Hawkes Process Modeling of Adverse Drug Reactions with Longitudinal Observational Data

Yujia Bao YUJIA@CS.WISC.EDU
Department of Mathematics
University of Wisconsin-Madison
Madison, WI, USA

Zhaobin Kuang ZKUANG@WISC.EDU
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI, USA

Peggy Peissig PEISSIG.PEGGY@MCRF.MFLDCLIN.EDU
Marshfield Clinic Research Foundation
Marshfield, WI, USA

David Page PAGE@BIOSTAT.WISC.EDU
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
Madison, WI, USA

Rebecca Willett WILLETT@DISCOVERY.WISC.EDU
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
Madison, WI, USA

Abstract

Adverse drug reaction (ADR) discovery is the task of identifying unexpected and negative events caused by pharmaceutical products. This paper describes a log-linear Hawkes process model for ADR discovery from longitudinal observational data such as electronic health records (EHRs). The proposed method leverages the irregular time-stamped events in EHRs to represent the time-varying effect of various drugs on the occurrence rate of adverse events. Experimental results on a large-scale cohort of real-world EHRs demonstrate that the proposed method outperforms a leading approach, multiple self-controlled case series (Simpson et al., 2013), in identifying benchmark ADRs defined by the Observational Medical Outcomes Partnership.

1. Introduction

Adverse drug reaction (ADR) discovery is the task of finding unexpected and negative effects of drugs prescribed to patients. ADR discovery is a major public health challenge. It is estimated that ADRs cause 4.2-30% of hospitalizations in the United States and Canada, with an approximated relevant annual cost of 30.1 billion US dollars in the United States (Sultana et al., 2013). Although the U.S. Food and Drug Administration (FDA) has established one of the most rigorous drug preapproval procedures in the world, many potential

ADRs of a drug may not be identified in its developmental stage. During the preapproval clinical trials, a drug might be tested on just a few thousand people. Therefore, ADRs with low occurrence rates are likely not to be identified in this relatively small population. However, these ADRs might occur and even become a public health hazard after the drug is introduced to the market, where potentially millions of people with much more diverse profiles are taking the drug. Therefore, postmarketing surveillance methods that can quickly and effectively detect potential ADRs are highly desirable to address this major public health challenge.

Modern postmarketing surveillance (Robb et al., 2012; Findlay, 2015; Hripcsak et al., 2015) leverages machine learning and data mining algorithms for ADR discovery (Ryan et al., 2012; Norén et al., 2013a,b; Ryan et al., 2013a,b; Schuemie et al., 2013; Suchard et al., 2013) on large-scale longitudinal observational databases (LODs) such as insurance claim databases and electronic health records (EHRs), where drug prescription records, adverse health outcome occurrences, and demographic information from millions of individuals are collected as time-event pairs. A leading model used for ADR discovery from LODs is the multiple self-controlled case series (MSCCS, Simpson et al. 2013). In MSCCS, we only consider individuals with at least one occurrence of an adverse health outcome of interest as cases. By estimating the occurrence rates of the adverse events when the individuals are exposed (or not exposed) to various drugs, each individual can serve as his/her own control, potentially linking the elevation of the occurrence rate of adverse events to the exposure of particular drugs and providing evidence for ADR discovery.

While MSCCS has gained tremendous empirical success (Simpson et al., 2013; Suchard et al., 2013) in identifying benchmark ADRs defined by the Observational Medical Outcomes Partnership (OMOP), the model relies on somewhat restrictive assumptions:

- **Drug Era Construction**: In MSCCS, upon the prescription of a drug to a patient, the patient is assumed to be under the exposure of the drug for a continuous period of time called drug era (Reisinger et al., 2010). Since in most EHRs, only the time-stamped drug prescription records are available, drug eras are usually constructed manually based on heuristics that incorporate adjacent time-stamped drug prescription records of the same drug. A data-driven, drug-era-free approach that directly leverages the time-stamped information in the EHRs is hence highly desirable to represent the influence of a particular drug upon the occurrence of an adverse event.

- **Time-Invariant Drug Effect**: Standard MSCCS also assumes that during a drug era, the effect of the drug on the occurrence rate of the adverse event remains constant. This obviously is an over-simplification in practice, as different drugs exhibit different pharmacokinetics and exert different dynamic impacts at different times. While efforts have been made to extend self-controlled case series to address time-varying drug effects for a single drug (Schuemie et al., 2016), modeling time-varying drug effects on adverse events for multiple drugs in large-scale LODs remains underdeveloped.

