Assessing Individual's Response to the Nonlinear Health Insurance Plan: Evidence From A Hawkes Process Framework

YUHAO LI
Economics and Management School, Wuhan University

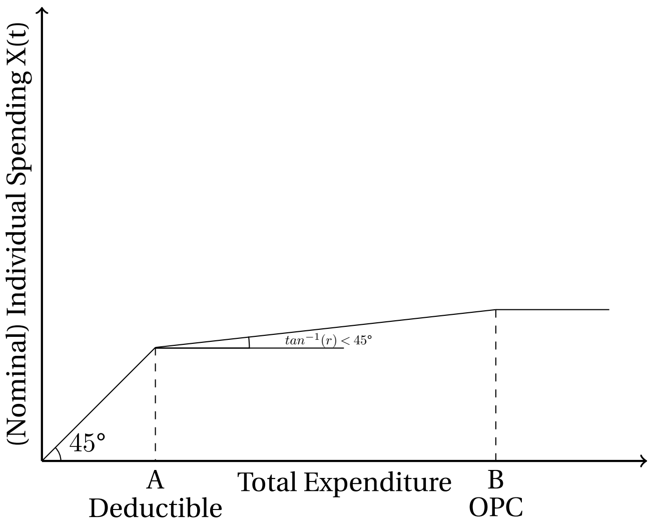
This paper studies an individual's responsiveness to the outpatient utilization under a nonlinear health insurance contract. Specifically, we aim to investigate how a change of the shadow price of medical services would affect an individual's doctor-visit probability given all the history information. We define the episode-varying shadow price as the conditional expectation of the end-of-year co-insurance rate given the cumulative individual spending at current time. Conditional on a stochastic cumulative individual spending process creates econometric challenges as the data generating process is a doubly stochastic process. We represent occurrence times of each doctor visit as points in a Hawkes process. The structure of the Hawkes process enables us to account for various history-dependent effects, including the shadow price effect. In addition, the Hawkes process framework also provides a way to analyze the cluster pattern of an individual's doctor visits. Studying the RAND Health Insurance Experiment data, we found that individuals do understand the price incentive of a nonlinear contract. We also found that setting a free insurance plan as a control, a same canonical individual who is in a cost-sharing insurance plan would have fewer doctor visit clusters, and the average number of doctor visits within each cluster shrinks as well.

KEYWORDS. Health Insurance, Shadow Price, Hawkes Process, Dynamic Behavior.
JEL CLASSIFICATION. C41, C51, I12, I13.

1. INTRODUCTION

Cost-sharing policies in a health insurance contract are widely used in practice as tools to control the moral hazard, i.e., the additional health care that is purchased when an individual became insured. The most commonly used cost-sharing polices are the deductible, the co-insurance rate and the out-of-pocket fee cap (OPC). In a typical setup, individuals need to cover all medical expenditures until the deductible. Once this threshold is passed, the co-insurance policy is applied, where individuals pay a part of the expenditures based on the co-insurance rate. Finally, if the total expenditure paid by the individual passes the OPC, no cost (or very little cost) would be paid by this individual. Figure 1 illustrates such a typical non-linear budget constraint. The bulk of the

Yuhao Li: liyuhao.econ@whu.edu.cn
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.



The total expenditure is the sum of individual spending and expenditures paid by the insurance. Point A and B are the deductible threshold and OPC, respectively. When the total expenditure is below A, the co-insurance is 100% (individuals pay all cost) and the slope is 1. Between A and B, a co-insurance rate (the slope) $0 < R < 1$ is applied. Whenever the total expenditure is beyond B, there is no more cost for individuals (the slope is 0).

FIGURE 1. Non-linear Individual Spending

evidences suggest that introducing cost-sharing tools do reduce spending. More specifically, the reduction is achieved mainly through quantity whereby individuals purchase fewer medical care services, instead of the price shopping whereby individuals search for cheaper providers without compromising the quantity (Brot-Goldberg et al., 2017) . Thus, assessing an individual's response to a nonlinear health insurance plan amounts to assessing how this individual adjusts her medical consumption quantity under a nonlinear price system.

