Modelling Dynamic Individual Outpatient Claims using Hawkes Process*

Yuhao Li LIYUHAO.ECON@WHU.EDU.CN
Economics and Management School
Wuhan University

Draft Time: October, 2022

Abstract

Outpatient activities are often clustered in time and are state dependent. This paper proposes a Hawkes process model to study dynamic individual outpatient claim data. Specifically, I construct Hawkes processes that consist of outpatient activities for individuals who hold different health insurance plans. Introducing individual heterogeneities into the model complicates the estimation procedure, as an underlying process could be non-stationary. I re-introduce a minimum distance estimator with simplified proofs to estimate the model. Finite sample properties are investigated in simulations. The approach is applied to studying the RAND Health Insurance Experiment data. The empirical results may shed light on the dynamic mechanism of the outpatient claim data.

JEL Classification: C41, C51, I12, I13

Keywords: Hawkes Process, Claim Frequency, Health Insurance, State Dependence

1. Introduction

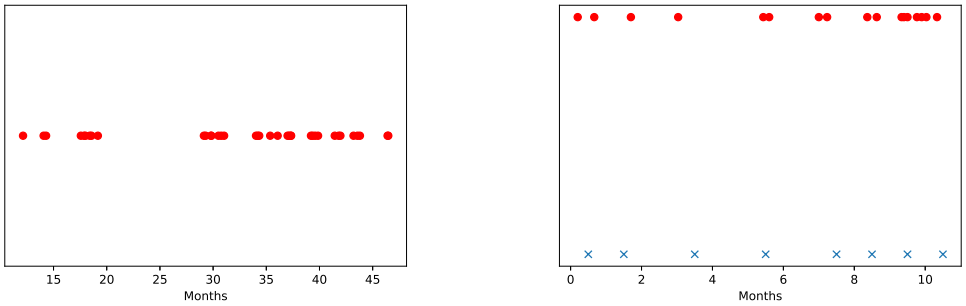
2. Introduction

In this paper, I present a Hawkes process framework to model and analyze individual-level outpatient claim data in health insurance. The Hawkes process is a counting process with the self-exciting property, i.e., past claims will affect the occurrence of future claims. This new framework can (1) model cluster pattern in the claim data, (2) study marginal effect at a representative value, and (3) allow a researcher to specify and estimate an individual-specific and history-dependent shadow coinsurance rate.

Before going into detail, I display two figures that record detailed claim times for two individuals under different cost-sharing insurance plans. Details about the data can be

∗. I am grateful to Miguel A. Delgado for supports and guidance throughout this project, and to Winfried Stute for his inspiration and valuable comments. I also thank conference and seminar participants at the EEA-ESEM Lisbon, IAAE Montreal, UC3M, Liaoning University and SUFE. All errors are my own.

found in Section 4. Careful analysis of these figures is useful and necessary, as it reveals some important aspects of the nature of the claim data:



(a) Individual 1, free plan

(b) Individual 2, 95% coinsurance rate plan

Figure 1: Claim times

- Claims are clustered in time, as indicated by red dots in Figure 1. The cluster pattern is not unique to those individuals displayed here, but ubiquitous in the data set.
- Clustered occurrence times imply the data is non-stationary. Specifically, durations between claims are not stationary.
- The source of such non-stationarity might be state dependence, i.e., previous claims would have effects on future ones. Put it another way, the dynamics among claims could be the reason for such a cluster pattern.
- Furthermore, the dynamic mechanism involved could be complicated. For example, the medical follow-up check is one dynamic channel. In a cost-sharing insurance plan, the moral hazard could be another dynamic channel: Individuals might have an incentive to consume more medical services because the cumulative cost is closer to the out-of-pocket cap after each outpatient consumption.

Similar patterns can also be found in automobile insurance data, see Pinquet (2000); Seal (1969). Understanding and modelling dynamic properties of the claim data is important. For an actuarial audience, being able to predict claims as accurately as possible is an essential task. Actuaries require accurate predictions for pricing, and for estimating future company liabilities. For economists, dynamic insurance data could help identify moral hazard, see Abbring et al. (2003a,b). Yet, existing actuarial literature often ignore the dynamic nature of the data when studying the claim frequency. In the economic literature, structure discrete choice models are used to study medical claim frequency. The key assumption made in those literatures is the Markov dynamic mechanism between past claims and future ones. But this assumption contradicts the non-stationary observation mentioned

before. To reconcile this conflict, a researcher might aggregate claims (for example, aggregate daily claims into monthly dichotomous data, with one indicating the existence of claims and zero indicating the absence of claims) to form a stationary series. The blue 'x' in subfigure (b) of Figure 1 illustrates such a method. However, in this way, dynamic mechanisms are inevitably altered, and consequently, results might not be plausible. In Section 6, I will discuss some drawbacks of the mentioned methods in detail.

This paper advocates the Hawkes process (Hawkes, 1971) as an alternative approach to modelling dynamic individual-level insurance data. The Hawkes process is a self-exciting process, i.e., the intensity rate of an occurrence is conditional on a sigma-algebra generated by the process up to current time. This sigma-algebra can also include time-invariant elements (e.g., individual covariates) and even external shocks. The self-exciting structure is ideal for modelling dynamic mechanisms. The Hawkes process is also a branching process (Brémaud and Massoulié, 2001). Its branching mechanism might be summarized as follows. First, a Poisson process would independently generate 'immigrants'. Second, a given immigrant event can give birth to subsequent offspring events, and an offspring event might also be the ancestor of future generation offspring events. Thus, the branching structure gives rise to a cluster interpretation.

Existing applications of the Hawkes process can be found in finance, see Bacry et al. (2015); Bowsher (2007); Chavez-Demoulin et al. (2005), in seismology, see Zhuang et al. (2002) and in criminology, see Mohler et al. (2012). Particularly, in the field of insurance, the Hawkes process is used to model the risk process (e.g., the Cramer-Lundberg process). To the best of my knowledge, Stabile and Torrisi (2010) is the first work to consider a risk model with Hawkes claims arrivals, Cheng and Seol (2020); Dassios and Zhao (2012); Jang and Dassios (2013); Zhu (2013) extend the model in various directions. Swishchuk et al. (2021) use the Hawkes process to model arrivals of legal expenses insurance claims.

In these applications, researchers usually specify a Hawkes process to model one observation process. For example, in finance, it is bid or ask times for one stock; in seismology, it is earthquake occurrence times in a region; in criminology, it is crime events in one area; and in insurance, it is the ruin of one insurance company. This study differs from the existing Hawkes process literature in the sense that instead of one observation, I have multiple observations. In the dataset, n individuals are observed, and for each individual, his or her outpatient claim times consist of an observation process. I use a minimum distance method instead of the conventional maximum likelihood method to estimate parameters.

The paper is organized as follows. Section 2 introduces necessary concepts about the Hawkes process. In section 3, I outline the estimation method. In section 4, I demonstrate how to model dynamic outpatient claim data under different cost-sharing plans using Hawkes processes. Section 5 provides interpretations and results of this empirical study. In a divergence from typical formatting, in section 6, I will discuss some drawbacks of existing

methods with a purpose of highlighting advantages of the proposed approach. I also discuss some limitations of the new framework. Section 7 concludes.

3. The Hawkes Process

Suppose, for an individual, we are given a collection of increasing random points $T_1 < T_2 < \cdots$ observed over time. Let $N(t)$ for $t \in \mathcal{T} = (0, T]$ denote the number of T_j that fall below t:

$$N(t) = \sum_{j=1}^{\infty} \mathbb{I}\{T_j \leq t\}$$

where $\mathbb{I}\{A\}$ is an indicator with the value equal to one if an event A occurred, and zero else wise.

In this study, $N(t)$ is a counting process that records this individual's outpatient times. Define the (conditional) intensity $\lambda(t)$ as:

$$\lambda(t) = \lim_{h \to 0} \frac{Pr\{N(t, t+h] > 0 \mid \mathcal{F}(t-)\}}{h}$$

$\lambda(t)$ measures the instantaneous conditional probability of the occurrence of an event given all the history information, and it uniquely characterizes the probability structure of the corresponding counting process, see Proposition 7.2.IV of Daley and Vere-Jones (2007). In this definition, $\mathcal{F}(t-)$ is a filtration that contains information up to a time just before t. The choice of the filtration $\mathcal{F}(t-)$ is important in the context of the counting process analysis. In case of $\mathcal{F}(t-) = \emptyset$ and $\lambda(t) = \lambda, \forall t \in \mathcal{T}$, we end up with a Poisson process with rate λ. While in this study, I am interested in the filtration that includes a sigma-algebra generated by the process itself, i.e.,

$$\sigma(N(s) : s \leq t) \subseteq \mathcal{F}(t)$$

Counting processes with this self-generated sigma-algebra is called the self-exciting process. The Hawkes process is one well known self-exciting process, and its conditional intensity has the following specification:

$$\lambda(t) = \lambda_0 + \int_0^t g(t-s) dN(s) \tag{1}$$

$$= \lambda_0 + \sum_{j: t_j < t} g(t - t_j) \tag{2}$$

where λ_0 is a time-invariant parameter, and $g : \mathbb{R} \to \mathbb{R}^+$ is called the memory kernel. One popular kernel specification is the exponential function (Embrechts et al., 2011; Hawkes, 1971):

$$g(t) = \alpha \exp(-\mu t), \quad \alpha, \mu > 0$$

The construction of this process involves a *branching* mechanism, which naturally leads to a cluster interpretation:

- A Poisson process N^0 with rate λ_0 independently generates *immigrants* (i.e., independent events) $s_j \in \mathcal{T}$.
- Each immigrant might generate a cluster $C_j = C_{s_j}$, which is a random set formed by the *offspring* of s_j. I use Figure 2 to help illustrate this concept. In this particular realization of outpatient activities, there are three clusters $\{T_1, \ldots, T_6\}, \{T_7, T_8\}$ and $\{T_9\}$. $\{T_1, T_7, T_9\}$ are immigrant events (Gen_0) in the first, second and third cluster, respectively. Within each cluster, there might be more than one generation of offspring events.
- Given the immigrants, the centered clusters

$$C_j - s_j = \{\tau - s_j : \tau \in C_j\}, s_j \in N^0$$

  are i.i.d, and independent of N^0
- The Hawkes process $N(t), t \in \mathcal{T}$ consists of the union of all clusters.