To circumvent the aforementioned predicaments of MSCCS, we propose a log-linear Hawkes process (Hawkes, 1971a,b) for adverse drug reaction discovery with longitudinal observational data. A central component of the Hawkes process is its flexible representation

power to depict self-excitation and mutual-excitation of past events of various types to future events via triggering influence functions. Specifically, we propose using dyadic influence functions in lieu of the construction of drug eras to represent the effect of a drug on the future occurrence rate of an adverse event. In this way, the influence of a drug on an adverse event is modulated by the gap between the drug prescription time and the adverse event occurrence.

To the best of our knowledge, this work is the first attempt to model longitudinal observational data as a log-linear Hawkes process for ADR discovery. Experimental results on a real-world EHR demonstrate that the proposed method outperforms MSCCS in various settings.

2. Modeling framework

For each patient $p \in \{1, \ldots, P\}$, we observe $N_p > 0$ events. The i^{th} event is described by its time, $\tau_{p,i}$, and type, $m_{p,i}$, where $\tau_{p,i} \leq \tau_{p,i+1}$ for $i = 1, 2, \ldots, N_p - 1$. The times are generally discretized by EHR software to be accurate within eight hours. Assuming a sampling period of length $\Delta = 8$ hours, we let $x_{p,m,t}$ be the number of events at any time $\tau \in [\Delta t, \Delta(t+1))$ of type m for patient p. Event types m belong to a set $\mathcal{M} = \mathcal{D} \cup \mathcal{O}$, where $\mathcal{D}$ is the set of possible drug prescription events and $\mathcal{O}$ is the set of adverse health outcomes.

A complicating factor in predicting ADRs is that we do not know when a patient is actively taking a drug; we can only observe when the drug is prescribed, and different prescriptions can have different durations. This challenge has been noted before (Kuang et al., 2016a). A heuristic proposed in the Common Data Model (CDM, Reisinger et al. 2010) by Observational Medical Outcome Partnership (OMOP) is to assume that each drug has a *time-at-risk window*, which is comprised of (a) the drug era, or the times when a patient is assumed to be taking a drug based on the prescription date recorded in the EHR, and (b) the drug exposure window, or the times when a patient is assumed to still be reacting to a drug even though the prescription has ended.

In this paper, we denote the length of the time-at-risk window as L. That is, L is a measure of real time (hours), and L/Δ is a measure of the number of discrete time intervals (*e.g.,* 8-hours periods) in which the EHR data is stored.

Throughout this paper, we model the outcome events as realizations of a point process with time-varying rate λ. ADR analysis is the process of estimating λ from data and determining which factors from a patient's EHR most contribute accurate predictions of ADRs. In the below, we first describe the commonly-used *Multiple Self-Controlled Case Series* (MSCCS, Simpson et al. 2013) and then our proposed *log-linear Hawkes* model.

2.1 Multiple Self-Controlled Case Series Model

Multiple self-controlled case series (MSCCS, Simpson et al. 2013) is one of the leading methods for ADR discovery. Given L, the MSCCS model can be specified as follows. First, define

$$\tilde{x}_{p,m,t} := \begin{cases} 1, & \text{if } \exists s \in \{t - L/\Delta + 1, \ldots, t\} \text{ such that } x_{p,m,s} > 0 \\ 0, & \text{otherwise} \end{cases};$$

then $\tilde{x}_{p,m,t}$ indicates whether patient p was prescribed drug m at any point in the past L/Δ time units up until time t. We may then model the log-rate of ADR $o \in \mathcal{O}$ for patient p at time t as

$$\log \lambda_{p,o,t} = \Delta b_{p,o} + \sum_{d \in \mathcal{D}} w_{o,d} \tilde{x}_{p,d,t} \tag{MSCCS}$$

for some unknown weights $\{w_{o,d}\}_{d \in \mathcal{D}}$ and unknown baseline event rate $b_{p,o}$, which can be different for each patient.

Given this rate, we model our observations of ADRs using a Poisson distribution, so that the probability of patient p experiencing outcome o at time t is

$$\mathbb{P}(x_{p,o,t} | \lambda_{p,o,t}) = \frac{e^{-\lambda_{p,o,t}} \lambda_{p,o,t}^{x_{p,o,t}}}{x_{p,o,t}!}. \tag{1}$$

The model in (MSCCS) says that the log of this rate parameter is the sum of a patient-specific baseline rate and a weighted combination of the different events the patient is simultaneously experiencing. *The weights $\{w_{o,d}\}_{\substack{o \in \mathcal{O} \\ d \in \mathcal{D}}}$ indicate how well we may predict outcome o based on a patient being on drug d.*

While this model is popular in the literature and practice (Simpson et al., 2013; Suchard et al., 2013), choosing the time-at-risk window L can still confound analysis. The time-at-risk window L is generally chosen based on side information about common drug prescription durations, or is treated as a tuning parameter to be chosen based on data. If L is small, then the model behaves as if the patient is not on the drug L hours after the prescription is recorded, thus potentially masking longer-term causal effects. In contrast, if L is large, then it is difficult to distinguish the effect of a drug prescribed recently and a different drug prescribed in the distant past; in fact, MSCCS would treat those prescriptions as equal.

