneity and model the relationships among various types of events via a DNN from event history. We derive NSCCS from an MLE perspective. The advantages of NSCCS are then thoroughly explained. Finally, we discuss the optimization procedure used to solve NSCCS with a special treatment based on the use of piecewise constant conditional intensity functions.

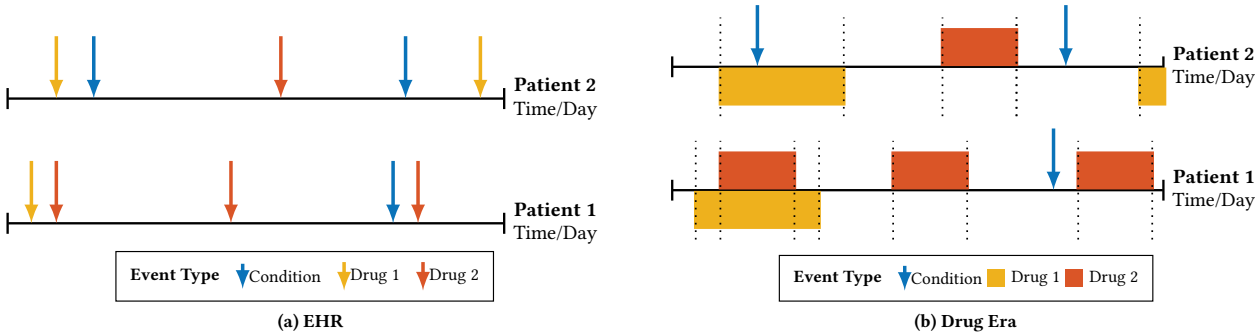


Figure 1: (a) Visualization of the EHRs from two patients. Arrows in different colors represent different event types such as prescriptions of different drugs and diagnoses of different conditions. (b) The drug eras corresponding to the drug prescription events. Dotted lines represent separation of time intervals during which binary drug exposure statuses of various drugs stay unchanged.

3.1 Individual-Heterogeneous Conditional Intensity Functions

We are interested in designing a $\lambda^*_{po}(t)$ such that the conditional intensity function can not only account for individual heterogeneity but also have flexibility in modeling potentially intricate relationships between occurrences of various event types and the occurrences of an outcome. To this end, we consider a conditional intensity function of the form:

$$\lambda^*_{po}(t) = \exp\left[\alpha_{po} + f^*_o(t,p)\right], \tag{2}$$

where α_{po} represents the log baseline conditional intensity of the individual p for the outcome type o, and $f^*_o(t,p) \triangleq f_o\left(t; \mathcal{H}_p(t)\right)$ represents the potentially highly intricate log conditional intensity contributed by the event history $\mathcal{H}_p(t)$ of events of various types prior to time t. Unlike the patient specific baseline, the function $f_o(\cdot)$ is shared across different individuals; thus it aims to model the global effects of the event history to the outcome o.

This parameterization of conditional intensity captures two aspects of influences due to different sources via α_{po} and $f^*_o(t,p)$. First, intrinsic individual heterogeneity for various outcomes can be accounted for by α_{po}'s. Such an explicit modeling of baseline intensity is especially suitable for real-world event data that are usually collected from a massive number of diverse individuals. In ADR discovery, different patients might have different risks of having a heart attack at the same age because of their different health conditions, genetics, and so forth. Yet these factors that lead to individual heterogeneity are not even necessarily observed. Nonetheless, the α_{po}'s introduced are agnostic to whether the factors are observed or unobserved; eventually a comprehensive baseline effect will be estimated based on all the observed and unobserved factors that are time-invariant throughout the subject's trajectory.

Second, the influence of personalized history of events is modeled via the use of globally shared $f^*_o(t,p)$. In the history $\mathcal{H}_p(t)$, various types of events could interact with each other, and exert potentially nonlinear synergistic effects towards the occurrences of the outcomes. For example, many of the most unexpected and detrimental ADRs are due to drug-drug interactions [22, 36]. Furthermore, the exact occurrence time of the events in the history could also make a difference in the eventual influence on the occurrences of the outcomes. For example, if the ADR in question is an acute effect of a particular drug, then the corresponding drug prescription events that occur right before the speculated adverse reaction event should deserve more attention. In contrast, if an ADR is due to long-term and high-dosage use of a medication, then all the corresponding drug prescription events in the past should be taken into consideration. All these complications demand a functional form with substantial flexibility and representation power. DNNs are the choice in our work to model $f^*_o(t,p;\boldsymbol{\Theta})$, where we rewrite $f^*_o(t,p)$ with the dependency on $\boldsymbol{\Theta}$, the parameterization of a DNN.

3.2 Deriving NSCCS from an MLE Perspective

To estimate α_{po}'s and $\boldsymbol{\Theta}$, maximum likelihood estimation (MLE) can be used. Specifically, define

$$A_{po}(\boldsymbol{\Theta}) \triangleq \log \int_{l_p}^{u_p} \exp f^*_o\left(t,p;\boldsymbol{\Theta}\right) dt, \tag{3}$$

and define $\boldsymbol{\alpha}$ as the vector of all α_{po}, for all $p \in [P]$ and $o \in [O]$. Then by (1), (2), (3), and the assumption that $\mathcal{H}_p$ are independent of each other for distinct $p \in [P]$, the joint log-likelihood function across all individuals is:

$$\begin{aligned}
\ell(\boldsymbol{\alpha},\boldsymbol{\Theta}) &\triangleq \sum_{p=1}^{P} \log \mathrm{P}\left(O_p;\boldsymbol{\alpha},\boldsymbol{\Theta}\right) \\
&= \sum_{p=1}^{P}\sum_{o=1}^{O}\left[\sum_{j=1}^{n_p} \log \lambda^*_{po}\left(t_{pj}\right)\cdot \mathbb{I}\left(s_{pj}=o\right) - \int_{l_p}^{u_p} \lambda^*_{po}\left(t\right) dt\right] \\
&= \sum_{p=1}^{P}\sum_{o=1}^{O}\left[\sum_{j=1}^{n_p} \left(\alpha_{po} + f^*_o(t_{pj},p;\boldsymbol{\Theta})\right)\cdot \mathbb{I}\left(s_{pj}=o\right)\right. \\
&\qquad \left. - \int_{l_p}^{u_p} \exp\left(\alpha_{po} + f^*_o(t,p;\boldsymbol{\Theta})\right) dt\right] \\
&= \sum_{p=1}^{P}\sum_{o=1}^{O}\left[\sum_{j=1}^{n_p} f^*_o(t_{pj},p;\boldsymbol{\Theta})\cdot \mathbb{I}\left(s_{pj}=o\right)\right.
\end{aligned}$$

$$+ \sum_{j=1}^{n_p} \alpha_{po} \cdot \mathbb{I}\left(s_{pj} = o\right) - \exp\left(\alpha_{po}\right) \exp\left(A_{po}(\boldsymbol{\Theta})\right) \Bigg].$$