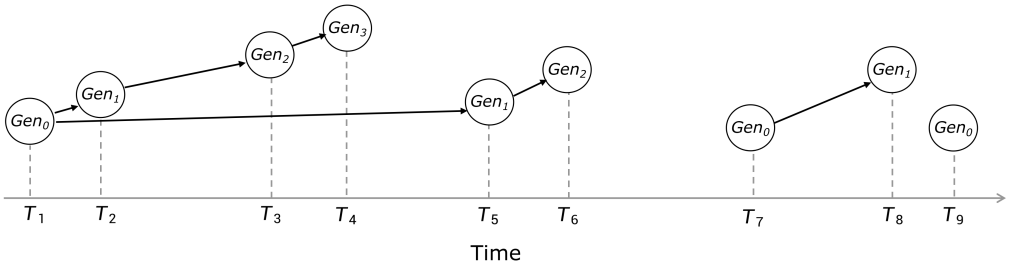


Figure 2: A possible cluster realization

Furthermore, suppose the Hawkes process is stationary, then the corresponding intensity would eventually reach a steady value λ^*:

$$\begin{aligned}\lambda^* &= \lambda_0 + \int_0^\infty g(t)\Lambda^*(dt) \\ &= \lambda_0 + \lambda^* \int_0^\infty g(t)dt\end{aligned}$$

and hence,

$$\lambda^* = \frac{\lambda_0}{1 - \int_0^\infty g(t)dt} \tag{3}$$

The term $n^* = \int_0^\infty g(t)dt$ is called the branching ratio. A Hawkes process is stationary if $0 < n^* < 1$. In the case of an exponential kernel, the stationary condition is:

$$n^* = \int_0^\infty \alpha \exp(-\mu t)dt = \frac{\alpha}{\mu} \in (0, 1)$$

The branching ratio n^* can be interpreted as the average number of offspring per event, and thus is an indicator of the cluster size. To see it, suppose the sampling size is normalized to one, then the term $\lambda^* dt$ is the proportion of sampled events. Among them, there are $\lambda_0 dt$ parent events generated by the Poisson process with rate λ_0; thus, there are $\lambda_0 dt$ families (clusters). By Equation 3, the expected size per family is $N_\infty = 1/(1 - n^*)$. Let A_l be the expected number of events in $Generation_l$, and $A_0 = 1$ (the parent immigrant). Then expected size of a cluster N_∞ can also be defined as:

$$N_\infty = \sum_{l \geq 1} A_l \tag{4}$$

Suppose the average number of offspring per event is $\tilde{n}$, then there is an inductive relationship $A_l = A_{l-1}\tilde{n}$. With $A_0 = 1$, deriving:

$$A_l = A_0 \left(\tilde{n}\right)^l = \left(\tilde{n}\right)^l, l \geq 1 \tag{5}$$

$$N_\infty = \frac{1}{1 - \tilde{n}} \tag{6}$$

Thus, $n^* = \tilde{n}$ measures the endogeneity degree. The case $n^* < 1$ implies that N_∞ is bounded and further implies that a cluster would eventually die out almost surely.

Lastly, the expected number of occurred events can be obtained if the Hawkes process is stationary:

$$\mathbb{E}N((0, t]) = \frac{\lambda_0 t}{1 - n^*}$$

4. Minimum Distance Estimation of the Hawkes Process

Conventionally, researchers would use maximum likelihood methods to estimate the intensity function $\lambda(t)$ of a stationary Hawkes process $N(t), t \in (0, T]$. The associated log-likelihood function is well known:

$$\log L(t_1, t_2, \ldots, t_{N(T)}) = \sum_{j=1}^{N(T)} \log \lambda(t_j) - \int_0^T \lambda(t)dt$$

where $t_1, t_2, \ldots, t_{N(T)} \in (0, T]$ are observed events, see Ogata (1998) for details.

When having multiple observations $N_i(t), i \in \{1, 2, \ldots, n\}, t \in (0, T]$, one must consider the individual heterogeneity. Considering a Hawkes process with exponential kernel, when having individual heterogeneities, the intensity specification for an individual i might be:

$$\lambda_i(t) = \lambda_0 + \int_0^t \phi(z_i) \exp(-\mu(t - t_{ij}))$$

where z_i is a vector of individual heterogeneities. The associated branching ratio is $n_i^* = \phi(z_i)/\mu$, which may or may not be in the stationary region $(0, 1)$. When $n_i^* > 1$ (corresponds to a non-stationary Hawkes process), the maximum likelihood method is invalid. Even some stationary mechanisms exist, pooling n-observation processes into one process and then using maximum likelihood might also be implausible. This is because the likelihood function given occurrence times is valid only if the underlying process is simple, i.e., points are almost everywhere pairwise distinct. It is not uncommon that two insurees would visit a doctor on the same day, and hence the pooled process would not be simple.

4.1 Estimation Theories

Kopperschmidt and Stute (2013) propose a minimum distance method to estimate self-exciting processes under multiple observations. Their method relies on the Doob-Meyer decomposition result:

$$N_i(t) = \Lambda_i(t) + M_i(t)$$

where $\Lambda_i(t) = \int_0^t \lambda_i(s)ds$ is the cumulative intensity function, also known as the compensator, and $M_i(t)$ is a martingale with zero mean: $\mathbb{E}M_i(t|\mathcal{F}_i(t-)) = 0$. Their proofs of asymptotic properties are based on U-statistic arguments. In this section, I provide an alternative proving strategy under the framework of the *generalized weighted Cramer-von Mises distance estimator* introduced by ÖZTÜRK and Hettmansperger (1997).

Denote $\theta \in \Theta \subset \mathbb{R}^k$ as parameters of interest, $\mathcal{T} = (0, T]$ as the time space, $\mathcal{M} = \{\Lambda(t; \theta) : t \in \mathcal{T}, \theta \in \Theta\}$ as the associated model for the cumulative intensity, and let

$$M(t; \theta) = \mathbb{E}\left(N_1(t) - \Lambda_1(t; \theta|\mathcal{F}_1(t-))\right)$$

be the moment restriction. By Doob-Meyer decomposition result, one has

$$M(t; \theta_0) = 0 \quad \forall t \in \mathcal{T}$$

where θ_0 is a vector of true parameters. I impose the following assumptions:

- A1. For each $\varepsilon > 0$.

$$\inf_{||\theta - \theta_0|| \geq \varepsilon} ||\mathbb{E}\Lambda(\cdot, \theta_0) - \mathbb{E}\Lambda(\cdot, \theta)||_{\mathbb{E}\Lambda(\cdot, \theta_0)} > 0$$

  where

$$||f||_\mu = \left[\int_{\mathcal{T}} f^2(t)\mu(dt)\right]^{1/2}$$

  is a semi-norm.

- A2. The process $(t, \theta) \to \Lambda(t, \theta)$ is continuous with probability one.
- A3. $\Lambda(t; \theta)$ is bounded in t and θ.
- A4. $\Theta \subset \mathbb{R}^k$ is compact.

Assumption A1 is a weak identification condition. A2 suggests that $\Lambda(\cdot, \theta)$ has a (random) Lebesgue intensity $\lambda(\cdot, \theta)$ with values in an appropriate Skorokhod Space. This guarantees continuity (but not differentiability) of $\Lambda(t, \theta)$ in t and allows for unexpected jumps in the intensity function. A3 is used in ÖZTÜRK and Hettmansperger (1997), and A4 is standard in the literature.