Another interpretation of the MSCCS model is that EHRs have *missing data* about which drugs patients are taking at what times. The $\tilde{x}_{p,d,t}$'s can be considered a combination of the original data and *imputed events* which may or may not be real. *Injecting artificial events into a patient's EHR poses significant risks for biased analysis leading to false conclusions.*

2.2 Hawkes model

We propose to use a Hawkes process model (Hawkes, 1971a,b; Daley and Vere-Jones, 2003) as an alternative to MSCCS. Hawkes processes have been used to model spike trains recorded from biological neural networks (Pillow et al., 2008), interactions within a social network (Hall and Willett, 2013), pricing changes within financial networks (Chavez-Demoulin and McGill, 2012), power failures in networked electrical systems (Ertekin et al., 2015), crime and military engagements (Linderman and Adams, 2014), and in a variety of other settings.

The log-linear Hawkes process shares several features with MSCCS, but is a continuous-time model that can *account for the influence of past events on future events.* To specify the Hawkes process, we first define a collection of K *influence functions*, $\{\phi_k(\cdot)\}_{k=0}^{K-1}$. The Hawkes process can be expressed in terms of any influence functions, and we describe our

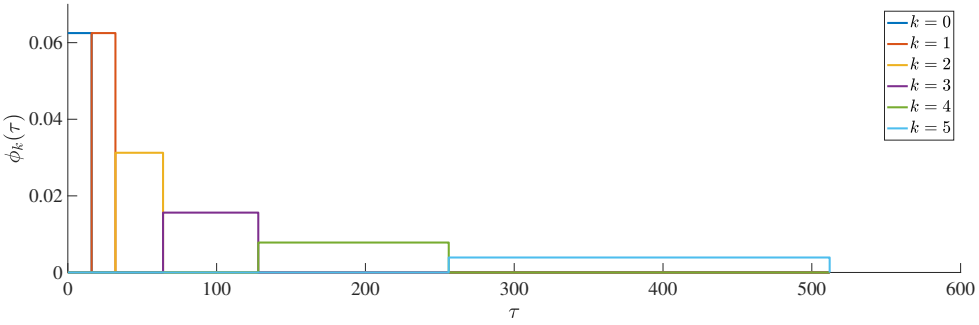

Figure 1: Dyadic influence functions for $L = 512$ and $K = 6$.

specific choice for ADR analysis in Section 2.2.1. Given these influence functions, the log-rate of the Hawkes process has the form

$$\log \lambda_{p,o}(\tau) = b_{p,o} + \sum_{d \in \mathcal{D}} \sum_{k=0}^{K-1} w_{o,d,k} \sum_{\substack{i \le N_p: \\ \tau_{p,i} \le \tau \\ m_{p,i} = d}} \phi_k(\tau - \tau_{p,i}). \tag{2}$$

Similar to MSCCS, the log of this rate parameter is the sum of a patient-specific baseline rate and a weighted combination of patient features. Unlike MSCCS, *the Hawkes model naturally accounts for the influence of past events without requiring the analyst to inject artificial events into a patient's EHR to account for the (unknown) time-at-risk window.* The weights $\{w_{o,d,k}\}_{\substack{o \in \mathcal{O},\, d \in \mathcal{D}, \\ k \in \{1,\ldots,K\}}}$ indicate how well we may predict outcome o based on a patient being on drug d according to the k^{th} influence function. Each influence function reflects how the influence of a past event changes based on how much time has elapsed since that event. We generally expect that more recent events have more bearing on a patient's risk of an ADR.

2.2.1 Choice of Influence Functions

For ADR analysis, we propose choosing the influence functions (ϕ_k's) to be piecewise constant functions supported on bounded intervals. Specifically, let K be the number of influence functions in our Hawkes model, and let L be the length of the maximum time-at-risk window.

For each k, we define an interval $I_k = [a_k, b_k)$ that satisfies the constraint that the collection of all K intervals cover the entire time-at-risk window $[0, L)$ (that is, $\bigcup_{k=0}^{K-1} I_k = [0, L)$). Then

$$\phi_k = \frac{1}{b_k - a_k} \mathbf{1}_{\{\tau \in I_k\}}.$$

Note that these ϕ_k's all integrate to one and are orthogonal to one another. By picking different pairs (a_k, b_k), we can jointly model short-term and long-term effects.