Due to the estimation of $\boldsymbol{\alpha}$ in $\ell(\boldsymbol{\alpha}, \boldsymbol{\Theta})$, given a fixed form of $f_o^*(t, p; \boldsymbol{\Theta})$, the number of parameters that need to be estimated scales with $P \times O$. Such a scaling will lead to tremendous challenges in practice where P as well as O may be large. To resolve this issue, the idea behind the self-controlled case series study design is to implicitly compute $\hat{\boldsymbol{\alpha}}$, where for a given $\boldsymbol{\Theta}$,

$$\hat{\boldsymbol{\alpha}} \triangleq \arg\max_{\boldsymbol{\alpha}} \ell(\boldsymbol{\alpha}, \boldsymbol{\Theta}). \tag{4}$$

Notice that given $\boldsymbol{\Theta}$, $\ell(\boldsymbol{\alpha}, \boldsymbol{\Theta})$ is concave with respect to $\boldsymbol{\alpha}$. Therefore, at optimality,

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\alpha}, \boldsymbol{\Theta})}{\partial \alpha_{po}} &= \sum_{j=1}^{n_p} \mathbb{I}\left(s_{pj} = o\right) - \exp\left(\alpha_{po}\right) \exp\left(A_{po}(\boldsymbol{\Theta})\right) = 0 \\ &\Rightarrow \quad m_{po} = \exp\left(\alpha_{po}\right) \exp\left(A_{po}(\boldsymbol{\Theta})\right) \\ &\Rightarrow \quad \hat{\alpha}_{po} = \log m_{po} - A_{po}(\boldsymbol{\Theta}), \end{aligned} \tag{5}$$

where $m_{po} \triangleq \sum_{j=1}^{n_p} \mathbb{I}(s_{pj} = o)$ is the total number of occurrences of event type o during the observation time interval $[l_p, u_p]$. Therefore, $m_{po} \geq 0$. When $m_{po} = 0$, by (5), $\hat{\alpha}_{po} = -\infty$. In this case, the data from the individual p does not contribute to $\ell(\boldsymbol{\alpha}, \boldsymbol{\Theta})$ via the outcome o. Otherwise, when $m_{po} > 0$, $\hat{\alpha}_{po}$ is finite. Plugging (5) in $\ell(\boldsymbol{\alpha}, \boldsymbol{\Theta})$ yields,

$$\begin{aligned} \ell(\hat{\boldsymbol{\alpha}}, \boldsymbol{\Theta}) &= \sum_{p=1}^{P} \sum_{o=1}^{O} \Bigg[\sum_{j=1}^{n_p} \big(\log m_{po} - A_{po}(\boldsymbol{\Theta}) \\ &\qquad + f_o^*(t_{pj}, p; \boldsymbol{\Theta}) \big) \cdot \mathbb{I}\left(s_{pj} = o\right) - m_{po} \Bigg] \\ &= \sum_{p=1}^{P} \sum_{o=1}^{O} \Bigg[\sum_{j=1}^{n_p} f_o^*(t_{pj}, p; \boldsymbol{\Theta}) \cdot \mathbb{I}\left(s_{pj} = o\right) \\ &\qquad + m_{po} \log m_{po} - m_{po} A_{po}(\boldsymbol{\Theta}) - m_{po} \Bigg] \\ &= \sum_{p=1}^{P} \sum_{o=1}^{O} \Bigg[\sum_{j=1}^{n_p} f_o^*(t_{pj}, p; \boldsymbol{\Theta}) \cdot \mathbb{I}\left(s_{pj} = o\right) - m_{po} A_{po}(\boldsymbol{\Theta}) \Bigg], \end{aligned}$$

where the last equality holds up to some constant. In this way, only *cases* (i.e., individuals with at least one occurrence of any outcome $o \in [O]$) can eventually contribute to $\ell(\hat{\boldsymbol{\alpha}}, \boldsymbol{\Theta})$, and each case serves as its own control. Therefore, we call

$$\begin{aligned} \tilde{\ell}(\boldsymbol{\Theta}) \triangleq & \ell(\hat{\boldsymbol{\alpha}}, \boldsymbol{\Theta}) \\ = & \sum_{p=1}^{P} \sum_{o=1}^{O} \Bigg[\sum_{j=1}^{n_p} f_o^*(t_{pj}, p; \boldsymbol{\Theta}) \cdot \mathbb{I}\left(s_{pj} = o\right) - m_{po} A_{po}(\boldsymbol{\Theta}) \Bigg], \end{aligned} \tag{6}$$

the objective for the *neural self-controlled case series* model. The name stems from the fact that the model leverages an SCCS study design to adjust for individual heterogeneity, and uses a DNN to capture the relationship of the occurrences among various event types. Notice that $\tilde{\ell}(\boldsymbol{\Theta})$ is only dependent on $\boldsymbol{\Theta}$, and by (4) and (5), solving following optimization problem

$$\boldsymbol{\Theta}^\dagger \triangleq \arg\max_{\boldsymbol{\Theta}} \tilde{\ell}(\boldsymbol{\Theta}) \tag{7}$$

will yield the maximum likelihood estimator $\boldsymbol{\Theta}^\dagger$ for the objective $\ell(\boldsymbol{\alpha}, \boldsymbol{\Theta})$, with the corresponding MLE for $\boldsymbol{\alpha}$, denoted as $\boldsymbol{\alpha}^\dagger$, acquired via (5) by letting $\boldsymbol{\Theta} = \boldsymbol{\Theta}^\dagger$. Therefore, solving the $\boldsymbol{\alpha}$-free optimization problem in (7) is equivalent to jointly maximizing over $\boldsymbol{\alpha}$ and $\boldsymbol{\Theta}$ in $\ell(\boldsymbol{\alpha}, \boldsymbol{\Theta})$, with the benefit that (7) is a more compact problem and hence is more efficient to solve. The implications of NSCCS abound. They are discussed for the rest of this section.