By Assumption A1, one has

$$P\left(M(t; \theta) = 0\right) < 1, \quad \theta \neq \theta_0$$

thus, $M(t; \theta) \neq 0$ in a non-null space of $\mathcal{T}$, and one has

$$\int_{\mathcal{T}} M(t; \theta_0)^2 \mathbb{E}\Lambda(dt; \theta_0) = 0$$

but

$$\int_{\mathcal{T}} M(t; \theta)^2 \mathbb{E}\Lambda(dt; \theta_0) \neq 0 \quad \forall \theta \neq \theta_0$$

Hence,

$$\theta_0 = \arg\min_{\theta \in \Theta} \int_{\mathcal{T}} M(t; \theta)^2 \mathbb{E}\Lambda(dt; \theta_0)$$

By Lemma 3 of Kopperschmidt and Stute (2013), the above equation can be re-written as:

$$\theta_0 = \arg\min_{\theta \in \Theta} ||\mathbb{E}\Lambda(\cdot, \theta_0) - \mathbb{E}\Lambda(\cdot, \theta)||^2_{\mathbb{E}\Lambda(\cdot, \theta_0)}$$

By Lemma 5 of the same paper, one has

$$||\bar{N}_n - \bar{\Lambda}_n(\cdot, \theta)||^2_{\bar{N}_n} \xrightarrow{p} ||\mathbb{E}\Lambda(\cdot, \theta_0) - \mathbb{E}\Lambda(\cdot, \theta)||_{\mathbb{E}\Lambda(\cdot, \theta_0)}$$

where

$$\bar{N}_n = \frac{1}{n} \sum_{i=1}^{n} N_i, \quad \bar{\Lambda}_n(\cdot, \theta) = \frac{1}{n} \sum_{i=1}^{n} \Lambda_i(\cdot, \theta)$$

We can write the minimum distance estimator as

$$\begin{aligned} \hat{\theta}_n &= \arg\min_{\theta \in \Theta} ||\bar{N}_n - \bar{\Lambda}_n(\cdot, \theta)||^2_{\bar{N}_n} \\ &= \arg\min_{\theta \in \Theta} \int_{\mathcal{T}} \bar{M}_n(t; \theta)^2 \bar{N}_n(dt) \end{aligned}$$

where $\bar{M}_n(t; \theta) = \bar{N}_n(t) - \bar{\Lambda}_n(t, \theta)$, and $M_i(t, \theta) = N_i(t) - \Lambda_i(t; \theta)$. The quantity $||\bar{N}_n - \bar{\Lambda}_n(\cdot, \theta)||^2_{\bar{N}_n}$ represents an overall measure of fit of $\bar{\Lambda}_n(\cdot, \theta)$ to $\bar{N}_n$. This objective function is a weighted Cramér-von Mises statistic, which can be interpreted as a minimum distance estimator.

Theorem 1 *Under Assumptions A1-A4, one has*

$$\hat{\theta}_n \xrightarrow{a.s} \theta_0$$

Proof See Appendix A ■

In order to obtain the asymptotic normality, additional assumptions are required.

- A5. $\Lambda(t,\cdot)$ is once differentiable in a neighborhood of θ_0 and satisfies $\dot{\Lambda}(t,\theta)$ is square integrable w.r.t $\mathbb{E}\Lambda(\cdot,\theta_0)$ where $\mathcal{N}_0$ is a neighborhood of θ_0 and $\dot{\Lambda}(t,\theta) = \partial\Lambda(t;\theta)/\partial\theta$.
- A6. $\theta_0 \in int(\Theta)$.

Assumption A5 is a standard smoothness condition. A5 is unchanged if one replaces $\Lambda(t,\cdot)$ by $M(t;\cdot)$, and $\dot{\Lambda}(t,\theta)$ by $\dot{M}(t;\theta) = \mathbb{E}\dot{M}_1(t;\theta) = \partial M(t;\theta)/\partial\theta$. Assumption A6 is standard.

Theorem 2 *Under Assumptions A1-A6, one has*

$$\sqrt{n}\left(\int_{\mathcal{T}} \dot{M}(t;\theta_0)\dot{M}(t;\theta_0)^\top \mathbb{E}\Lambda(dt;\theta_0)\right)(\hat{\theta}_n - \theta_0) \xrightarrow{d} \int_{\mathcal{T}} \dot{M}(t;\theta_0) B_\Gamma \mathbb{E}\Lambda(dt;\theta_0)$$

where B_Γ denotes a centered Gaussian process with covariance structure given by $\Gamma(t,s) = \mathbb{E}(M_1(t;\theta_0)M_1(s;\theta_0))$.

Proof See Appendix A. ■

This theorem naturally leads to the following corollary.

Corollary 3 *Under Assumptions A1-A6, one has*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0,\Omega)$$

where

$$\begin{aligned}\Omega = &\left(\int_{\mathcal{T}} \dot{M}(t;\theta_0)\dot{M}(t;\theta_0)^\top \mathbb{E}\Lambda(dt;\theta_0)\right)^{-1} \times \\ &\int_{\mathcal{T}}\int_{\mathcal{T}} \dot{M}(t;\theta_0)\dot{M}(s;\theta_0)^\top \Gamma(t,s)\mathbb{E}\Lambda(dt;\theta_0)\mathbb{E}\Lambda(ds;\theta_0)\times \\ &\left(\int_{\mathcal{T}} \dot{M}(t;\theta_0)\dot{M}(t;\theta_0)^\top \mathbb{E}\Lambda(dt;\theta_0)\right)^{-1}\end{aligned}$$

Proof This result follows immediately from Theorem 2 and the fact that the integrated weighted Gaussian process follows a normal distribution. ■

Remark. A transformation of the expression for Ω might simplify its estimation. Notice that

$$\int_{\mathcal{T}} \dot{M}(t;\theta_0) B_\Gamma \mathbb{E}\Lambda(dt;\theta_0) = \sqrt{n} \int_{\mathcal{T}} \bar{M}_n(t;\theta_0) \dot{M}(t;\theta_0) \mathbb{E}\Lambda(dt;\theta_0) + o_p(1)$$

and

$$\begin{aligned}
\sqrt{n} \int_{\mathcal{T}} \bar{M}_n(t;\theta_0) \dot{M}(t;\theta_0) \mathbb{E}\Lambda(dt;\theta_0) &= \sqrt{n} \int_{\mathcal{T}} \bar{M}_n(t;\theta_0) \mathbb{E} \frac{\partial}{\partial \theta} \Lambda(t;\theta) \mathbb{E}\Lambda(dt;\theta_0) \mid_{\theta=\theta_0} \\
&= \sqrt{n} \int_{\mathcal{T}} \int_s^T \mathbb{E} \frac{\partial}{\partial \theta} \Lambda(t;\theta) \mathbb{E}\Lambda(dt;\theta_0) \bar{M}_n(ds;\theta_0) \mid_{\theta=\theta_0} \\
&= \sqrt{n} \int_{\mathcal{T}} \psi(s) \bar{M}_n(ds;\theta_0)
\end{aligned}$$

where

$$\psi(s) = \int_s^T \mathbb{E} \frac{\partial}{\partial \theta} \Lambda(t;\theta) \mathbb{E}\Lambda(dt;\theta_0)$$

Here, the second equation is the outcome of applying Fubini's Theorem. Thus,

$$\sqrt{n} \left(\int_{\mathcal{T}} \dot{M}(t;\theta_0) \dot{M}(t;\theta_0)^\top \mathbb{E}\Lambda(dt;\theta_0) \right) (\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, C)$$

where C is a $k \times k$ matrix with entries

$$C_{ij} = \int_{\mathcal{T}} \psi_i(t) \psi_j(t) \mathbb{E}\Lambda(dt;\theta_0)$$

Notice that $\psi(s)$ can be estimated by

$$\hat{\psi}(s) = \int_s^T \frac{\partial}{\partial \theta} \bar{\Lambda}_n(t;\theta) \bar{N}_n(dt) \mid_{\theta=\hat{\theta}} = \frac{1}{N_n} \sum_{l:t_l>s} \frac{\partial}{\partial \theta} \bar{\Lambda}_n(t;\theta) \mid_{\theta=\hat{\theta}}$$

where N_n and t_l are the number of events and event times of the average process $\bar{N}_n((0,T])$, respectively. Similarly, C_{ij} is estimated by

$$\hat{C}_{ij} = \int_{\mathcal{T}} \hat{\psi}_i(t) \hat{\psi}_j(t) \bar{N}_n(dt) = \frac{1}{N_n} \sum_{l=1}^{N_n} \hat{\psi}_i(t_l) \hat{\psi}_j(t_l)$$

The term $\left(\int_{\mathcal{T}} \dot{M}(t;\theta_0) \dot{M}(t;\theta_0)^\top \mathbb{E}\Lambda(dt;\theta_0) \right)$ can be estimated in the same way, and hence, its estimation expression is omitted here.

4.2 Simulation Studies

In this subsection, I conduct Monte Carlo simulations to investigate the finite sample performance of the proposed estimator. The data generating process is the epidemic type aftershock sequence (ETAS) model. The ETAS model was first introduced by Ogata and Katsura (1988) and ever since has been widely used in seismology literature (Zhuang et al.,

2002). The model extends the classical Hawkes model and includes the marks, it characterizes both the earthquake times and magnitudes. The intensity of a ETAS model, for its simplest form, could be:

$$\lambda(t) = \mu + \sum_{j:t_j<t} e^{\alpha x_j} \left(1 + \frac{t - t_j}{c}\right)^{-1}$$

where x_j is the magnitude of an earthquake occurring at time t_j, and the mark density, for simplicity, is assumed to be i.i.d:

$$f(x|t, \mathcal{F}_{t-}) = \delta e^{-\delta x}$$

The above data generating process can be simulated using the R package 'PtProcess' (Harte, 2010).[1]. Details on the simulation algorithm can be found in Appendix B. I set the true parameters as $\mu = 0.007$, $\alpha = 1.98$, $c = 0.008$ and $\delta = log(10)$ and generate $N = 50, N = 100$, $N = 200$ and $N = 400$ individual counting processes. The time-intervals are set to be $(0, 100], (0, 500]$ and $(0, 3000]$. For each simulation setting, we run $B = 1000$ repeats. I report their standard deviation (SD), median of absolute deviation (MAD), 95% confidence interval coverage rate (CI95) and 90% confidence interval coverage rate (CI90). The results are presented below. As the number of observations N increases, the estimators become more stable and their empirical coverage rates get closer to the theoretical ones. It is also noticeable that the performance of estimators is insensitive to the number of events per person. (We increase the length of the time horizon to increase such a number under the same true parameters.)