In our experiments, we focus on ϕ_k's where the intervals are chosen as follows. Define

$$\alpha_k := \begin{cases} 2^{K-1}/L, & k = 0, \\ 2^{K-k}/L, & k = 1, \ldots, K-1, \end{cases}$$

and the intervals

$$I_k := \begin{cases} \left[0, 1/\alpha_k\right), & k = 0, \\ \left[1/\alpha_k, 2/\alpha_k\right), & k = 1, \ldots, K-1 \end{cases}.$$

Then we define

$$\phi_k(\tau) = \alpha_k \mathbf{1}_{\{\tau \in I_k\}}$$

where $\mathbf{1}_{\{A\}} = \begin{cases} 1, & A \text{ true}, \\ 0, & A \text{ false} \end{cases}$ is the indicator function.

We refer to the above choice of influence function functions (depicted in Figure 1) as *dyadic influence functions* because they are supported on dyadic intervals that correspond to dividing the interval $[0, L)$ in half repeatedly.

2.2.2 Hawkes processes with dyadic influence functions

In this subsection, we examine the Hawkes model of (2) in the specific case of dyadic influence functions. In particular, we note that for $\tau \in [\Delta t, \Delta(t+1))$,

$$\sum_{\substack{i \le N_p: \\ \tau_{p,i} \le \tau \\ m_i = d}} \phi_k(\tau - \tau_{p,i}) = \sum_{\substack{i \le N_p: \\ \tau_{p,i} \le \tau \\ m_i = d}} \alpha_k \mathbf{1}_{\tau - \tau_{p,i} \in I_k} = \alpha_k \sum_{s:(t-s)\Delta \in I_k} x_{p,d,s}.$$

Define

$$z_{p,d,t,k} := \Delta \alpha_k \sum_{s:(t-s)\Delta \in I_k} x_{p,d,s}.$$

Then by sampling (2) via integration over intervals of length Δ, we have

$$\log \lambda_{p,o,t} = \Delta b_{p,o} + \sum_{d \in \mathcal{D}} \sum_{k=0}^{K-1} w_{o,d,k} z_{p,d,t,k}. \qquad \text{(DyadicHawkes)}$$

The total influence of drug $d \in \mathcal{D}$ on outcome $o \in \mathcal{O}$ can be measured by $\sum_{k=0}^{K-1} w_{o,d,k}$.

Note that the weights are independent of the patient p and the times t, so that within a collection of EHRs, we have a large number of training samples that can be used to infer the weights. Also note that the sufficient statistics of the data, $z_{p,d,t,k}$, are simple functions of the data and independent of outcome o. Hence these statistics can be pre-computed once and used for all outcomes of interest.

Note that (DyadicHawkes) is a generalization of a log-linear Poisson autoregressive processes (Zhu and Wang, 2011), for which Hall et al. (2016) have recently derived sample complexity bounds.

2.3 Comparing the two models

Contrasting the classical model in (MSCCS) and our proposed Hawkes model with dyadic influence functions in (DyadicHawkes), we see that both model the log of the event rate $\lambda_{p,o,t}$ as a linear combination of sufficient statistics of the past data (either the $\tilde{x}$'s or the z's, respectively). Despite this superficial similarity, the models exhibit very different behaviors. In particular, the $\tilde{x}$'s in (MSCCS) can be thought of as the collection of observed events plus *artificial, simulated* events injected into the model. In particular, we can think of every day when a patient is taking a drug but the drug is not noted that day in the EHR as *missing data*. The MSCCS approach essentially imputes values for the missing data by assuming all people are taking all drugs for the same amount of time. Clearly this imputation is inaccurate, and these inaccuracies can bias inference of which drugs are causing which ADRs.

In contrast, the Hawkes model in (DyadicHawkes) does not require us to explicitly impute missing data. The idea is that different drugs may have different impacts after different delays after the initial prescription, and different potential delays are captured by the different ϕ_ks. In effect, when we learn the parameters $\{w_{o,d,k}\}_{\substack{o\in\mathcal{O},d\in\mathcal{D}\\ k\in\{0,\dots,K-1\}}}$, we are learning the strength of the impact of drug d when the time since it was prescribed is on the order of 2^k. Thus this model is more flexible than the MSCCS model.

Note that (MSCCS) is similar (but not equivalent to) (DyadicHawkes) for $K = 1$ if the same value of L is used. In particular, if a patient was prescribed a drug multiple times in the past L hours, then MSCCS would treat this a single drug occurrence in the time-at-risk window. In contrast, the Hawkes model suggests the multiple prescriptions have a cumulative effect. Since the number of prescriptions within a time-at-risk window L is generally small, these models can have similar empirical performances for $K = 1$.