In terms of the *modeling* via DNNs of time-stamped multi-type event data collected from a massive number of heterogeneous individuals, NSCCS provides a principled framework with tremendous flexibility and representation power to model the complex and nonlinear relationships among various event types while simultaneously addressing individual heterogeneity. The resultant DNN is robust to time-invariant confounding issues, and could potentially represent more causally faithful relationships compared to a standard DNN that does not adjust for time-invariant confounding issues (See Section 4 for empirical comparison). Furthermore, NSCCS is flexible because the primary requirement of $\mathbf{f}^*(t, p; \boldsymbol{\Theta})$ is that it maps the history $\mathcal{H}_p(t)$ to an $O \times 1$ vector that represents the log intensity of the occurrences of the outcomes. Here,

$$\mathbf{f}^*(t, p; \boldsymbol{\Theta}) \triangleq \begin{bmatrix} f_1^*(t, p; \boldsymbol{\Theta}) & f_2^*(t, p; \boldsymbol{\Theta}) & \cdots & f_O^*(t, p; \boldsymbol{\Theta}) \end{bmatrix}^\top.$$

Such a mild requirement opens up various possibilities in using event sequence modeling architectures [7, 23, 38] in an NSCCS setting.

In terms of the *optimization* of DNNs, NSCCS essentially provides a compact formulation of the MLE objective for training DNNs that account for individual heterogeneity. The parameterization that needs to be learned under NSCCS is exactly the same as the one for a regular DNN, and hence introduces minimal overhead during learning while enjoying the extra benefit of robustness to time-constant confounding issues. As for computational complexity, by (6), only the individuals that experience at least one occurrence of event type $o \in O$ are needed for estimation. These individuals are called cases. Therefore, the number of samples used for training only grows with respect to the number of cases. Such parsimony in the utilization of data is especially valuable in the big data setting, where the number of samples could easily be on the order of hundreds of millions, while we are only interested in predicting the occurrences of event types that are rarely observed (adverse drug reaction discovery, anomaly detection, etc.).

A concern about NSCCS compared to its aforementioned counterparts is the interpretability issue. Such a concern can be mitigated by recent advances in interpreting black box model [15, 20, 27, 30, 34, 37], as well as using interpretable architecture [1, 37] or attention mechanism [3, 39, 40] to model $\mathbf{f}^*(t, p; \boldsymbol{\Theta})$. Therefore, while interpretability is of great importance when attributing the effect of one event type on another, especially in application domains such as healthcare, NSCCS strives to offer an accurate description of the possibly unfathomable data generation process whose oversimplified interpretation might lead to potential misspecification.

3.3 Optimization

We apply a stochastic gradient descent based (SGD) algorithm to solve the optimization problem in (7). The use of SGD requires the gradient evaluation of $A_{po}(\boldsymbol{\Theta})$ defined in (3), in which the most

computationally intensive step is to evaluate $\int_{l_p}^{u_p} \exp f_o^*(t,p;\Theta)\,dt$. However, since $f_o^*(t,p;\Theta)$ is generated by a DNN, the integral in (3) generally has no closed-form solution. Numerical integration [5] could be applied for approximation the intergal; yet it is inefficient and generally suffers from large variances, resulting in low convergence rate in the SGD-based optimization.

When $f_o^*(t,p;\Theta)$ takes a piecewise constant functional form, the integral in (3) can be explicitly computed for efficient evaluation. This is the approach we take in our present formulation of NSCCS. Specifically, let I_{pk} represent the union of time intervals such that for all $t \in I_{pk}$, $f_o^*(t,p;\Theta) \equiv \gamma_{opk}$ for some γ_{opk}. Suppose that $[l_p, u_p] = \bigcup_{k=1}^{K_p} I_{pk}$, where K_p represents the total number of I_{pk}'s for the individual p, and $I_{pk} \cap I_{pk'} = \emptyset$, for all $k, k' \in [K_p]$ and $k \neq k'$. Hence $u_p - l_p = \sum_{k=1}^{K_p} |I_{pk}|$, where $|I_{pk}|$ denotes the total time spanned by I_{pk}. In this way, $A_{po}(\Theta)$ can be substantially simplified and efficiently computed as:

$$A_{po}(\Theta) = \log \sum_{k=1}^{K_p} |I_{pk}| \exp \gamma_{opk}. \tag{8}$$

3.4 Drug Eras and Piecewise Constant Conditional Intensity

A major organization promoting ADR discovery from EHRs is the Observational Health Data Sciences and Informatics (OHDSI) [11], which offers a Common Data Model (CDM) [26] to convert different EHRs into the same format so that an ADR detection algorithm can be run across various EHRs without modification to achieve proactive drug safety surveillance. The piecewise constant formulation of $f_o^*(t,p;\Theta)$ in NSCCS is very suitable to deal with EHRs in CDM format, where the drug prescription sequences in the raw EHRs as illustrated in Figure 1a are converted into corresponding *drug eras* [26], as shown in Figure 1b.

Drug eras represent periods of time a patient is believed to persistently taking a drug as directed; hence they represent relatively long-term periods of stable influence of a drug on potential condition occurrences. Using drug eras, at time t we are aware of the *binary* exposure status of different drugs for any given patient p. We hereby use $\mathbf{x}_p(t) \in \{0,1\}^D$ to represent a $D \times 1$ vector of binary drug exposure statuses, where D is the number of drugs in question (and hence $S = D + O$). Furthermore, with $\mathbf{x}_p(t)$, we can let

$$f_o^*(t,p;\Theta) \triangleq g_o(\mathbf{x}_p(t);\Theta), \tag{9}$$

where $g_o(\mathbf{x}_p(t);\Theta)$ is modeled by a DNN with the parameterization Θ. Note that given t, the number of possible configurations of $\mathbf{x}_p(t)$'s is 2^D. Therefore, $g_o(\mathbf{x}_p(t);\Theta)$ is piecewise constant, and hence $A_{po}(\Theta)$ can also be computed in a piecewise constant fashion efficiently according to (8).