5. Modelling Dynamic Outpatient Claim Data

I use the Hawkes process framework to study outpatient claims under two different insurance plans in this section. These plans mainly differ in cost-sharing policies. One insurance plan offers a generous zero-coinsurance rate, i.e., all medical costs are covered by the insurance company, while the other insurance plan imposes a 95% coinsurance rate where insurees have to pay most of the medical cost.

5.1 Data

The data comes from the well-known RAND Health Insurance Experiment (RAND HIE), one of the most important health insurance studies ever conducted. The HIE project was started in 1971 and was funded by the Department of Health, Education, and Welfare. The company randomly assigned 5809 people to insurance plans that either had no

1. https://cran.r-project.org/package=PtProcess

Table 1: Minimum Distance Estimator Results, with $T = 100$

$N = 400$	True	Estimator	SD	MAD	CI95	CI90
μ	0.007	0.006747	0.002320	0.001530	95.2%	92.9%
α	1.98	1.980313	1.687546	0.326757	95.1%	94%
c	0.008	0.010274	0.016460	0.006809	95.4%	93.9%
$N = 200$						
μ	0.007	0.006313	0.002893	0.001907	95.2%	92.4%
α	1.98	1.979364	2.092911	0.316262	97.1%	96.2%
c	0.008	0.011875	0.023568	0.007983	96.7%	95.4%
$N = 100$						
μ	0.007	0.013175	0.005717	0.003802	81.5%	75.7%
α	1.98	1.719879	2.227818	0.926524	92.2%	89.6%
c	0.008	0.020892	0.036641	0.016629	89%	86.9%
$N = 50$						
μ	0.007	0.012732	0.006974	0.004389	85.9%	82.9%
α	1.98	1.874360	3.961052	1.036084	95.6%	93.5%
c	0.008	0.021302	0.045482	0.016142	89.2%	87.2%

Table 2: Minimum Distance Estimator Results, with $T = 500$

$N = 400$	True	Estimator	SD	MAD	CI95	CI90
μ	0.007	0.006829	0.001273	0.000783	95.5%	92.7%
α	1.98	1.985477	0.256038	0.071041	96.4%	95.9%
c	0.008	0.008305	0.005284	0.001915	96.1%	95.1%
$N = 200$						
μ	0.007	0.007056	0.001783	0.001321	92.5%	89.6%
α	1.98	1.977045	0.448665	0.217622	91.9%	90.6%
c	0.008	0.009059	0.008174	0.004485	91.5%	89.9%
$N = 100$						
μ	0.007	0.006608	0.0022961	0.001927	90.1%	86%
α	1.98	1.761040	0.850601	0.671524	86.6%	83%
c	0.008	0.016624	0.017485	0.012113	86.7%	83.5%
$N = 50$						
μ	0.007	0.006672	0.002964	0.002222	90.3%	87.9%
α	1.98	1.761366	2.207844	0.778182	91.4%	88.7%
c	0.008	0.018084	0.025082	0.013142	90.6%	87.8%

Table 3: Minimum Distance Estimator Results, with $T = 3000$

$N = 400$	True	Estimator	SD	MAD	CI95	CI90
μ	0.007	0.006957	0.000627	0.000432	94.9%	92. 5%
α	1.98	1.978269	0.073311	0.039946	93.5%	90.8%
c	0.008	0.008131	0.001724	0.000937	93.9%	91.7%
$N = 200$						
μ	0.007	0.006963	0.000832	0.000727	92.4%	87.2%
α	1.98	1.992719	0.104450	0.067616	91.2%	89.8%
c	0.008	0.007930	0.002337	0.001600	90.7%	88.3%
$N = 100$						
μ	0.007	0.006847	0.001146	0.000909	93.4%	90.9%
α	1.98	1.964071	0.165430	0.088718	92.1%	90.1%
c	0.008	0.008571	0.003605	0.002196	92.3%	90.5%
$N = 50$						
μ	0.007	0.006810	0.001541	0.001389	89.1%	84.9%
α	1.98	1.974604	0.276515	0.226873	87.9%	83.7%
c	0.008	0.008980	0.005476	0.004328	86.9%	83.1%

cost-sharing, 25%, 50% or 95% coinsurance rates. The out-of-pocket cap (OPC) varied among different plans, too. The HIE was conducted from 1974 to 1982 in six sites across the USA: Dayton, Ohio, Seattle, Washington, Fitchburg-Leominster and Franklin County, Massachusetts, and Charleston and Georgetown County, South Carolina. These sites represent four census regions (Midwest, West, Northeast, and South), as well as urban and rural areas.

Because of the nonlinear structure of our model, to ease the burden of computation, I only use data from Seattle, which has the largest medical claim records available. For demonstration purpose, I will only focus on two insurance plans: the zero coinsurance plan and the 95% coinsurance plan. In the first plan, an insuree needs not pay anything, while in the second plan, an insuree needs to pay 95% of the medical care costs if the cumulative medical cost is under the OPC. The OPC in this plan is 150 USD per person or 450 USD per family.[2] Once the OPC is reached, the insurance plan would cover all the costs. The OPC and the coinsurance rate in this plan only applied to ambulatory services; inpatient services were free. Both plans cover a wide range of services. Medical expenses include services provided by non-physicians such as chiropractors and optometrists, and prescription drugs and supplies. There is no deductible in this insurance contract.

The time unit is annual. For example, if an insurance contract begins on Jan-01-1977 and the date of a doctor visit is Oct-01-1977, the time stamp is then 0.748 (years). When

2. In 1973 dollars.

preparing the dataset, I delete all records with missing time information. When analyzing the cost-sharing plan, I restrict the dataset within the contract year 1977-1978, since cost-sharing policies are renewed annually. However, this restriction is not needed for the free plan, as there is no within-year cost-sharing policy. For the free plan, the time horizon ranges from 1975 to 1980. At the end, there are 243 individuals in the free plan with 7638 claims over the five years and 131 individuals in the cost-sharing plan with 1103 claims over the 1977-1978 contract year.

The demographic covariates included in the model are age, sex, education (in terms of schooling years) and log-income. For simplicity, I fixed all ages at the enrollment time. Thus, all covariates are time-independent. Other restrictions on the dataset include 1) individuals who are younger than 18 or are older than 60 are excluded from the sample; 2) if the value of a doctor visit cost is not available, I replace it with zero; 3) if information on the education is unknown, I replace it with the average education level.

5.2 Modelling dynamic claim data for a free plan

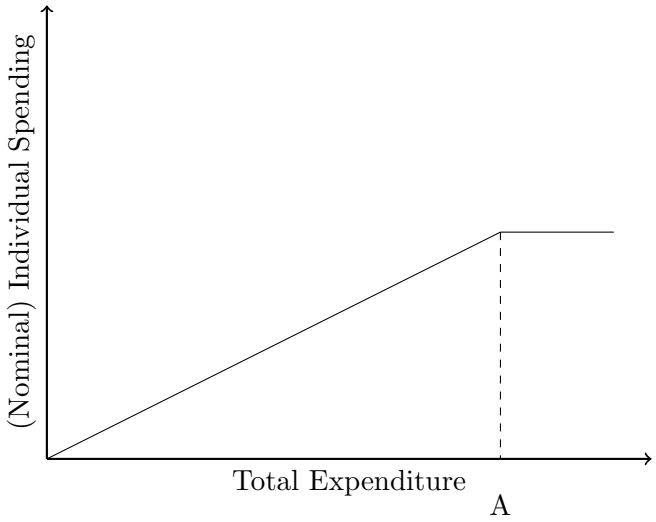
Figure 1 indicates that outpatient claims are clustered in a free insurance plan. One possible explanation is that one outpatient activity might trigger occurrences of future activities. A typical example is the medical follow-up check. The self-exciting part of the Hawkes process is particular suitable for modelling such dynamic mechanism. To this end, I specify the intensity function for an individual i who holds a free insurance contract as:

$$\lambda_i^{P0}(t) = \exp(\lambda_0) + \exp(z_i^\top \gamma) \sum_{j:t_{ij}<t} \alpha \exp(-\mu(t - t_{ij})), \quad t \in \mathcal{T}$$

where the superscript $P0$ denotes the free insurance plan, $\exp(\lambda_0)$ measures background risk of visiting a doctor, z_i is a vector of individual heterogeneities. I assume they enter into the model linearly via an exponential link function. Finally, I specify the exciting function using the conventional exponential kernel with parameters α, μ. Notice that there is no need to specify a constant term for the γ parameters, since the parameter α captures a base level of the intensity function, i.e., $\alpha = \exp(\gamma_0)$.

5.3 Modelling dynamic data for a cost-sharing plan

The cost-sharing plan in this study differs from the free plan by a 95% coinsurance rate. Such a cost-sharing design creates a non-linear budget constraint, as illustrated by Figure 3. How individuals respond to medical costs has been a central question in health economics as well as actuarial research. The bulk of evidence suggests that introducing cost-sharing tools reduces medical spending. More specifically, the reduction is achieved mainly through quantity whereby individuals purchase fewer medical care services, instead of price shopping whereby individuals search for cheaper providers without compromising the quantity (Brot-Goldberg et al., 2017). In addition, costs of medical services are conventionally as-



The total expenditure is the sum of individual spending and expenditures paid by the insurance. Point A is the OPC. When the total expenditure is below A, the co-insurance rate (the slope) $r = 0.95$ is applied. Whenever the total expenditure is beyond A, there is no more cost for individuals (the slope is 0).