Measuring a consumer's responsiveness to the medical care price is a central issue in health economics and a key ingredient in the optimal design of health insurance contracts. Historically, literature studying the price elasticity of health insurance contracts often assume that individuals only respond to the 'spot' price. For example, Cutler and Zeckhauser (2000) summarize about thirty studies that adopt this assumption. Perhaps, the most famous result in this strand of literature is Manning et al. (1987), Keeler and Rolph (1988), where they obtain the price elasticity of -0.2 in the RAND Health Insurance Experiment (RAND HIE). However, most health insurance contracts, including the ones in the RAND HIE, are highly nonlinear. Therefore, trying to summarize an individual's medical spending behavior with single price elasticity is not well-defined. As mentioned in Aron-Dine et al. (2015, 2013), 'It begs the question, with respect to which price?', and 'In general, there is no "right" way to summarize a nonlinear budget set with a single price'. In addition, the adoption of the spot price implicitly assumes that consumers may not appropriately understand the price incentive of their insurance contract.

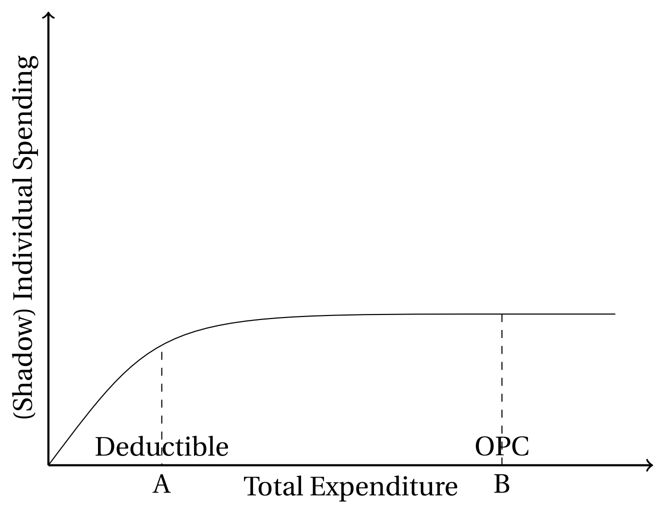
Recent literature deviate from this assumption, see Aron-Dine et al. (2013), Einav et al. (2015), Brot-Goldberg et al. (2017), but find mixed evidence on individuals' responsiveness to the dynamic incentives created by the cost-sharing health insurance plan. These literature often depend heavily on specific data structure and strong homogeneous assumptions. For example, Aron-Dine et al. (2015) use a firm-level data and exploit the fact that annual coverage usually resets every January, and individuals that join the firm in different months in a year will face the same initial 'spot' price of health care but different expected end-of-year prices. They also maintain a strong assumption that 'individual have no private information about their health shocks'. Brot-Goldberg et al. (2017) also use a firm-level data and leverage a natural experiment when the firm requires its employees to switch from a free insurance plan to a nonlinear, high-deductible plan. They divide people into different cells by observed characters and calculate an individual's shadow price (expected end-of-year price conditional on current spending and health status) using observations within each cell. This practice implicitly assumes that individuals among the same cell are homogeneous. Moreover, they only use health status and age to construct the sextile.

This paper describes a new framework designed for studying a consumer's responsiveness to medical price. Specifically, we aim to investigate how a change of a shadow price would affect an individual's probability of doctor visiting. We will focus on an individual's outpatient medical consumption quantity, as inpatient consumptions are infrequent and are often associated with large expenditures that exceed the OPC easily (hence, decreasing future price to near zero). What we contribute to the existing studies is the usage of an episode-variant shadow price instead of the spot price or a fixed end-of-year price. Within this framework, we would model and compare an individual's doctor-visit decisions under a free plan and those under a cost-sharing plan. Our strategy requires less restrictive data structure and fewer model assumptions. Moreover, unlike previous literature that use static models, we are able to describe an individual's dynamic medical spending based on the shadow price that is conditional on a person's own year-to-date accumulative spending $X(t)$. For a given individual i, the shadow price at time t is defined as the conditional expected co-insurance rate given current cumulative individual cost $X_i(t)$:

$$P_i^s(t) = \mathbb{E}(R_{EOY} \mid X_i(t)) = R(X_i(t))$$

where $P_i^s(t), R_{EOY}$ are the shadow price at time t and the end-of-year co-insurance rate, respectively. $0 \leq R(X_i(t)) \leq 1$ with $R^{'} < 0$. The intuition behind this definition is simple: If $X_i(t)$ is under the deductible threshold, every health care consumption will lead to an increase of $X_i(t)$, making an individual cross the threshold more easily, thus, making the next purchase cheaper. The shadow price is therefore decreasing with every health care consumption. We then could use the stochastic cumulative individual spending $X(t)$ to study her responsiveness to the corresponding non-linearity budget constraint.