Note that the number of weights to be inferred in (MSCCS) is equal to the number of drugs being evaluated. The number of weights to be inferred in our Hawkes model (DyadicHawkes) is equal to the product of the number of drugs, $|\mathcal{D}|$, and K, the number of different influence functions in the model. Thus while using the Hawkes process with multiple influence functions can reduce bias in estimating ADRs, (DyadicHawkes) has a larger (by a factor of K) parameter space than (MSCCS). We adjust for this larger parameter space in our inference method by using sparsity regularization, as described below.

3. Inference approach

Let $\boldsymbol{b} \coloneqq (b_{p,o})_{p\in\{1,\dots,P\},\, o\in\mathcal{O}}$ and $\boldsymbol{w} \coloneqq (w_{o,d,k})_{o\in\mathcal{O},d\in\mathcal{D},\, k\in\{0,\dots,K-1\}}$ denote the model parameters. (The model parameters for MSCCS can be represented this way with $K = 1$.) Using the Poisson likelihood in (1), we have that the negative log-likelihood of patient p's occurrences of outcome o is proportional to

$$\ell_{p,o,t}(b_{p,o}, \boldsymbol{w}) \coloneqq \lambda_{p,o,t} - x_{p,o,t} \log \lambda_{p,o,t} \tag{3}$$

Define

$$\ell(\boldsymbol{b}, \boldsymbol{w}) = \frac{1}{P} \sum_{p=1}^{P} \sum_{o\in\mathcal{O}} \sum_{t} \ell_{p,o,t}(b_{p,o}, \boldsymbol{w}).$$

Note that in our Hawkes model (DyadicHawkes), $\ell_{p,o,t}$ is piecewise constant over t, so the log-likelihood can be efficiently computed via data squashing (Madigan et al., 2002; Simpson et al., 2013). In order to avoid overfitting and obtain an interpretable result, we induce sparsity by adding an ℓ_1 (LASSO) penalty (Tibshirani, 1996) on $\boldsymbol{w}$, resulting in the following optimization problem:

$$\min_{\boldsymbol{b},\boldsymbol{w}} \ell(\boldsymbol{b}, \boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1, \tag{4}$$

where $\lambda > 0$ is a tuning parameter controlling the level of sparsity.

The objective function in (4) is convex and can be minimized using a variety of approaches (*cf.* Wright et al. (2009)). Empirically we find that alternating between minimizing $\boldsymbol{b}$ (which has a closed form solution) and updating $\boldsymbol{w}$ using FISTA (Beck and Teboulle, 2009) yields fast convergence and quickly computable updates.

4. Experiments

4.1 OMOP task

To evaluate methods for ADR discovery, OMOP established a challenge problem of ranking drug-outcome pairs as possible ADRs. From ten different drugs and ten different outcomes, 53 drug-outcome pairs are labeled by OMOP as ground-truth true or false ADRs based on information on drug labels, for example calling warfarin-bleeding and ACE inhibitor-angioedema true pairs while calling ACE inhibitor-bleeding a false pair. From this ground truth, any algorithm that can rank drug-condition pairs from most- to least-likely ADRs can be evaluated via an ROC curve. This task is extremely difficult because many ADRs are (thankfully) rare, in addition to all the ordinary challenges of causal discovery from LODs, such as confounding by other measured or unmeasured variables, which may also vary over time.

4.2 Data description

We employ a de-identified version of Marshfield Clinic health system's EHR, which has been used for clinical care since the mid 1980s, serving primary, secondary, and tertiary care clinicians throughout Central and Northern Wisconsin (Powell et al., 2012). The system uses a variety of data gathering techniques to capture and code patient encounter information including diagnoses, laboratory results, procedures, medications, and vital sign measurements such as height, weight, blood pressures, etc. This longitudinal data is linked for each patient and exists in electronic form back to the early 1960s. Data consist of date-stamped events such as diagnosis codes and drug prescriptions; dates are encoded as patient age in 1/1000 years, for privacy reasons.

We extract ten drug prescription records and ten diagnosis records from the de-identified EHRs according to the definitions of the vocabularies used in the OMOP ground truth. We admit a patient into the cohort if the length of the observation for the patient is at least three months. The resulting cohort contains 327,824 patients with 1,940,681 adverse health outcome occurrences and 11,211,769 drug prescription records. The average observation duration for the patients in our cohort is 9.1 years. Following the design of MSCCS, we