Figure 3: Non-linear Individual Spending

sumed to be i.i.d log-normal, see Handel et al. (2015); Keeler and Rolph (1988). Thus, assessing an individual's response to a cost-sharing policy amounts to assessing how this individual adjusts her medical consumption quantities. Historically, literature studying the price elasticity of health insurance contracts often assumes that individuals only respond to the 'spot' price, see Cutler and Zeckhauser (2000); Keeler and Rolph (1988); Manning et al. (1987). Recent literature deviate from this assumption, and assume that individuals might respond to the 'shadow price' as well, see Aron-Dine et al. (2013); Brot-Goldberg et al. (2017); Einav et al. (2015), but finds mixed evidence on individuals' responsiveness to the dynamic incentives created by the cost-sharing health insurance plan.

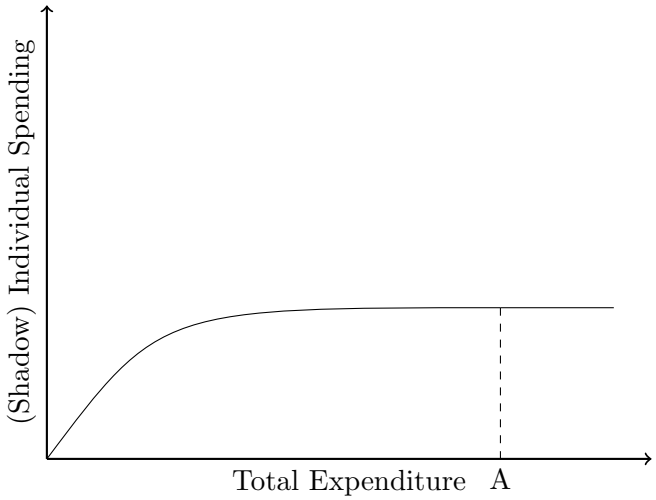
This study adopts the second approach. Specifically, we define the 'shadow coinsurance rate' for an individual i as:

$$r(x_i(t)) = \mathbb{E}(r_{EOY} \mid x_i(t))$$

where $x_i(t)$ is the cumulative individual cost up to time t, and r_{EOY} is the coinsurance rate at the end of a year. Figure 4 illustrates this shadow rate curve.

Note that $x_i(t)$ is an individual-specific and history-dependent stochastic process, which implies that a desired model would depend on a stochastic filtration. To make things more realistic (yet more complicated), I specify $x_i(t)$ as the summation of outpatient costs and drug-purchase costs:

$$x_i(t) = \sum_{j=1}^{N_i(t-)} o_j + \sum_{j=1}^{N_i^1(t-)} d_j$$



The total expenditure is the sum of individual spending and expenditures paid by the insurance. Point A is the OPC. When the total expenditure is below A, the shadow coinsurance rate (the slope) $0 < r(x) < 1$ is a function of cumulative individual spending with $r' < 0$. Whenever the total expenditure is beyond A, there is no cost for individuals.

Figure 4: Non-linear Individual Shadow Price

where $N_i(t)$ is a counting process that consists of outpatient activities, while $N_i^1(t)$ is a counting process formed by drug-purchase times. o_j and d_j are outpatient costs and drug-purchase costs, respectively. Since the primary focus is the outpatient counting process, the drug purchase process, along with its marks d_j, is an external shock to the model, and will create jumps in the interested intensity function. These jumps, however, do not impose any threat to our estimation due to Assumption A2.

I specify the intensity function for an individual i as:

$$\lambda_i^{P95}(t) = \exp(\lambda'_{0,i}) + \exp(z_i^\top \gamma) \sum_{j:t_{ij}<t} p_i(x_i(t_{ij}))\alpha \exp(-\mu(t - t_{ij}))$$

where $P95$ stands for the 95% coinsurance plan,

$$\exp(\lambda'_{0,i}) = \exp(\lambda_0 + \rho \mathbb{I}\{x_i(t) > 150\})$$

measures the background intensity of claims before and after an individual exceeds the OPC threshold. The term

$$p_i(x_i(t)) = (1-b)\exp(\beta_1(x_i(t) - 150))\mathbb{I}\{x_i(t) \leq 150\} + 1 - \mathbb{I}\{x_i(t) \leq 150\}$$

is the compound price effect. It consists of three parts. The first part, $1 - b$, measures an individual's responsiveness towards the spot coinsurance rate, while the second part, $\exp(\beta_1(x_i(t) - 150)) = 1 - r(x_i(t))$, is a parametric specification of the shadow coinsurance

rate. Here, to simplify the analysis, I assume that the OPC threshold is uniformed at 150 USD per individual. When $\beta_1 > 0$, $\exp(\beta_1(x_i(t)-150))$ is less than one if the cumulative cost is below the threshold, but it will gradually approach to one, since $x_i(t)$ is non-decreasing with time t. The first two parts enter into the model if an individual's cumulative cost is less than the OPC threshold. The last part, $1-\mathbb{I}\{x_i(t) \leq 150\}$, normalizes the compound price effect to unit if the cumulative cost exceeds the OPC threshold. The rest of the parameters and specifications are identical to the case of the free plan.

6. Results and Interpretations

I will first present the interpretation of the model, and then list estimation results. The interpretation consists of two parts. First, how to interpret coefficients of individual heterogeneities, and second, how to interpret the cluster structure. Empirical results presented here will shed light on the cluster structure of outpatient claims, as well as the effect of accumulated claim severity on the individual claim frequency.

6.1 Interpreting the model

The interpretation of individual heterogeneity effects is not straightforward due to the nonlinear nature of $\exp(z_i^\top \gamma)$. However, one may fix a period and treat the counting process as a count data. The interpretation is then identical to that of the marginal effect at a representative value (MER) of a count data model. Let $Y_{it} = N_i(t)$ be the number of events that occurred before time t. Let scalar z_{ij} denote the $j-th$ covariate. Differentiating

$$\frac{\partial \mathbb{E}(Y_{it}|Z_i = z_i)}{\partial z_{ij}} = \gamma_j \mathbb{E}(\Lambda_i(t|Z_i = z_i) - \exp(\lambda_0)t)$$

by the exponential structure of $\exp(z_i^\top \gamma)$. For example, if $\hat{\gamma_j} = 0.2$, $\bar{\Lambda}_n(t|Z_i = z_i) - \exp(\hat{\lambda}_0)t = 2.5$, then a one-unit change in the $j-th$ covariate increases the expectation of Y_t by 0.5 units.

The interpretation of cluster structures demands separate discussions for the free plan and the cost-sharing plan. The introduction of individual heterogeneities in the free plan complicates the cluster structure. Different individuals might have different cluster structures:

$$n_i^* = \frac{\exp(z_i^\top \gamma)\alpha}{\mu}$$

but the expected branching ratio over individuals is well-defined:

$$\mathbb{E}n_i^* = \frac{\alpha \mathbb{E}\exp(z_i^\top \gamma)}{\mu}$$

By stationarity, the expected numer of claims during a time interval $(0, T]$ is

$$\mathbb{E}N((0,T]) = \frac{\exp(\lambda_0)T}{1-\mathbb{E}n_i^*}$$

which can be estimated by replacing $\mathbb{E}n_i^*$ with $1/n \sum_{i=1}^n \exp(z_i^\top \hat{\gamma})\hat{\alpha}/\hat{\mu}$.

Interpreting the cluster structure of the cost-sharing plan is challenging, since not only individual heterogeneities enter into the model, the compound price effect $p_i(t)$ also affects the cluster structure:

$$\begin{aligned} n_i^* &= \int_0^\infty \exp(z_i^\top \gamma) p_i(x_i(t)) \alpha \exp(-\mu t) dt \\ &= \exp(z_i^\top \gamma) \sum_{j=1}^{N_i(T)} \frac{p_i(x_i(t_{ij}))\alpha}{\mu} \left(\exp(-\mu t_{i(j-1)}) - \exp(-\mu t_{ij})\right) \end{aligned}$$

here, we use the fact that $x_i(t)$ is piecewise-constant. $n_i^* < 1$ as long as $\alpha \exp(z_i^\top \gamma)/\mu < 1$. Finding an expression for $\mathbb{E}n_i^*$ is nontrivial, since $x_i(t)$ is also stochastic. Nevertheless, one can find lower and upper bounds for a branching ratio in the cost-sharing plan,

$$\frac{\exp(z_i^\top \gamma)\alpha p_i(0)}{\mu} \leq n_i^* \leq \frac{\exp(z_i^\top \gamma)\alpha}{\mu}$$

The lower bound is the branching ratio of a Hawkes process if one 'freezes' the compound price at $p_i(0)$, while the upper bound is the branching ratio of a process in the free plan. Consequently, the lower and upper bounds for the expected counts $\mathbb{E}N_i((0,T])$ can be computed accordingly. Using the simulation method documented in Appendix B, one can also estimate $\mathbb{E}N_i((0,T])$ via parametric bootstrap once parameters have been estimated.