4 EXPERIMENTS

In this section, we present our empirical evaluation of NSCCS on a real-world EHR dataset with a benchmark task of ADR discovery created by the Observational Medical Outcomes Partnership (OMOP) [32], the predecessor of OHDSI.

4.1 Experimental Setup

The Benchmark Task. The OMOP benchmark defines 10 major drug types and 10 adverse condition types, and provides ground truth for 53 out of $10 \times 10 = 100$ pairs. In detail, 9 drug-condition pairs are confirmed as positive cases (i.e., ADRs) and 44 pairs are as negative controls. The overall goal of the task is to test the abilities of various models to rank ADR pairs over negative control pairs.

Data Source and Cohort Design. We used a large-scale, de-identified, and IRB-approved EHR dataset collected from Marshfield Clinic[1] as a data source. The raw EHR data include condition diagnoses, drug prescriptions, laboratory procedures and results, and other vital physical and physiological measurements. We focused on the condition diagnoses and drug prescriptions, and we extracted the event sequences from the raw EHRs that contain the prescription of the 10 drug classes and/or diagnosis of 10 adverse condition classes, both based on the vocabulary definition used in the OMOP ground truth. This resulted in a cohort containing 327,842 patients with a total of 1,940,681 outcome events and 11,211,769 of drug events. For an outcome event and a drug event that share the same time-stamp (which is possible since the minimum time resolution of time-stamps in this dataset is a day), we broke the tie by always considering the outcome event to occur before the drug event. This strategy reflects the commonly observed fact that a same-day drug prescription tends to be the result of a same-day condition diagnosis. For other cases of ties, we assigned the orders among events randomly.

We then constructed time varying features $\mathbf{x}_p(t)$ from the event sequence of each patient by adopting the standard construction of drug eras [26]. Also following the standard practice, we extended each drug era by a pre-specified risk window. The risk windows we considered are `None`, `1 Month`, `3 Months`, `6 Months`, and `Lasting`, where the `Lasting` risk window means that the drug effect of each drug era is assumed to last throughout the whole remaining observation time interval of the patient.

Models for Comparison. We compared our proposed method (`NSCCS`) with three other baseline methods. These baselines can all be viewed as special cases of `NSCCS`, with either a simplified choice of α_{po} or a restrictive form of $g_o(\cdot)$. We describe them in detail as follows:

- `MSCCS` (Multiple self-controlled case series [31]): arguably the leading method for ADR discovery, `MSCCS` uses a conditional intensity of the following parametric form:

  $$\lambda_{po}^*(t) = \exp[\alpha_{po} + \boldsymbol{\beta}_o^T \mathbf{x}_p(t)].$$

  Note that this is a special case of our model achieved by setting all g_o to be linear functions. The resultant function $\mathbf{g}(\cdot)$ is equivalent to a single linear-layer neural network.
- `DNN` (Regular feed-forward neural network without accounting for individual heterogeneity): this baseline method assumes the conditional intensity has the form:

  $$\lambda_{po}^*(t) = \exp\left[\alpha_o + g_o(\mathbf{x}_p(t);\Theta)\right].$$

  The α_o is shared across all patients, and thus the base occurrence rate of each outcome o is assumed to be homogeneous.

[1] https://www.marshfieldclinic.org

- NSCCS-S (NSCCS trained on each outcome separately): this baseline is the closest to NSCCS, except that it considers each outcome separately and optimizes the loss in (7) separately. Thus, the learned g_o's do not share their hidden representations.

Other baseline methods for ADR, such as [8, 24], were not included either due to scalability issues or the restrictive single-drug-single-outcome setting.

Evaluation Metric. We computed the area under curve (AUC) of the receiver operating characteristic (ROC) curve using the OMOP ground truth and the ranking generated from each trained model. As the input to $\mathbf{g}$ is a binary vector of size D, and each dimension indicates whether the corresponding drug is in effect at the time, a reasonable quantity to measure the influence of drug d on outcome o is

$$g_o(x_d = 1, \mathbf{x}_{\neg d} = \mathbf{0}) - g_o(x_d = 0, \mathbf{x}_{\neg d} = \mathbf{0}),$$

where $\mathbf{x}_{\neg d}$ is the feature vector excluding dimension d. All statistical tests for difference of AUCs were based on the DeLong test [6].

Hyper-parameters. For NSCCS and DNN, we chose $\mathbf{g}(\cdot)$ to be a feed-forward neural network with a linear output layer. We fixed the activation functions to be tanh and optimized the loss using Adam [14] with step size 0.001. A validation set of 10,000 patients was used to monitor the training procedure and to conduct early stopping. We then tuned the following hyperparameters:

- *Network size*: networks of one, two, and three hidden layers are considered, with the number of hidden units for each layer (from input to output) being $\{32\}$, $\{32, 128\}$, and $\{32, 32, 128\}$, respectively.
- *Batch size*: the number of patients used for optimizing at each step. We considered it to be 128 or 256.
- *Patience*: the number of times of observing worse validation loss before early stopping. We considered it to be 50 or 100.

Tuning these hyperparameters, along with the five aforementioned choices of risk windows, resulted in 60 experimental configurations for NSCCS and DNN. Due to its linear form, MSCCS does not need to adjust the network size and thus only had 20 experimental configurations.