To the best our knowledge, Keeler et al. (1977) is the first theoretical paper to study a consumer's optimal choice under the non-linear medical price schedule. Using a dynamic programming model, they show that the shadow price of the $j-$th episode is a function of demands prior to this episode (hence the cumulative individual spending). The shadow price theory has profound implications on estimating medical demand. First, it suggests one should not use the nominal price, since the difference between the nominal price and the shadow price is not randomly generated. An incorrectly chosen nominal price would lead to a biased estimation. Second, as the shadow price is a function of the accumulative individual spending, individuals will make medical service utilization decisions in a sequential and contingent way. Figure 2 illustrates this situation.



Point A and B are the deductible threshold and OPC, respectively. When the total expenditure is below B, the price (the slope) $0 < R(X) < 1$ is a function of cumulative individual spending with $R' < 0$. Whenever the total expenditure is beyond B, there is no more cost for individuals.

FIGURE 2. Non-linear Individual Shadow Price

The adoption of the cumulative individual spending $X(t)$ creates econometric challenges. First, $X(t)$ is cumulated from the beginning of an insurance contract. Thus, conventional Markov-type dynamic mechanisms (i.e., p-th order lags, also known as the limited memory) would fail. Second, and perhaps most importantly, $X(t)$ is an individual-specific and history-dependent stochastic process. This feature implies that our model would depend on a stochastic filtration, and any likelihood-based modelling strategy would fail. Although the likelihood function is also conditional on individual-specific filtrations, these filtrations are deterministic. In addition, $X(t)$ is calculated as the summation of costs of each medical consumption, and hence would contain the information of N_i, the number of an individual i's doctor-visits within a pre-defined time interval. Ignoring the randomness of N_i by simply fixing it for all individuals would create a selection bias problem.

Our proposed framework is able to study such sequential and contingent consumption decisions and to overcome the mentioned econometric difficulties. This framework is based on a counting process called the Hawkes process. In this framework, the observation unit is a counting process in a form of the step function: it is piecewise constant with jumps (of size one) occurred at the time when a doctor visit happened. This data structure contains rich information. For example, fix arbitrary time t, the value of a counting process indicates the number of events occurred thus far. Meanwhile, the distance between two consecutive jumps is the doctor-visiting duration. When building econometric models of medical consumption, we would specify the intensity function instead of the likelihood function of a Hawkes process. The intensity function measures the instantaneous conditional probability of the occurrence of an event given all the history information. Furthermore, the intensity function completely characterize the corresponding Hawkes process.

The Hawkes process is state dependent (also known as self-exciting), i.e., some past events would affect the future ones. In our context, we are interested in two self-exciting channels. The first channel is the episode triggering effect, i.e., a doctor visit is caused by previous ones. A typical example is the recheck examination. The second channel is the shadow price effect, which is determined by the cumulation of past costs. We expect that with the shadow price decreasing, an individual would respond to a medical consumption more positively. Nevertheless, not all doctor visits are consequences of past experience. Some might occur independently. The Hawkes process could describe such a mixed self-exciting structure, and therefore, is an ideal tool to analyse the dynamic mechanism of an individual's doctor visits.

Due to this mixed self-exciting property, we would expect some doctor visits can form a cluster or a family where there is one independent initial event and several offspring events. Analyzing the cluster patterns is important for resource planning, allocation and the evaluation of the appropriateness, medical needs and efficiency of health care services (Hu et al., 2012). This paper will provide insights on how cost-sharing policies affect the number of clusters as well as the size of a cluster.

The paper is organized as follows. Section 2 describes the data and presents a preliminary result that suggests the existence of the state dependence. Section 3 discusses the specification of the model in detail, and Section 4 introduces the estimation method. In section 5, we report the main results and robustness checks. In section 6, we discuss the needs to use the proposed framework by investigating the probabilistic structure of the data. We also discuss the reason of not using a likelihood based estimation method. Lastly, section 7 concludes the paper.