6.2 Results

With these interpretations, I now list the estimation results, presented in Table 4. First, let's focus on the free plan. There are two types of results, one on the explanatory variables and one on the cluster structure. For explanatory variables, I include age, sex, education (in terms of schooling years) and log-income as individual covariates. I introduce age2 and edu2 to model the nonlinear effect of age and education, and they are defined as $age^2/100$ and $edu^2/100$, respectively.

- *Age.* At first, intensity values will decrease as age increases. After one passes the age of 46, intensity values and age are positively correlated. It is well-known that youngsters are more risky compared to their mid-age counterparts. While, as individuals begin to age, they become physically weaker and are more prone to sickness.

- *Sex.* Females seem to be more likely to visit the doctor.

- *Education and Income.* Income is positively correlated with the use of medical service. The result on education, by and large, suggests an u-shape relation between education and the outpatient utilization. The better education one obtains, the fewer outpatient activities one would conduct. The turning point is around 15 schooling years, roughly

the third year of the college. After which, better education implies higher chance to visit a doctor. One explanation is that higher education often associates with a healthier lifestyle, which reduces the hazard rate of visiting a doctor. While a college degree is highly correlated with high income, which gives individuals the ability to cover the opportunity cost related to the absence from work.

Next, let's focus on the cluster parameters, namely α and μ. A Wald test on $\alpha = \mu = 0$ yields the statistic with a value around 189.1, suggesting to reject the null. The estimated branching ratio and cluster size are $\hat{n}^* = \hat{\alpha}/\hat{\mu} = 0.782$, $\hat{N}_\infty = 1/(1-\hat{n}^*) = 4.587$, respectively for a representative individual whose heterogeneities are normalized to one.

Estimated coefficients in the cost-sharing plan are quite similar to the ones in the free plan. This is hardly a surprising given the random assignment experiment design of the Rand HIE project. The standard errors in the cost-sharing plan are large, but I believe this is due to the relative small sample size of the plan. A Wald test on the null hypothesis that all the individual heterogeneity coefficients are zero yields the statistic with a value over 6140, strongly rejecting the null. Estimated cluster parameters $\hat{\alpha}$ and $\hat{\mu}$ are also close to the ones in the free plan and are statistically indifferent.

The most interesting results in the cost-sharing plan are compound price effect estimators. First, the background intensity $\exp(\lambda_0')$ does not change before and after one passes the OPC threshold (one does not reject the null that $\rho = 0$). Second, both b and β_1 are significantly different from zero, implying individuals would respond to both the shadow coinsurance rate and the spot coinsurance rate. As mentioned before, one could 'freeze' the compound price effect at $p_i(0)$ to calculate a lower bound for the branching ratio. Thus, the lower bound for the cluster size is estimated at $1/(1-(1-\hat{b})\exp(\hat{\beta}_1(0-15))\hat{\alpha}/\hat{\mu}) = 1.203$ for a representative individual.

7. Discussion

In this section, I will first compare the Hawkes process framework with other approaches that used to model individual claim frequency, namely, the count data model and the dynamic discrete choice model. In the second part of this section, I will discuss some limitations of the Hawkes process framework.

7.1 Comparison with other models in the individual claim frequency literature

The claim frequency is vital to determine the risk premium, and often is measured by the number of claims of an individual over a period. Naturally, count data models are widely applied in the insurance literature, see Desjardins et al. (2001); Englund et al. (2009); Pinquet (1997). The Poisson model is the most classical count data model. Given the individual unobserved risk profile, the number of claims is assumed to be Poisson distributed. However, Poisson models are quite restrictive, as it requires the data to be equidispersion.

Table 4: Main Results

	Free Plan	*Cost-Sharing Plan*
α	5.173	4.09
	(3.423)	(0.451)
μ	6.619	7.036
	(3.664)	(0.942)
age	-0.49	-0.552
	(0.06)	(1.632)
age2	0.532	0.432
	(0.08)	(3.44)
male	-0.424	-0.862
	(0.265)	(0.753)
edu	-1.13	-0.961
	(0.095)	(2.888)
edu2	3.861	3.503
	(0.347)	(11.421)
log income	1.82	1.962
	(0.137)	(1.126)
λ_0	-0.342	-0.422
	(0.016)	(2.714)
ρ		0.065
		(5.895)
b		0.509
		(0.038)
β_1		0.035
		(0.007)

I replace the cumulative cost $x(t)$ with $x(t)/10$ in the model to avoid overflow in computing. Consequently, the OPC threshold is replaced by 15.0.

Recognizing this limitation, one might model inter-arrival times (durations) with various distributions (e.g., Gamma, Weibull, Mittag-Leffler) to allow for overdispersion or underdispersion. The heterogeneous Gamma-Poisson (negative binomial) model is routinely used to model overdispersed count data. Although rare in insurance data, several econometric studies have developed models to fit underdispersion data, see Cameron and Trivedi (2013) for reference. One source of overdispersion is the excess of zeros. The claim data in automobile insurance is one typical example. The no claim discount system, which is widely adopted by automobile insurers, makes policyholders seldom make a claim if the loss is small. Zero-inflation and hurdle models are often used in such circumstances.

A count data is usually the outcome of an underlying counting process. Here, I will show that the Hawkes process can generate counts that are overdispersion and are zero-excessive. To do so, I simulate the same ETAS model[3] with 1000 repeats and report the count result. In this simulation study, I set the parameters at $\mu = 0.007, \alpha = 0.5$ and $c = 0.018$. The time interval is $\mathcal{T} = (0, 200]$. Figure 5 presents the simulation result. For a given time interval, 24.6% of individuals report no claim. While the average of counts is 1.743, the variance of counts is 2.766, so the data is overdispersion.

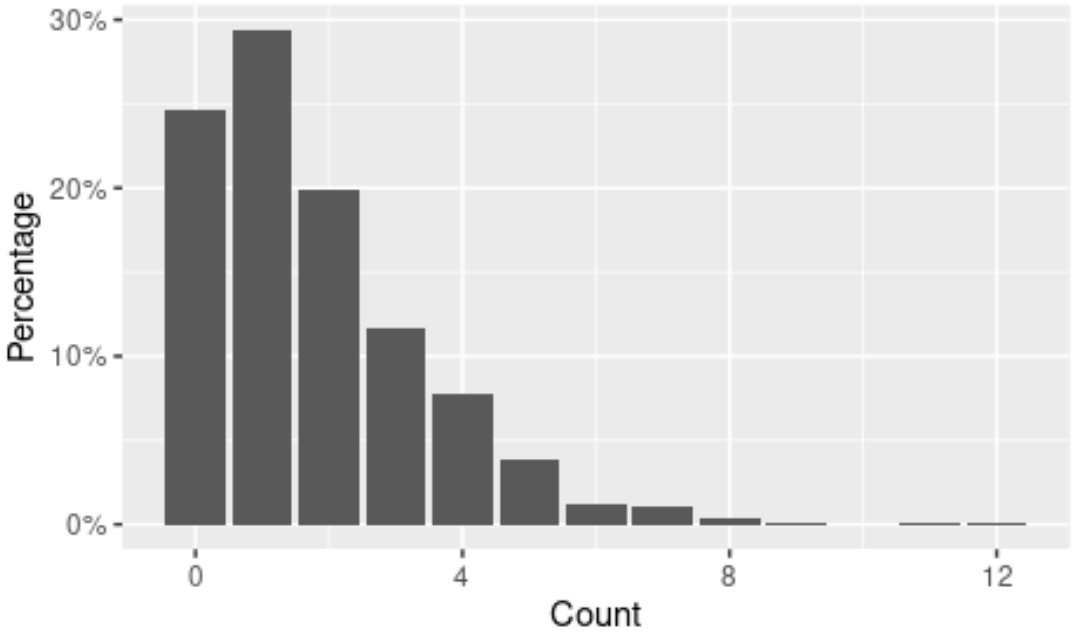


Figure 5: Histogram of Counts

In health economics literature, the structure dynamic discrete choice model is widely used to model the occurrence probability of an individual's medical activity. For example, Cronin (2019) developed a within-year-decision-making model to quantify the moral haz-

3. See Section 3.2 for details.

ard effect of health insurance on medical expenditure. However, this model assumes that individuals decide only at the beginning of each month. Thus, the author is effectively aggregating the daily data into a monthly data. Such aggregation would alter data properties, as illustrated in the right panel of Figure 1.

Within the context of outpatient activity choices, the dynamic discrete choice model can be summarized as follows.

1. Specify a base intensity for an individual i, e.g., $v_i = \exp(z_i^\top \gamma)$.

2. In each period, update the cumulative individual cost x_{it} as well as a self-exciting state Q_{it}:
$$Q_{it+1} = \alpha d_{it} + (1-\delta) Q_{it}$$
where d_{it} is an indicator that takes value of one if individual i conducts an outpatient activity on period t and zero otherwise. δ is a discount rate.

3. Let $h(x_{it})$ be a parametric specification for the coinsurance rate, T be the terminal period. The shadow coinsurance rate $r(x_{it}, Q_{it}, v_i) = \mathbb{E}(h(x_{it})|x_{it}, Q_{it}, v_i)$ can be solved recursively backward from a terminal function:
$$r(x_{iT}, Q_{iT}, v_i) = h(x_{iT})$$
with the law of iterated expectations:
$$r(x_{it}, Q_{it}, v_i) = \kappa_{it}\mathbb{E}(r(x_{it+1}, \alpha + (1-\delta)Q_{it}, v_i)|d_{it} = 1) + (1-\kappa_{it})r(x_{it}, (1-\delta)Q_{it}, v_i)$$
where
$$\kappa_{it} = \mathbb{E}(d_{it}|r(x_{it}, Q_{it}, v_i), Q_{it}, v_i)$$
is the conditional probability of conducting an outpatient activity in this period given current states.