4.2 Results and Discussion

4.2.1 Overall Performance. In existing literature about developing methods for ADR discovery, it is customary to examine the overall performance of the developed methods under various experimental configurations [21, 25, 29, 33]. Thus in Figure 2 we show the boxplot for the distributions of AUCs of NSCCS and of the other three baseline methods, where the box shows the quartiles of data points while the whiskers extend to show the rest of the distribution, except for "outlier" points that are past the low and high quartile by 1.5 interquartile range. Both versions of NSCCS significantly outperform MSCCS on AUC by substantial margins, both in terms of the median and the two quantiles ($p < 0.05$). This confirms the importance of adopting a flexible $\mathbf{f}^*$ for modeling the complex influence of history on the occurrence rate of outcome. DNN, due to its failing to incorporate individual-specific baselines, has much lower AUC than alternative methods, which emphasizes the necessity of addressing individual heterogeneity in real-world, large-scale EHR datasets. For NSCCS-S, while its AUCs have a higher third quartile

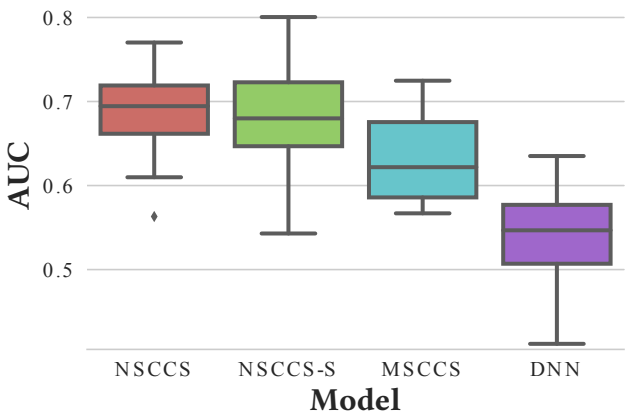


Figure 2: Overall performance of NSCCS and baselines measured by AUC among different experimental configurations.

than NSCCS, NSCCS exhibits higher median and smaller variance of AUCs, which are also important advantages when deploying ADR discovery methods in practice.

4.2.2 Best Performer. Figure 3 shows the ranking of adverse drug reactions assigned by the three methods with their best configurations summarized in Table 1. Across all 9 ADR pairs, the rankings assigned by NSCCS consistently attain (or are closest to) the best rankings.

Table 1: The best configurations of the four methods that attain the best performance on AUC.

Method	# of Layers	Risk Window	Batch Size	Patience
NSCCS	3	6 Months	128	50
NSCCS-S	2	3 Months	128	100
MSCCS	N/A	Lasting	256	50
DNN	3	3 Months	256	100

4.2.3 Model Selection and Generalization. To validate how well NSCCS can *generalize* to unseen adverse conditions, we performed leave-one-condition-out cross-validation (LOCOCV). Following the same procedure described in [2, 18], for each outcome o, we selected the best configuration of a model across all its experimental configurations in a sub-dataset with outcome o left out. We then trained the model with the best configuration on the full dataset and obtained the ranking of drugs for outcome o. Repeating this procedure O times yielded predictions on all outcomes and enables the evaluation of LOCOCV AUCs. As seen in Table 2, the LOCOCV AUC of NSCCS is significantly higher than those of all three baselines ($p < 0.05$) and exceeds the runner-up by a margin of over 0.1.

4.2.4 Sensitivity Analysis. We also investigated how sensitive the performance of NSCCS is with respect to three set of experimental configurations: (a) network complexity, (b) risk window, and (c) optimization.

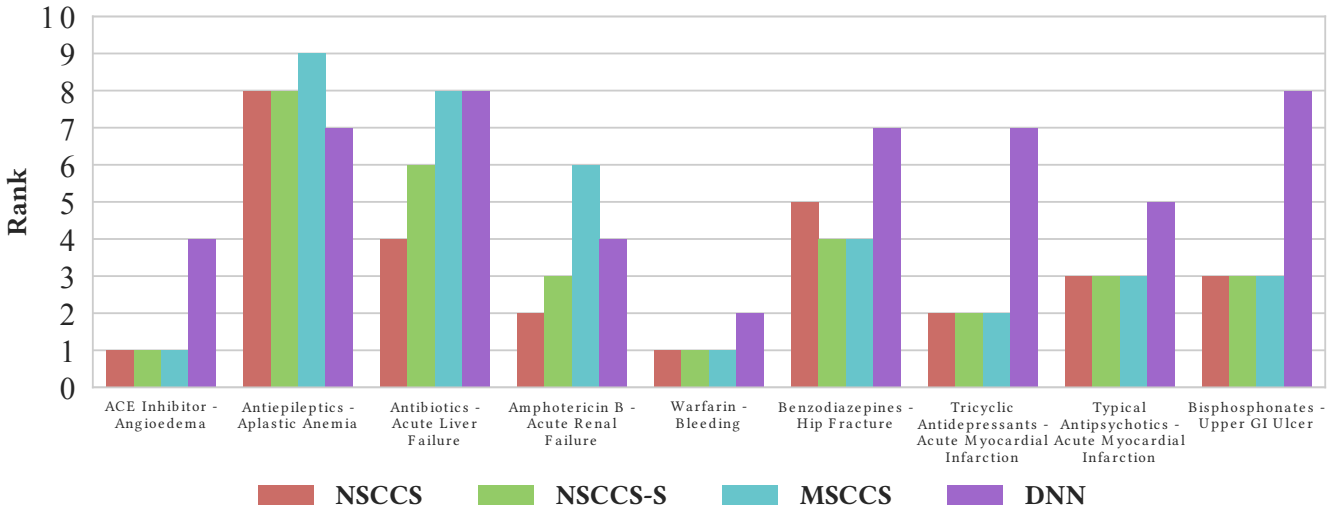
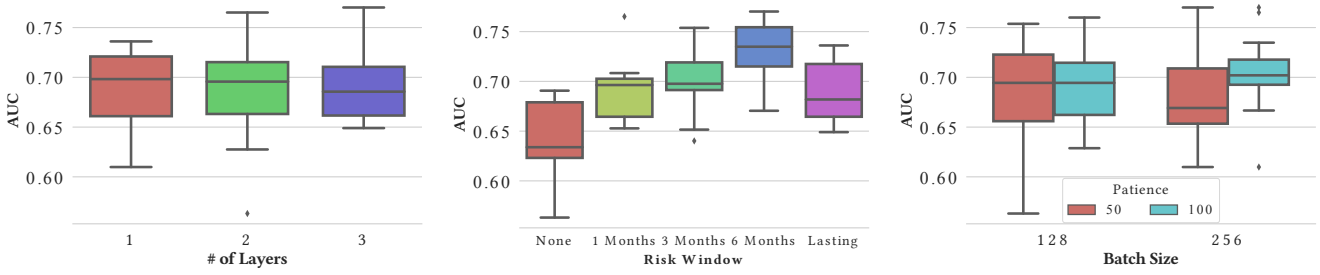


Figure 3: Ranks of true ADR-causing drugs among all ten drugs for 9 true ADR pairs. Across all 9 ADR pairs, the rankings assigned by NSCCS consistently attain (or are closest to) the best rankings.