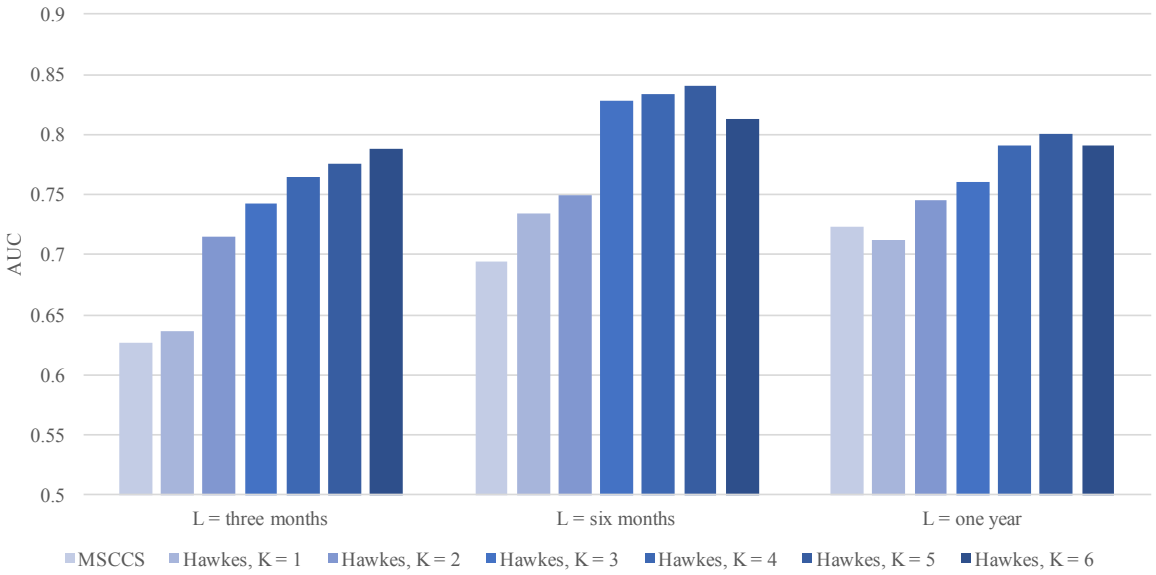

Figure 2: Area under the curve for MSCCS and the Hawkes with various L and K

restrict our attention to patients with at least one occurrence of the outcome o when we are inferring the weight for that outcome.

4.3 Metrics

Since the log-likelihood for both models is separable across different health outcomes, the influence is not directly comparable among different outcomes. We define the normalized score $S_{o,d}$ for each drug-outcome pair in MSCCS and the Hawkes process model as following:

$$S_{o,d} = \frac{w_{o,d}}{\sqrt{\sum_{d \in \mathcal{D}} w_{o,d}^2}} \text{ in MSCCS}, \quad S_{o,d} = \frac{\sum_{k=0}^{K-1} w_{o,d,k}}{\sqrt{\sum_{d \in \mathcal{D}} \left(\sum_{k=0}^{K-1} w_{o,d,k} \right)^2}} \text{ in Hawkes.}$$

For quantitative metrics, we report the area under the curve (AUC) of receiver operating characteristics (ROC) using the OMOP ground truth and the scores defined above.

4.4 Evaluation

To choose the shrinkage parameter λ for both MSCCS and the Hawkes process model, we perform leave-one-condition-out cross-validation (LOCOCV): for each of the ten outcomes, we adaptively pick $\lambda \in \{0, 10^{-8}, 10^{-7}, 10^{-6}\}$ that perform the best on the other nine conditions.

Figure 2 presents the AUC of MSCCS and Hawkes with various K for $L \in \{$three months, six months, one year$\}$. Note that for fixed L, both MSCCS and Hawkes make use of the information in the past L hours to model the occurrence of adverse health outcomes at each time. We observe that the Hawkes process model consistently outperforms MSCCS when more then one influence function is used, effectively indicating that modeling drug-