4. With these equations and some additional computation, a likelihood for all individual-periods $\{d_{it}\}_{i,t}$ can be formed, and the estimators can be obtained using the maximum likelihood method.

Although such structure models provide clear identification and interpretation of the model parameters, key specifications in this framework would lead to a data generating process that is quite different from the one observed.

- In step 2, a self-exciting state is updated in a Markov way, i.e., the future event only depends on the current event. This specification will generate stationary data, hence, the cluster pattern one observes in Figure 1 is lost.

- In step 3, the shadow price is calculated recursively and implicitly, and consequently, there is no closed form specification of the count expectation. Furthermore, there is no clear algorithm to simulate counts based on this recursive specification. This limits the application in actuarial practice.

7.2 Limitations of the Hawkes process framework

There are two limitations in the Hawkes process framework presented in this paper. First, I do not consider the unobserved heterogeneity in the model. Since the seminal work of Heckman (1981), it is clear that unobserved heterogeneity could generate the so called 'spurious state dependence': events are correlated via this latent term, yet once conditional on it, events are independent. Researchers often impose random effects assumptions to distinguish between true and spurious state dependence. The random effects would require the unobserved heterogeneity follows a particular distribution and is uncorrelated with other explanatory variables. However, this approach is difficult to implement in this study, since the exciting part, by construction, is correlated with this latent element. Thus, a fixed effect approach is more suitable in this context. However, neither where to put this effect in the model, nor how to 'cancel' the fixed effect in the model is clear at this moment. One workaround is to use 'the number of events occurred before' as a proxy to the unobserved heterogeneity, but this would inevitably introduce measurement errors in the model, and make the estimators unstable.

The second limitation is the lack of proper goodness-of-fit tests. Conventionally, the Random Time Change Theorem has been proposed for specification test. This theorem states that, for a counting process $N(t)$ with points $t_1, t_2, \ldots, t_m$, one can transform these occurrence times to:

$$Y_j(\theta_0) = \int_0^{t_j} \lambda(s;\theta_0)ds, \quad j = 1, 2, \ldots, m$$

where $\lambda(t;\theta_0)$ is the corresponding intensity function, such that $\{Y_j\}_{j=1,2,\ldots,m}$ form a Poisson process with unit rate, and the transformed duration $\Delta_j(\theta_0) = Y_{j+1}(\theta_0) - Y_j(\theta_0)$ are i.i.d and should follow an exponential distribution with unit rate.

Suppose $\hat{\theta}$ is a vector of root-n consistent estimators of θ_0, one might obtain a Q-Q plot using $\{\Delta_j(\hat{\theta})\}_j$, or perform a Kolmogorov-Smirnov type test to check if the model is corrected specified. However, these methods are not rigorous because the estimation effect is not taken into account: The random time change theorem is valid only if the model is correctly specified. After replacing θ_0 with $\hat{\theta}$, neither the distribution of $\Delta_j(\hat{\theta})$ is known, nor the independence among $\{\Delta_j(\hat{\theta})\}_j$ holds.

8. Conclusion

In this paper, I use the Hawkes process as an alternative model to model individual outpatient claim data under different insurance plans. Using this new framework, I model and analyze data cluster patterns, estimate marginal effect at a representative value, and study an individual-specific and history-dependent shadow coinsurance rate. The Hawkes process is a counting process with the self-exciting property, i.e., past claims will affect the occurrence of future claims. It also has a branching structure, which naturally leads to a

cluster interpretation. These properties make the Hawkes process an ideal tool to model claim data.

Existing literature use the maximum likelihood method to estimate one stationary Hawkes process. In this study, outpatient activities of each individual consist of an observation process. For each process, I include individual heterogeneities in the model, this makes the underlying process not necessarily stationary. The Doob-Meyer decomposition result is used to construct a continuous moment restriction, and I use a minimum distance method to estimate the model. Consistency and asymptotic normality results are proven, and Monte Carlo simulations have been implemented. The simulation results suggest good finite sample performances.

I illustrate the use of the model based on the Rand Health Insurance Experiment data. Specifically, I model two insurance plans with different cost-sharing policies: one with zero coinsurance rate and one with a 95% coinsurance rate. For the first plan, the data cluster might be a consequence of the state dependent effect among claims, while for the cost-sharing plan, the moral hazard might also be a source. I found that for a representative individual whose heterogeneities are normalized to one, the cluster size is 4.587 in the free plan, while in the cost-sharing plan, the lower bound for the cluster size is 1.203. In addition, I estimated marginal effects as well as the shadow coinsurance rate. The estimation results suggest that individuals understand the cost-sharing nature of a health insurance and would respond to both the shadow and spot coinsurance rates.

Lastly, I compare the new approach with count data and structure dynamic discrete choice models. Both types of models are unable to characterize the cluster pattern of the claim data. For the count data model, the reason is obvious, since it aggregates the whole counting process into one variable. For dynamic discrete choice models, the Markov specification of the self-exciting state is difficult to reproduce cluster patterns observed in the data. On the other hand, I show, through simulation, that overdispersed and zero-excessive count data can be generated by a Hawkes process. I also discuss some limitations of the proposed model. First, it is difficult to introduce the unobserved heterogeneity in a random effect way, since the self-exciting part is, by construction, correlated with the latent individual variable. In addition, it is not clear where to put a fixed effect into the model, nor does one know the transformation to cancel this fixed effect. The second limitation is the lack of proper goodness-of-fit tests. Conventionally used methods are based on the random time change theorem, which is valid only if the parameters are the true ones. Replacing true parameters with their estimators introduces the estimation effects, making the distribution of the transformed times unknown.

References

Abbring, J. H., P.-A. Chiappori, and J. Pinquet (2003a): "Moral hazard and dynamic insurance data," *Journal of the European Economic Association*, 1, 767–820.

Abbring, J. H., J. J. Heckman, P.-A. Chiappori, and J. Pinquet (2003b): "Adverse selection and moral hazard in insurance: Can dynamic data help to distinguish?" *Journal of the European Economic Association*, 1, 512–521.

Aron-Dine, A., L. Einav, and A. Finkelstein (2013): "The RAND health insurance experiment, three decades later," *Journal of Economic Perspectives*, 27, 197–222.

Bacry, E., I. Mastromatteo, and J.-F. Muzy (2015): "Hawkes processes in finance," *Market Microstructure and Liquidity*, 1, 1550005.

Bowsher, C. G. (2007): "Modelling security market events in continuous time: Intensity based, multivariate point process models," *Journal of Econometrics*, 141, 876–912.

Brémaud, P. and L. Massoulié (2001): "Hawkes branching point processes without ancestors," *Journal of applied probability*, 38, 122–135.

Brot-Goldberg, Z. C., A. Chandra, B. R. Handel, and J. T. Kolstad (2017): "What does a deductible do? The impact of cost-sharing on health care prices, quantities, and spending dynamics," *The Quarterly Journal of Economics*, 132, 1261–1318.

Cameron, A. C. and P. K. Trivedi (2013): *Regression analysis of count data*, vol. 53, Cambridge university press.

Chavez-Demoulin, V., A. C. Davison, and A. J. McNeil (2005): "Estimating value-at-risk: a point process approach," *Quantitative Finance*, 5, 227–234.

Cheng, Z. and Y. Seol (2020): "Diffusion approximation of a risk model with non-stationary Hawkes arrivals of claims," *Methodology and Computing in Applied Probability*, 22, 555–571.

Cronin, C. J. (2019): "Insurance-induced moral Hazard: a dynamic model of within-year medical care decision making under uncertainty," *International Economic Review*, 60, 187–218.

Cutler, D. M. and R. J. Zeckhauser (2000): "The anatomy of health insurance," *Handbook of health economics*, 1, 563–643.

Daley, D. J. and D. Vere-Jones (2007): *An introduction to the theory of point processes: volume II: general theory and structure*, vol. 1,2, Springer Science & Business Media.

Dassios, A. and H. Zhao (2012): "Ruin by dynamic contagion claims," *Insurance: Mathematics and Economics*, 51, 93–106.

Desjardins, D., G. Dionne, and J. Pinquet (2001): "Experience rating schemes for fleets of vehicles," *ASTIN Bulletin: The Journal of the IAA*, 31, 81–105.

Einav, L., A. Finkelstein, and P. Schrimpf (2015): "The response of drug expenditure to nonlinear contract design: evidence from medicare part D," *The quarterly journal of economics*, 130, 841–899.

Embrechts, P., T. Liniger, and L. Lin (2011): "Multivariate Hawkes processes: an application to financial data," *Journal of Applied Probability*, 48, 367–378.

Englund, M., J. Gustafsson, J. P. Nielsen, and F. Thuring (2009): "Multidimensional credibility with time effects: An application to commercial business lines," *Journal of Risk and Insurance*, 76, 443–453.

Handel, B. R., J. T. Kolstad, and J. Spinnewijn (2015): "Information frictions and adverse selection: Policy interventions in health insurance markets," Tech. rep., National Bureau of Economic Research.