(a) Network Complexity **(b) Risk Window** **(c) Optimization**

Figure 4: Sensitivity analysis of the performance of NSCCS on AUC with respect to various experimental configurations.

Table 2: Performance of leave-one-condition-out cross-validation (LOCOCV) for four methods.

Method	NSCCS	NSCCS-S	MSCCS	DNN
AUC	0.779	0.663	0.665	0.528

Regarding the impact of network complexity, Figure 4a shows that with the increase in the number of hidden layers, the AUC distribution of NSCCS shifts towards higher values, and the AUCs of the top performers for n hidden layers consistently increase as the n increases. Therefore, the AUC of NSCCS in general benefits from the increased network complexity. This phenomenon is consistent with the intuition that increased network complexity is beneficial for NSCCS to capture more complicated dependencies among occurrences of different event types, and hence to identify associations that are more causally credible. Nonetheless, such benefits might not be as substantial if the current network architecture is complicated enough already to capture the important interactions. This perspective can be confirmed by the fact that the median AUC for the networks with two hidden layers is higher than the median AUC for the networks with three hiddenlayers. Yet We also notice that with the exception of one outlier, even the poorest performers benefit from the increase of network complexity as well. This suggests that nonlinear dependencies among various event types seem to be intrinsic to the dataset in question, and increasing the network complexity is arguably a safe bet to ensure that these dependencies are captured.

Regarding the impact of risk window, Figure 4b shows that as the length of the risk windows increases from None to 6 Months, the performance of NSCCS on AUC also increases. This positive association implies that a reasonable risk window design is beneficial for NSCCS to capture ADR signals. On the other hand, when the Lasting risk window is used, the performance of NSCCS drops. A reasonable explanation is that NSCCS is capable of modeling highly

nonlinear interactions among distinct event types. However, when the Lasting risk window is in use, any drug prescription that occurs even in the distant past could still have unattenuated interactions with a much more recent prescription of a different drug. Such a feature construction via the Lasting risk window is hence prone to introduce spurious correlations that are captured by the highly expressive NSCCS. This justification also explains why linear methods that do not model interaction between different drugs, such as MSCCS, tends to perform well when a Lasting risk window is in use [31, 33]. Therefore, for those linear models the main effect of various drugs can be accumulated over time through the Lasting risk window design while avoiding the over-fitting issue induced by spurious correlations.

Finally, regarding the impact of optimization, Figure 4c shows that when the smaller batch size of 128 is in use, we observe higher variance in the AUC distribution, compared to that of the larger batch size of 256. This observation is consistent with the well-known behavior of stochastic-gradient-based methods: larger batch size helps to reduce the variance of the stochastic gradient, improving the quality of the stochastic approximation to the exact gradient. As also shown in Figure 4c, when the optimization algorithm is more patient, the performance of NSCCS on AUC tends to be better. This phenomenon is also somewhat expected as the optimization problem of NSCCS is highly nonlinear and non-convex, and a higher patience threshold implies that the algorithm tends to explore more in the hope of finding a better solution before early stopping.

4.2.5 Towards Temporary Heterogeneity. While NSCCS strictly follows an SCCS design and controls for the time-invariant confounders that act multiplicatively on the conditional intensities, its temporal homogeneity assumption may still be restrictive, due to the longitudinal nature of the data. In particular, Kuang et al. [18] consider the following conditional intensity form:

$$\lambda^*_{po}(t) = \exp[\alpha_{po}(t) + \boldsymbol{\beta}_o^T \mathbf{x}_p(t)],$$

where $\alpha_{po}(t)$ is the patient-specific time-varying baseline to estimate. They further introduce Baseline Regularization (BR) and other techniques, such as parameter tying, to avoid overfitting; they show the improvement of BR over its counterpart MSCCS.

Although NSCCS and BR make orthogonal extensions to MSCCS: non-linear models and time-varying baselines, respectively, we still desire to investigate how incorporating the modeling of time-varying baseline would affect the performance of NSCCS. Without rebuilding the existing formulation from scratch, we approximate this goal by adding an extra preprocessing step for constructing the cohort, analogous to the construction of time-dependent strata commonly used in longitudinal survival analysis [13]: we divided a patient's event sequence into multiple subsequences such that no consecutive events in the same subsequences are far apart by a time gap τ, and subsequences from the same patient were treated as from independent patients. As a result, each subsequence would correspond to a baseline parameter and be implicitly adjusted, and time-varying confounder issues can thereby be partly mitigated. We chose $\tau = 2$ year and trained NSCCS on the newly constructed cohort with the same 60 aforementioned configurations. For BR, we followed the same design as Kuang et al. [18], except for not imposing the the minimum duration constraint on each patient's observation window, resulting in 216 configurations.

With this amendment to allow temporal heterogeneity, NSCCS achieves a maximum AUC of 0.83 and a median of AUC of 0.73, improved from 0.77 and 0.70, respectively, as opposed to the maximum AUC of 0.81 and median AUC of 0.65 achieved by BR, which suggests that NSCCS can benefit from incorporating the time-varying individual baselines on ADR discovery. In the long run, we plan to incorporate temporal heterogeneity in a manner more tightly following BR, to see if that further improves performance. Bringing the approach from BR, which solves a convex problem, into non-convex NSCCS will require substantial further work.

5 CONCLUSION

We have proposed NSCCS for ADR discovery from EHRs. Our contribution to DNNs is to endow them with robustness to time-invariant confounders caused by individual heterogeneity; our contribution to self-controlled case series, which build linear models, is to enable learning of complex non-linear interactions that can influence outcomes such as ADRs. We conduct a comprehensive empirical evaluation on a real-world large-scale EHR with a benchmark ADR discovery task, and demonstrate that NSCCS achieves superior performance compared to other competitive baselines. NSCCS is hence an effective approach to identify ADR signals from EHR. In addition, because discovering causally faithful associations from longitudinal event data is a problem that applies to many domains, we anticipate this approach will also have applications beyond ADR discovery [10, 16, 17, 19, 41]. Future directions include, among others, adjusting for time-varying confounding, developing more sophisticated architectures to learn $f^*_o(t, ; \boldsymbol{\Theta})$ for all outcome o, and interpreting a DNN using higher order interaction intelligible models and justifying its effectiveness.