2. Data and Some Preliminary Results

In this section, we introduce the data set and provide some descriptive results that indicate a sign of serial correlation among doctor visits.

2.1 *The Data*

The data come 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 cost-sharing, 25%, 50% or 95% coinsurance rates. The out-of-pocket cap 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 the nonlinear structure of our model, to ease the burden of computation, we only use data from Seattle, which has the largest medical claim records available. We separate the data by two different insurance plans: zero coinsurance rate plan (free plan, denoted as P0), in which a patient does not pay anything; and a cost-sharing plan (denoted as P95) in which a coinsurance rate of 95% is applied and the OPC is 150 USD per person or 450 USD per family[1] (i.e., before exceeding the OPC, individuals need to pay 95% of the medical care cost. Once the OPC is reached, all costs are paid by the insurance.). The OPC and the coinsurance rate in this plan only applied to ambulatory services; inpatient services were free. Both plans covered 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 preparing the dataset, we delete all records with missing time information. When analyzing the cost-sharing plan, we restrict our 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, we have 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, we fixed all ages at the enrollment

[1] In 1973 dollars.

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 in the sample; 2) if the value of a doctor visit cost is not available, we replace it with zero; 3) if information on the education is unknown, we replace it with the average education level.

2.2 *Preliminary Results on Cluster*

The RAND HIE data is widely studied, we believe there is little interest in providing another descriptive summary. Instead, we would present some preliminary results on the outpatient doctor visits cluster pattern. The cluster pattern are characterized by the number of clusters and the average number of events within a cluster.

Apart from the individual heterogeneity, doctor visits might be correlated due to some state dependent effects. The shadow price is one channel of such correlation if individuals do respond to its variation, another channel could be the triggering effect, i.e., past episodes might trigger the occurrence of future ones. One could observe doctor visit clusters in a free insurance plan if the later channel is indeed valid. We use the DBSCAN (Density-based spatial clustering of applications with noise) to nonparametrically analyze the cluster pattern. This is a cluster algorithm that is widely used in computer science and statistical learning (Ester et al., 1996). For this algorithm, there are two inputs: Eps, the radius of one dense region, and minPts, the minimum number of points required to form a dense region. For the purpose of DBSCAN clustering, points are classified as core points, border points or noise points. Core and border points form a cluster via different definitions of 'reachable'. Noise points are the points that do not belong to any cluster. We provide details of this algorithm and the definition of a cluster in the online Appendix. The algorithm's ability to identify 'noise' points is particularly appealing to us as some acute episodes are small in scale and only need one doctor visit to fully recover.

We set $Eps = 21$ days and as a rule of thumb[2] $minPts = 2$. For the purpose of comparison, we restrict the time horizon in both plans to the 1977-1978 insurance year. For each individual (both free plan and cost-sharing plan), we run the DBSCAN algorithm, document the number of clusters, the average number of instances per cluster and the number of noise points. For each insurance plan, we compute the average number of clusters per person, the average number of instances per cluster per person and the average noise points per person. Table 1 summarizes the results.

TABLE 1. Cluster Analysis

	avg cluster number	avg cluster members	avg noise points
free plan	1.2287	4.55187	1.62332
Cost-sharing plan	0.862595	3.3625	1.47328

[2]The rule of thumb is minPts = dimension +1

Since there is no shadow price effect in the free plan, the above results regarding the free plan indicate that the existence of the triggering effect is highly likely. The effects of cost-sharing policies on cluster structure are threefold. First, they reduce the average number of clusters per person. That means for the initial episode, the cost-sharing policies suppress the first doctor visiting behaviors. Second, within each cluster, they reduce the number of follow-up visits. Third, cost-sharing policies reduce the average number of noise points per person, i.e., they discourage individuals to use medical services when they have small episodes like minor injuries.