Harte, D. (2010): "PtProcess: An R package for modelling marked point processes indexed by time," *Journal of Statistical Software*, 35, 1–32.

Hawkes, A. G. (1971): "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, 58, 83–90.

Heckman, J. J. (1981): "Heterogeneity and state dependence," in *Studies in labor markets*, University of Chicago Press, 91–140.

Jang, J. and A. Dassios (2013): "A bivariate shot noise self-exciting process for insurance," *Insurance: Mathematics and Economics*, 53, 524–532.

Keeler, E. B. and J. E. Rolph (1988): "The demand for episodes of treatment in the health insurance experiment," *Journal of health economics*, 7, 337–367.

Kopperschmidt, K. and W. Stute (2013): "The statistical analysis of self-exciting point processes," *Stast. Sinica*, 23, 1273–1298.

Lewis, P. A. and G. S. Shedler (1979): "Simulation of nonhomogeneous Poisson processes by thinning," *Naval Research Logistics Quarterly*, 26, 403–413.

Manning, W. G., J. P. Newhouse, N. Duan, E. B. Keeler, and A. Leibowitz (1987): "Health insurance and the demand for medical care: evidence from a randomized experiment," *The American economic review*, 251–277.

Mohler, G. O., M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita (2012): "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*.

Ogata, Y. (1981): "On Lewis' simulation method for point processes," *IEEE Transactions on Information Theory*, 27, 23–31.

——— (1998): "Space-time point-process models for earthquake occurrences," *Annals of the Institute of Statistical Mathematics*, 50, 379–402.

Ogata, Y. and K. Katsura (1988): "Likelihood analysis of spatial inhomogeneity for marked point patterns," *Annals of the Institute of Statistical Mathematics*, 40, 29–39.

Öztürk, Ö. and T. P. Hettmansperger (1997): "Generalised weighted Cramér-von Mises distance estimators," *Biometrika*, 84, 283–294.

Pinquet, J. (1997): "Allowance for cost of claims in bonus-malus systems," *ASTIN Bulletin: The Journal of the IAA*, 27, 33–57.

——— (2000): "Experience rating through heterogeneous models," in *Handbook of Insurance*, Springer, 459–500.

Rao, R. R. (1962): "Relations between weak and uniform convergence of measures with applications," *The Annals of Mathematical Statistics*, 659–680.

Seal, H. L. (1969): "Stochastic theory of a risk business," Tech. rep.

Stabile, G. and G. L. Torrisi (2010): "Risk processes with non-stationary Hawkes claims arrivals," *Methodology and Computing in Applied Probability*, 12, 415–429.

Swishchuk, A., R. Zagst, and G. Zeller (2021): "Hawkes processes in insurance: Risk model, application to empirical data and optimal investment," *Insurance: Mathematics and Economics*, 101, 107–124.

Zhu, L. (2013): "Ruin probabilities for risk processes with non-stationary arrivals and subexponential claims," *Insurance: Mathematics and Economics*, 53, 544–550.

Zhuang, J., Y. Ogata, and D. Vere-Jones (2002): "Stochastic declustering of space-time earthquake occurrences," *Journal of the American Statistical Association*, 97, 369–380.

A. Proofs

A.1 Theorem 1

Since $\int_{\mathcal{T}} M(t;\theta)^2 \mathbb{E}\Lambda(dt;\theta_0)$ has an unique minimizer at $\theta = \theta_0$, then, using the theory of M-estimator, one just need to show uniformly in θ

$$\int_{\mathcal{T}} \bar{M}_n(t;\theta)^2 \bar{N}_n(dt) \xrightarrow{a.s} \int_{\mathcal{T}} M(t;\theta)^2 \mathbb{E}\Lambda(dt;\theta_0)$$

Notice that A2 implies $\bar{M}_n(t;\theta)$ is continuous in θ. This result holds by applying the continuous mapping theorem and the fact that $\bar{N}_n(t), \bar{\Lambda}_n(t;\theta)$ are sample mean of i.i.d nondecreasing process, a Glivenko-Cantelli argument yields, with probability one, uniform convergence of $\bar{N}_n(t), \bar{\Lambda}_n(t;\theta)$ to $\mathbb{E}N_1(t) = \mathbb{E}\Lambda(t;\theta_0), \mathbb{E}\Lambda_1(t;\theta)$, respectively, uniform in t and compact subsets of Θ.

A.2 Theorem 2

The following Lemma is needed to prove Theorem 3.

Lemma 1. Let θ^* be a consistent estimator of θ_0, then

$$\frac{1}{n}\sum_{i=1}^{n} \dot{M}_i(t;\theta^*) \xrightarrow{a.s} \mathbb{E}\dot{M}_1(t;\theta_0)$$

Proof See Rao (1962) ■

The first order condition of the minimization problem is

$$\sum_{l=1}^{N_n} \left(\sum_{i=1}^{n} \dot{M}_i(t_l;\hat{\theta}_n)\right)\left(\sum_{i=1}^{n} M_i(t_l;\hat{\theta}_n)\right) = 0$$

where in a similar notation, $\dot{M}_i(t;\theta) = \partial M_i(t;\theta)/\partial\theta$. By the mean value theorem, one can find an estimator $\theta_n^* = \gamma\hat{\theta}_n + (1-\gamma)\theta_0, \gamma \in (0,1)$ such that

$$\sum_{l=1}^{N_n} \left(\sum_{i=1}^{n} \dot{M}_i(t_l;\hat{\theta}_n)\right)\left(\sum_{i=1}^{n} M_i(t_l;\theta_0)\right) + G_n(\hat{\theta}_n - \theta_0) = 0$$

where

$$G_n = \sum_{l=1}^{N_n} \left(\sum_{i=1}^{n} \dot{M}_i(t_l;\hat{\theta}_n)\right)\left(\sum_{i=1}^{n} \dot{M}_i^{\top}(t_l;\theta_n^*)\right)$$

Therefore

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) &= n^3 G_n^{-1}\left(\frac{1}{n}\sum_{l=1}^{N_n}\left[\frac{1}{n}\sum_{i=1}^{n}\dot{M}(t_l;\hat{\theta}_n)\right]\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} M_i(t_l;\theta_0)\right]\right) \\ &= n^3 G_n^{-1}\int_{\mathcal{T}}\left[\frac{1}{n}\sum_{i=1}^{n}\dot{M}(t;\hat{\theta}_n)\right]\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} M_i(t;\theta_0)\right]\bar{N}_n(dt)\end{aligned}$$

Notice that

$$n^3 G_n^{-1} = \left(\frac{1}{n} \sum_{l=1}^{N_n} \left(\frac{1}{n} \sum_{i=1}^{n} \dot{M}_i(t_l; \hat{\theta}_n) \right) \left(\frac{1}{n} \sum_{i=1}^{n} \dot{M}_i^\top(t_l; \theta_n^*) \right) \right)^{-1}$$
$$= \left(\int_{\mathcal{T}} \left(\frac{1}{n} \sum_{i=1}^{n} \dot{M}_i(t; \hat{\theta}_n) \right) \left(\frac{1}{n} \sum_{i=1}^{n} \dot{M}_i^\top(t; \theta_n^*) \right) \bar{N}_n(dt) \right)^{-1}$$

thus, by LLN and Lemma1,

$$n^3 G_n^{-1} \xrightarrow{a.s} \left(\int_{\mathcal{T}} \dot{M}(t; \theta_0) \dot{M}(t; \theta_0)^\top \mathbb{E}\Lambda(t; \theta_0) \right)^{-1}$$

Similarly, by Lemma 1, $\frac{1}{n} \sum_{i=1}^{n} \dot{M}(t; \hat{\theta}_n) \xrightarrow{a.s} \dot{M}(t; \theta_0), \forall t \in \mathcal{T}$. Finally, by martingale CLT, one have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} M_i(\cdot; \theta_0) \Rightarrow B_\Gamma(\cdot)$$

where $\Rightarrow$ denotes weakly convergence, and B_Γ is a centered Gaussian process with covariance structure $\Gamma(t, s) = \mathbb{E}(M_1(t; \theta_0) M(s; \theta_0))$.

Putting everything together, one has the result stated in Theorem 2.

B. Simulation Details

We use the *thinning method* to generate the data. This method was first introduced by Lewis and Shedler (1979); Ogata (1981). The procedure consists of

1. Let τ be the start point of a small simulation interval
2. Take a small interval $(\tau, \tau + \delta)$
3. Calculate the maximum of $\lambda(t)$ in the interval as

$$\lambda_{max} = \max_{t \in (\tau, \tau + \delta)} \lambda(t)$$

4. Simulate an exponential random number ξ with rate λ_{max}
5. if

$$\frac{\lambda_g(\tau + \xi | \mathcal{F}_{t-})}{\lambda_{max}} < 1$$

go to step 6.

Else no events occurred in interval $(\tau, \tau + \delta)$, and set the start point at $\tau \leftarrow \tau + \delta$ and return to step 2

6. Simulate a uniform random number U on the interval $(0, 1)$

7. If

$$U \leq \frac{\lambda_g(\tau+\xi|\mathcal{F}_{t-})}{\lambda_{max}}$$

   then a new 'event' occurs at time $t_i = \tau + \xi$. Simulate the associated marks for this new event.

8. Increase $\tau \leftarrow \tau + \xi$ for the next event simulation

9. Return to step 2