Acknowledgements

The authors gratefully thank the anonymous reviewers for their helpful feedback. This work was supported by NIGMS grant 2RO1 GM097618.

REFERENCES

[1] David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing System*. Curran Associates Inc., 7786–7795. https://doi.org/10.1016/j.jdent.2008.01.001 arXiv:1806.07538

[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*, Vol. 68. 1–14.

[3] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing System*. 3504–3512. arXiv:1608.05745

[4] D J Daley and D Vere-Jones. 2003. *An Introduction to the Theory of Point Processes* (second ed.). Probability and its Applications (New York), Vol. I. Springer-Verlag, New York. https://doi.org/10.1007/978-0-387-49835-5 arXiv:arXiv:1011.1669v3

[5] Philip J Davis and Philip Rabinowitz. 2007. *Methods of Numerical Integration*. Courier Corporation.

[6] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 44, 3 (1988), 837–845. https://kaigi.org/jsai/webprogram/2017/pdf/814.pdf

[7] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 2016. Recurrent Marked Temporal Point Processes:

Embedding Event History to Vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1555–1564. https://doi.org/10.1145/2939672.2939875 arXiv:arXiv:1508.06655v1

[8] S Escolano, C Hill, and P Tubert-Bitter. 2013. A New Self-Controlled Case Series Method for Analyzing Spontaneous Reports of Adverse Events After Vaccination. *American Journal of Epidemiology* 178, 9 (nov 2013), 1496–1504. https://doi.org/10.1093/aje/kwt128

[9] C. P. Farrington and H. J. Whitaker. 2006. Semiparametric analysis of case series data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 55, 5 (nov 2006), 553–594.

[10] Sinong Geng, Zhaobin Kuang, Peggy Peissig, and David Page. 2018. Temporal poisson square root graphical models. *Proceedings of machine learning research* 80 (2018), 1714.

[11] Hripcsaka George, Jon D Duke, Nigam H Shah, and Et Al. 2015. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in Health Technology and Informatics* 216 (2015), 574–578. https://doi.org/10.3233/978-1-61499-564-7-574

[12] Rave Harpaz, William DuMouchel, Nigam H Shah, David Madigan, Patrick Ryan, Carol Friedman, and Author Manuscript. 2012. Novel Data Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology and Therapeutics* 91, 6 (jun 2012), 1010–1021. https://doi.org/10.1038/clpt.2012.50. Novel arXiv:NIHMS150003

[13] John D Kalbfleisch and Ross L Prentice. 2011. Time Dependence in the Relative Risk Model. In *Statistical Analysis of Failure Time Data*. Vol. 360. John Wiley & Sons, Chapter 6.4, 200–208.

[14] Diederik P Kingma and Jimmy Lei Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint* (2014), 1–15. arXiv:1412.6980

[15] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, Vol. 4. 2976–2987. arXiv:1703.04730 http://arxiv.org/abs/1703.04730

[16] Zhaobin Kuang. 2018. *Towards Learning with High Causal Fidelity from Longitudinal Event Data*. Ph.D. Dissertation. University of Wisconsin–Madison.

[17] Zhaobin Kuang, Yujia Bao, James Thomson, Michael Caldwell, Peggy Peissig, Ron Stewart, Rebecca Willett, and David Page. 2019. A machine-learning-based drug repurposing approach using baseline regularization. In *Computational Methods for Drug Repurposing*. Springer, 255–267.

[18] Zhaobin Kuang, Peggy Peissig, Vitor Santos Costa, Richard Maclin, and David Page. 2017. Pharmacovigilance via Baseline Regularization with Large-Scale Longitudinal Observational Data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1537–1546.

[19] Zhaobin Kuang, James Thomson, Michael Caldwell, Peggy Peissig, Ron Stewart, and David Page. 2016. Baseline regularization for computational drug repositioning with longitudinal observational data. In *IJCAI: proceedings of the conference*, Vol. 2016. NIH Public Access, 2521.

[20] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing System*. 4765–4774. arXiv:1705.07874 http://arxiv.org/abs/1705.07874

[21] David Madigan, Martijn J Schuemie, and Patrick B Ryan. 2013. Empirical performance of the case-control method: lessons for developing a risk identification and analysis system. *Drug Safety* 36, 1 (2013), 73–82. https://doi.org/10.1007/s40264-013-0105-z

[22] Lara Magro, Ugo Moretti, and Roberto Leone. 2012. Epidemiology and Characteristics of Adverse Drug Reactions Caused by Drug-Drug Interactions. *Expert Opinion on Drug Safety* 11, 1 (jan 2012), 83–94. https://doi.org/10.1517/14740338.2012.631910

[23] Hongyuan Mei and Jason M Eisner. 2017. The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process. In *Advances in Neural Information Processing System*. Long Beach, 1–21. arXiv:1612.09328 https://arxiv.org/abs/1612.09328http://arxiv.org/abs/1612.09328

[24] Ramin Moghaddass, Cynthia Rudin, and David Madigan. 2016. The Factorized Self-Controlled Case Series Method: An Approach for Estimating the Effects of Many Drugs on Many Outcomes. *Journal of Machine Learning Research* 17, 1 (2016), 1–24.

[25] G Niklas Norén, Tomas Bergvall, Patrick B Ryan, Kristina Juhlin, Martijn J Schuemie, and David Madigan. 2013. Empirical Performance of the Calibrated Self-Controlled Cohort Analysis within Temporal Pattern Discovery: Lessons for Developing a Risk Identification and Analysis System. *Drug Safety* 36, 1 (2013), 107–121. https://doi.org/10.1007/s40264-013-0095-x

[26] Stephanie J Reisinger, Patrick B Ryan, Donald J O'Hara, Gregory E Powell, Jeffery L Painter, Edward N Pattishall, and Jonathan A Morris. 2010. Development and Evaluation of a Common Data Model Enabling Active Drug Safety Surveillance Using Disparate Healthcare Databases. *Journal of the American Medical Informatics Association* 17, 6 (2010), 652–662. https://doi.org/10.1136/jamia.2009.002477

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144. https://doi.org/10.1145/1235 arXiv:1602.04938

[28] Melissa A Robb, Judith A Racoosin, Rachel E Sherman, Thomas P Gross, Robert Ball, Marsha E Reichman, Karen Midthun, and Janet Woodcock. 2012. The US Food and Drug Administration's Sentinel Initiative: Expanding the Horizons of Medical Product Safety. *Pharmacoepidemiology and Drug Safety* 21, S1 (2012), 9–11.