3. Econometric Model

This section will present our econometric model. We will first describe how to represent our data in a Hawkes process and the cluster structure of this process. For outpatient doctor visits, some of them would occur independently, while others might be correlated with previous episodes. This cluster structure can be summarised by Figure 3. In this particular realization of doctor visits, we have three clusters $\{T_1,...,T_6\},\{T_7,T_8\}$ and $\{T_9\}$. $\{T_1,T_7,T_9\}$ are the parent events (Gen_0) in the first, second and third cluster,

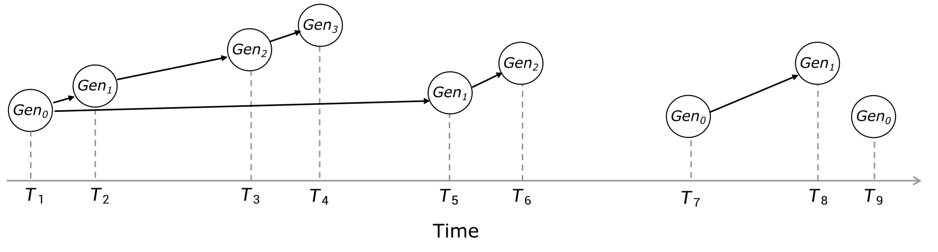


FIGURE 3. A possible cluster realization

respectively. Within each cluster, there might be more than one generations of children events.

Next, we focus on the model specification of the free insurance plan, where there is no individual expenditures, and the cluster structure is assumed to be the result of the triggering effect. Lastly, we present the model for the cost-sharing plan, where individual expenditures are introduced as marks. In terms of economic implication, we are interested in testing whether an individual would react to the change of these expenditures.

3.1 *Representing Data as the Hawkes Process and its Structure*

The Hawkes process is a special counting process. Fix an individual i, the counting process is defined as:

$$N_i(t) = \sum_{j=1}^{\infty} \mathbb{I}\{T_{ij} \leq t\}, \quad t_{i0} = 0$$

This is a step function with jumps happening in occurrence times of events. From Figure 4, one can conclude that the counting process contains rich information: It tells how

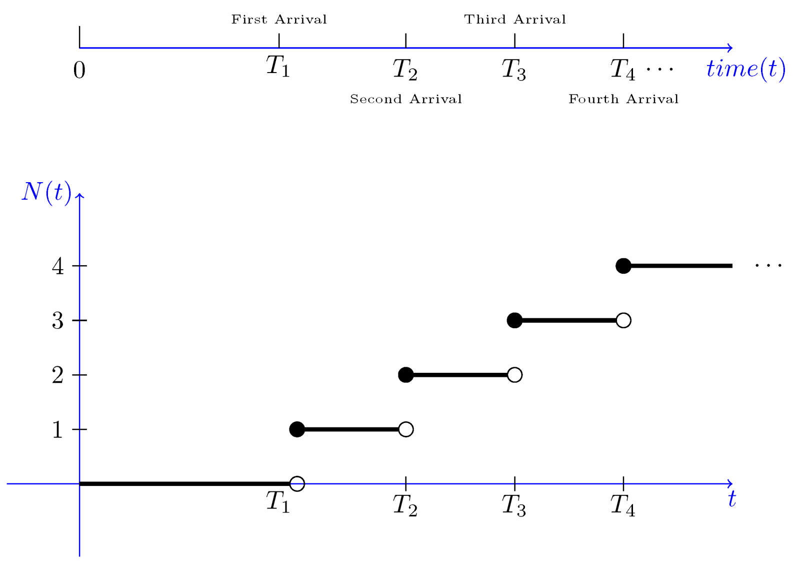


FIGURE 4. A possible counting process

many events has occurred so far, since the function value $N_i(t)$ is a count number by fixing a time t. It also completely describes the exact occurrence times of each event,

$$T_{ij} = \inf\{t \mid N_i(t) = j\} \quad j \geq 1$$

Throughout the paper, we always consider a simple counting process, i.e., there would be no common jumps at the same time and the jump size is always one.