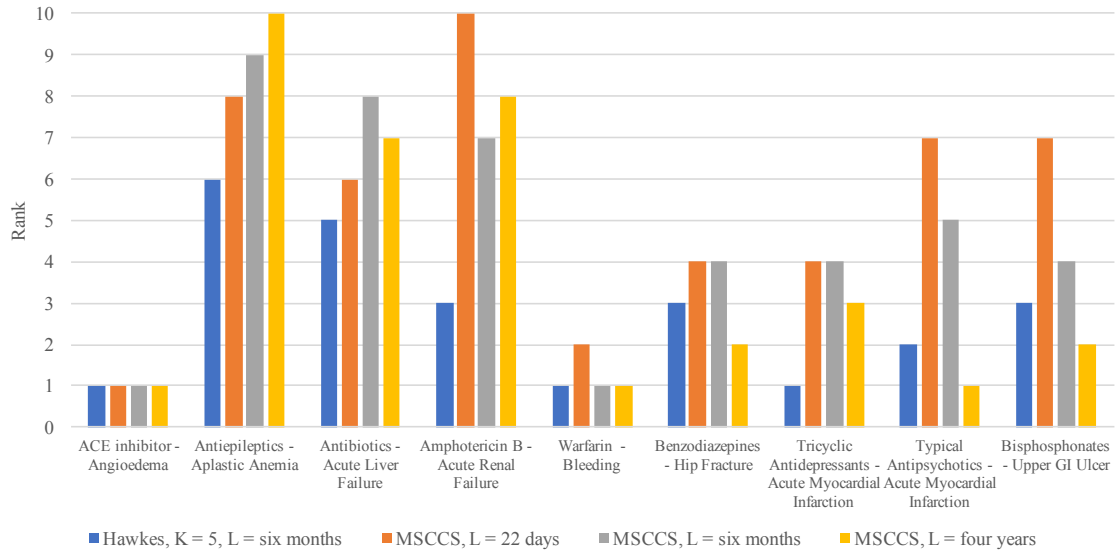

Figure 3: Rank of true ADR-causing drug among all ten drugs for each true ADR pair

dependent time-at-risk windows (captured in a data-dependent manner by the Hawkes model) is beneficial to ADR discovery.

In the literature of ADR discovery from LODs (Norén et al., 2013b; Simpson et al., 2013; Suchard et al., 2013; Schuemie et al., 2016), different methods are compared under their best settings. To test the highest AUC for both models, we vary L from 22 days to ten years and K from 1 to 7. The best performers of MSCCS reaches an AUC of 0.7449 at $L =$ four years, while the Hawkes process model reaches its best AUC of 0.8409 at $K = 5$, $L =$ six months. To demonstrate how well the Hawkes process model and MSCCS can predict unseen adverse drug reactions in practice, we perform LOCOCV to adaptively and jointly pick λ, L and K. The AUC after LOCOCV for MSCCS is 0.6970, while the AUC after LOCOCV for Hawkes is 0.8258, indicating that the expressive power of the the Hawkes process model better coincides with the ADR signals encoded in the data.

Figure 3 shows the rank of the true ADR-causing drug among all ten drugs for each of the nine true ADR pairs. Rank of one means the true ADR-causing drug is assigned the highest score among all ten drugs by the method. Notice that the eighth and ninth pairs are both associated with the same outcome, so ranking one true causing drug to the first place and the other true causing drug to the second place is the best one can do. We observe that although MSCCS with $L =$ four years performs reasonably well on the first pair and the last five pairs, it completely fails to discover the second true ADR pair. Actually, it even assigns a negative score to this true ADR pair, suggesting that the true causing drug inhibits the occurrence of the adverse outcome. On the other hand, MSCCS with $L = 22$ days attains better performance on the second and third pairs, but it cannot successfully learn the eighth and ninth true ADR pair due to the limitation of a short time-at-risk window. By using different influence functions, the Hawkes process model is able to better capture these long-term and short-term effects jointly and this results in an overall performance superior to MSCCS.

5. Discussion

We have proposed a log-linear Hawkes process model of adverse drug reactions with longitudinal observational data. Compared with the leading approach, multiple self-controlled case series, for ADR discovery with LODs, the proposed method offers tremendous flexibility in modeling time-varying effects of various drugs on the occurrence of adverse health outcomes. Experimental results demonstrate the superior performance of the proposed method over MSCCS in various experiment settings.

Notice that in our experiments, the increase of the time-at-risk window and the number of influence functions used in the models does not necessarily correspond to the improvement of ADR discovery performance. A reasonable explanation is that with prolonged time-at-risk windows, long-term fluctuation of the baseline occurrence rate of an adverse health outcome also needs to be taken into consideration. However, in the current modeling framework, for efficiency we only use a patient-specific yet time-invariant parameterization to model the baseline occurrence rate of an adverse health outcome. Therefore, incorporating time-varying baseline (Kuang et al., 2016b) to distinguish between baseline fluctuation and time-varying drug effects would be an important future research direction. Other future directions include improving the efficiency of model fitting via parallelism and stochasticity as well as designing different kernels to facilitate the incorporation of different clinical hypotheses.

Acknowledgments

The project described was supported by the NIH BD2K Initiative grant U54 AI117924, the NIGMS grant 2RO1 GM097618, and the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427, and by the NSF grant CCF-1418976. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF. The authors would like to thank Mark Craven, Zhanrong Du, and Sinong Geng for helpful discussions.

References

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Valérie Chavez-Demoulin and JA McGill. High-frequency financial data modeling using hawkes processes. *Journal of Banking & Finance*, 36(12):3415–3426, 2012.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes, Vol. I: Probability and its Applications*. Springer-Verlag, New York, second edition, 2003.