[29] Patrick B Ryan, Paul E Stang, J Marc Overhage, Marc A Suchard, Abraham G Hartzema, William DuMouchel, Christian G Reich, Martijn J Schuemie, and David Madigan. 2013. A Comparison of the Empirical Performance of Methods for a Risk Identification System. *Drug Safety* 36, SUPPL.1 (2013). https://doi.org/10.1007/s40264-013-0108-9

[30] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *International Conference on Machine Learning*. JMLR. org, 3145–3153. arXiv:1704.02685 http://arxiv.org/abs/1704.02685

[31] Shawn E Simpson, David Madigan, Ivan Zorych, Martijn J Schuemie, Patrick B Ryan, and Marc A Suchard. 2013. Multiple Self-Controlled Case Series for Large-Scale Longitudinal Observational Databases. *Biometrics* 69, 4 (dec 2013), 893–902. https://doi.org/10.1111/biom.12078

[32] Paul E Stang, Patrick B Ryan, Judith a Racoosin, and J Marc Overhage. 2010. Research and Reporting Methods Annals of Internal Medicine Advancing the Science for Active Surveillance : Rationale and Design for the Observational Medical Outcomes Partnership. *Annals of Internal Medicine* 153, 9 (2010), 600–606. https://doi.org/10.7326/0003-4819-153-9-201011020-00010

[33] Marc A Suchard, Ivan Zorych, Shawn E Simpson, Martijn J Schuemie, Patrick B Ryan, and David Madigan. 2013. Empirical Performance of the Self-Controlled Case Series Design: Lessons for Developing a Risk Identification and Analysis System. *Drug Safety* 36, 1 (2013), 83–93. https://doi.org/10.1007/s40264-013-0100-4

[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*. JMLR. org, 3319–3328. https://doi.org/10.1007/s10144-009-0162-4 arXiv:1703.01365

[35] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing System*. 3104–3112. https://doi.org/10.1007/s10107-014-0839-0 arXiv:1409.3215

[36] Nicholas P Tatonetti, P Ye Patrick, Roxana Daneshjou, and Russ B Altman. 2012. Data-Driven Prediction of Drug Effects and Interactions. *Science Translational Medicine* 4, 125 (2012), 125ra31————-125ra31. https://doi.org/10.1126/scitranslmed.3003377.Data-Driven

[37] Michael Tsang, Dehua Cheng, and Yan Liu. 2018. Detecting Statistical Interactions from Neural Network Weights. In *6th International Conference on Learning Representations, {ICLR} 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. https://openreview.net/forum?id=ByOfBggRZ

[38] Shuai Xiao, Junchi Yan, Stephen M. Chu, Xiaokang Yang, Hongyuan Zha, and Stephen M. Chu. 2017. Modeling the Intensity Function of Point Process via Recurrent Neural Networks. In *Proceedings of the 31st Conference on Artificial Intelligence (AAAI)*. 1597–1603. arXiv:1705.08982 http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14391http://arxiv.org/abs/1705.08982

[39] Shuai Xiao, Junchi Yan, Mehrdad Farajtabar, Le Song, Xiaokang Yang, and Hongyuan Zha. 2019. Learning Time Series Associated Event Sequences With Recurrent Point Process Networks. *IEEE Transactions on Neural Networks and Learning Systems* PP (2019), 1–13. https://doi.org/10.1109/TNNLS.2018.2889776

[40] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *International Conference on Machine Learning* (2015), 2048—-2057. https://doi.org/10.1109/72.279181 arXiv:1502.03044

[41] Wei Zhang, Thomas Kobber Panum, Somesh Jha, Prasad Chalasani, and David Page. 2020. CAUSE: Learning Granger Causality from Event Sequences using Attribution Methods. *arXiv preprint arXiv:2002.07906* (2020).

A THE DERIVATION OF THE LIKELIHOOD IN (1)

As we consider the p-th individual, the subscript p is dropped for notation simplicity. Suppose i outcome events have been observed and the time-stamp of the i-th outcome is t_i. Let the $T > t$ and $S \in [O]$ be the random variable of the time-stamp and the type of the next outcome, respectively. By the definition of conditional intensity, the probability of no outcome o occurring (i.e. *surviving*) in $[t, t + dt)$ is $1 - \lambda_o^*(t)dt$. When dt is an infinitesimal quantity, this probability is equal to $\exp(-\lambda_o^*(t)dt)$. As a result, the cumulative distribution function of T, conditioned on the past history $\mathcal{H}(t)$, is

given by

$$F^*(t) = 1 - P(T > t|\mathcal{H}(t)) = 1 - \exp\left(-\int_{t_i}^{t} \sum_o \lambda_o^*(t')dt'\right).$$

Taking the derivative of $F^*(t)$ then yields the conditional probability density function of T as

$$f^*(t) = \frac{dF^*(t)}{dt} = \exp\left(-\int_{t_i}^{t} \sum_o \lambda_o^*(t')dt'\right) \sum_o \lambda_o^*(t).$$

Given $T = t$ and $\mathcal{H}(t)$, the conditional probability of the type of the new outcome is $P(S = o|T = t, \mathcal{H}(t)) = \lambda_o^*(t)/\sum_{o'} \lambda_{o'}^*(t)$. Combined with (A), the probability of observing the new event at t with type o is

$$\Pr(T = t, O = o|\mathcal{H}(t)) = \prod_{o'} \exp\left(\log \lambda_{o'}^*(t) \cdot \mathbb{I}(o = o') - \int_{t_i}^{t} \lambda_{o'}^*(t')dt'\right).$$

Applying the chain rule and also considering the surviving period between the last outcome and the end of the observation time interval u, we obtain the likelihood for the whole outcome event sequence as in (1).