For any sub-martingale, including the counting process, we have the following Doob-Meyer decomposition result:

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

or

$$dN_i(t) = \lambda_i(t)dt + dM_i(t)$$

where $\Lambda_i(t)$ is the cumulative intensity or the compensator (hence, $\lambda_i(t) = d\Lambda_i(t)/dt$ is the intensity function), and $M_i(t)$ is a (local) martingale with respect to the given filtration $\mathcal{F}_i(t)$, and hence, trend-free. $\Lambda_i(t)$ 'compensates' the monotonicity of $N_i(t)$. Importantly, $\Lambda_i(t)$ is predictable (a practical implication of predictability is that we know the value of $\Lambda_i(t)$ one step ahead of time) and thus may serve as a predictor of $N_i(t)$. From the probability point of view, an intensity function conditional on a filtration $\mathcal{F}_i(t-)$ measures the instantaneous conditional probability of the occurrence of an event:

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

where the filtration $\mathcal{F}_i(t-)$ contains information up to a time just before t. Specifically, $\lambda_i(t)$ completely characterize the corresponding counting process.

The choice of the filtration $\mathcal{F}_i(t-)$ is important in the context of the counting process analysis. In this paper, we are interested in the filtration that includes a sigma-filed generated by the process itself, i.e.,

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

By construction, the corresponding conditional intensity function is a history-dependent rate function. It is important to realize that this conditional intensity is a stochastic process itself, since it is depend on the event history, which is stochastic. In the literature of counting process analysis (see Kass et al. (2014)), a conditional intensity function that depends on history or on any other stochastic process is often called a conditional intensity process, and the resulting counting process is called a doubly stochastic counting process.

The Hawkes process (Hawkes, 1971) is a well known doubly stochastic process, where its conditional intensity is characterized as:

$$\lambda_i(t \mid \mathcal{F}(t-)) = \lambda_0 + \int_0^t g(t-s)dN_i(s) \tag{1}$$

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

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

$$g(t) = \alpha \exp(-\mu t)$$

The Hawkes process is originally proposed to study earthquakes but soon generalizes to other areas such as finance(Bowsher, 2007) and criminology(Mohler et al., 2012). The construction of this process involves a *branching* mechanism: first, a Poisson process with rate λ_0 independently generates *immigrants* (i.e., independent events). The first event always comes from this Poisson process. Second, a given immigrant event can give birth to subsequent *offspring* events, an offspring event might also be the ancestor of future generation offspring events.

The branching mechanism would generate clusters in a stationary Hawkes process, in which the corresponding intensity would eventually reaches a steady value λ^*:

$$\frac{d\mathbb{E}N(t)}{dt} = \lambda^* = \lambda_0 + \lambda^* \int_0^\infty g(t)dt$$

and hence,

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

The term $n^* = \int_0^\infty g(t)dt$ is called the branching ratio. Clearly, the above result is well defined only if $n^* < 1$.

Most literature would assume that the underlying Hawkes process is stationary. The stationary property would enable a researcher to use likelihood based estimation method. We too, would impose this stationary restriction. However, our motivation is not about the estimation but about the cluster pattern of a stationary Hawkes process.

The branching ratio n^* can be interpreted as the average number of offsprings 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 $\lambda^*(t)dt$ 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, we have $\lambda_0 dt$ families (clusters), and the expected size per family is $1/(1-n^*)$. Let A_i be the expected number of events in $Generation_i$, and $A_0 = 1$ (the parent immigrant). Then expected size of a cluster $N_\infty = 1/(1-n^*)$ can also be defined as:

$$N_\infty = \sum_{i \geq 1} A_i \tag{3}$$

Suppose the average number of offsprings per event is $\tilde{n}$, then we can find a inductive relationship $A_i = A_{i-1}\tilde{n}$. With $A_0 = 1$, we derive:

$$A_i = A_0 \left(\tilde{n}\right)^i = \left(\tilde{n}\right)^i, i \geq 1 \tag{4}$$

$$N_\infty = \frac{1}{1-\tilde{n}} \tag{5}$$

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.

To summarize, the corresponding counting process $N(t)$ can be decomposed as

$$N(t) = N^0(t) + N^1(t)$$

where $N^0(t)$ is a Poisson process with constant rate of λ_0, while $N^1(t)$ is a self-exciting process whose events are triggered by both the Poisson process and the past events of its own. This branching interpretation is well suited for our outpatient utilization application, in which some episodes would arrive independently, while the other events are descendants of existing episodes.