Şeyda Ertekin, Cynthia Rudin, Tyler H McCormick, et al. Reactive point processes: A new approach to predicting power failures in underground electrical systems. *The Annals of Applied Statistics*, 9(1):122–144, 2015.

Steven Findlay. Health policy briefs: The fda's sentinel initiative. *Health Affiaris*, 2015.

Eric C Hall and Rebecca M Willett. Dynamical models and tracking regret in online convex programming. In *Proc. International Conference on Machine Learning (ICML)*, 2013. arXiv.org:1301.1254.

Eric C Hall, Garvesh Raskutti, and Rebecca Willett. Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*, 2016.

Alan G Hawkes. Point spectra of some self-exciting and mutually-exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:83–90, 1971a.

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

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216:574, 2015.

Zhaobin Kuang, James Thomson, Michael Caldwell, Peggy Peissig, Ron Stewart, and David Page. Computational drug repositioning using continuous self-controlled case series. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 491–500, New York, NY, USA, 2016a. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939715. URL `http://doi.acm.org/10.1145/2939672.2939715`.

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

Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *ICML*, pages 1413–1421, 2014. arXiv:1402.0914.

David Madigan, Nandini Raghavan, William Dumouchel, Martha Nason, Christian Posse, and Greg Ridgeway. Likelihood-based data squashing: A modeling approach to instance construction. *Data Mining and Knowledge Discovery*, 6(2):173–190, 2002.

G Niklas Norén, Tomas Bergvall, Patrick B Ryan, Kristina Juhlin, Martijn J Schuemie, and David Madigan. Empirical performance of the case–control method: lessons for developing a risk identification and analysis system. *Drug Safety*, 36(1):73–82, 2013a.

G Niklas Norén, Tomas Bergvall, Patrick B Ryan, Kristina Juhlin, Martijn J Schuemie, and David Madigan. 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):107–121, 2013b.

Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454:995–999, 2008.

Valerie Powell, Franklin M Din, Amit Acharya, and Miguel Humberto Torres-Urquidy. *Integration of medical and dental care and patient data*, volume 3. Springer Science & Business Media, 2012.

Stephanie J Reisinger, Patrick B Ryan, Donald J O'hara, Gregory E Powell, Jeffery L Painter, Edward N Pattishall, and Jonathan A Morris. 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):652–662, 2010.

Melissa A Robb, Judith A Racoosin, Rachel E Sherman, Thomas P Gross, Robert Ball, Marsha E Reichman, Karen Midthun, and Janet Woodcock. The us food and drug administration's sentinel initiative: expanding the horizons of medical product safety. *Pharmacoepidemiology and Drug Safety*, 21(S1):9–11, 2012.

Patrick B Ryan, David Madigan, Paul E Stang, J Marc Overhage, Judith A Racoosin, and Abraham G Hartzema. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the observational medical outcomes partnership. *Statistics in Medicine*, 31(30):4401–4415, 2012.

Patrick B Ryan, Martijn J Schuemie, Susan Gruber, Ivan Zorych, and David Madigan. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. *Drug Safety*, 36(1):59–72, 2013a.

Patrick B Ryan, Martijn J Schuemie, and David Madigan. Empirical performance of a self-controlled cohort method: lessons for developing a risk identification and analysis system. *Drug Safety*, 36(1):95–106, 2013b.

Martijn J Schuemie, David Madigan, and Patrick B Ryan. Empirical performance of lgps and leopard: lessons for developing a risk identification and analysis system. *Drug Safety*, 36(1):133–142, 2013.

Martijn J Schuemie, Gianluca Trifiro, Preciosa M Coloma, Patrick B Ryan, and David Madigan. Detecting adverse drug reactions following long-term exposure in longitudinal observational data: The exposure-adjusted self-controlled case series. *Statistical methods in medical research*, 25(6):2577–2592, 2016.

Shawn E Simpson, David Madigan, Ivan Zorych, Martijn J Schuemie, Patrick B Ryan, and Marc A Suchard. Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics*, 69(4):893–902, 2013.

Marc A Suchard, Ivan Zorych, Shawn E Simpson, Martijn J Schuemie, Patrick B Ryan, and David Madigan. Empirical performance of the self-controlled case series design: lessons for developing a risk identification and analysis system. *Drug Safety*, 36(1):83–93, 2013.

Janet Sultana, Paola Cutroneo, Gianluca Trifirò, et al. Clinical and economic burden of adverse drug reactions. *Journal of Pharmacology and Pharmacotherapeutics*, 4(5):73, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

F. Zhu and D. Wang. Estimation and testing for a poisson autoregressive model. *Metrika*, 73(2):211–230, 2011.