3.2 *A Model for the Free Insurance Plan*

As the cluster structure would depend on an individual's time-invariant heterogeneity, we need to control it to make the comparison between the free plan and the cost-sharing plan sensible. Thus, we begin our description of the model by first focusing on a canonical individual, whose time invariant covariates are normalized to ν. Suppose this individual signs a free insurance contract at time 0, and as long as the first event (doctor visit) has not occurred ($t < T_1$), we specify her intensity function as

$$\lambda(t) = \nu, \quad 0 \leq t < T_1 \tag{6}$$

When $t \geq T_1$, her intensity becomes

$$\lambda(t) = \nu + \int_0^t g(t-s)dN(s) \tag{7}$$

where $N(t)$ is the corresponding counting process. Notice that

$$\int_0^t g(t-s)dN(s) = \sum_{j:t_j<t} g(t-t_j)$$

In our case, we specify $g(t-t_j) = \alpha \exp(-\mu(t-t_j))$. This kernel describes the impact of the $j-th$ doctor visit. We expect that an individual is more likely to re-visit a doctor if the elapsed time from the previous doctor visiting is short, but this probability would gradually decrease as time goes by. Thus, we would expect that $\mu > 0$, and to ensure the non-negativity of an intensity function, we impose $\alpha > 0$. The stationary condition is:

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

Next, let's consider a specific individual i. When $t < T_{i1}$, we specify her intensity as:

$$\lambda_i(t) = \phi(z_i)\varepsilon_i\lambda(t) = \phi(z_i)\varepsilon_i\nu = \phi(z_i)\nu_i$$

where ε_i with $\mathbb{E}\varepsilon_i = 1$ is the idiosyncratic innovation. $\nu_i = \nu\varepsilon_i$ is the unobserved heterogeneity that may represent this individual's health status or a summary of the state dependent effect before time 0. z_i is a vector of observed individual covariates with $\phi(z_i) = \exp(z_i^\top \gamma)$. When $t \geq T_{i1}$, the intensity is written as:

$$\begin{aligned}\lambda_i(t) &= \phi(z_i)\left(\nu_i + \int_0^t g(t-s)dN_i(s)\right)\\ &= \phi(z_i)\left(\nu_i + \sum_{j:t_{ij}<t} g(t-t_{ij})\right)\end{aligned}$$

Here the unobserved heterogeneity ν_i is additively separated from the exciting function $\int_0^t g(g-s)dN_i(s)$. We restrict to this additive specification for two reasons: 1) for identification, we delay the identification discussion to later section, and 2) for a better distinguishing between the unobserved heterogeneity effect and the state dependent effect. Similar additive specification can been found in Kopperschmidt and Stute (2013). Notice that for the intensity function of individual i, the underlying process, conditional on the observed heterogeneity $\phi(z_i)$, may not be stationary.

To conclude, our intensity for the free insurance plan is specified as:

$$\lambda^{P0}(t) = \exp(z_i^\top \gamma)\nu_i, \quad t < T_1 \tag{8}$$

$$\lambda^{P0}(t) = \exp(z_i^\top \gamma)\nu_i + \exp(z_i^\top \gamma)\sum_{j:t_{ij}<t} \alpha \exp(-\mu(t-t_{ij})), \quad t \geq T_1 \tag{9}$$

3.3 *A Model for Cost Sharing Plan*

The major econometric difference between a free plan and a cost-sharing plan is the inclusion of an individual expenditure $X(t)$ as marks. $X(t)$ is a piecewise constant, non-decreasing stochastic process with multiple sources. Specifically, we note that the doctor visit fees are not the only component of $X(t)$. Another major source of individual expenditures is the drug purchase. Hence, $X(t)$ might be expressed as:

$$X(t) = \sum_{i=1}^{N(t-)} x_i + \sum_{i=1}^{N^1(t-)} y_i$$

where $N(t)$ is the doctor visit counting process with expenditure x_j for $j-th$ visit and $N^1(t)$ is the drug purchase counting process with expenditure y_j for $j-th$ drug purchase. The drug purchase process could be regarded as external shocks to the interested process.

With this in mind, we specify a canonical individual's intensity under a cost-sharing plan as:

$$\lambda(t) = b\nu, \quad t < T_1 \tag{10}$$

and

$$\lambda(t) = b\nu + b\int_0^t h(X(s))g(t-s)dN(s) \quad t \geq T_1 \tag{11}$$

where b is the cost-sharing effect. Like before, $g(t-t_j) = \alpha\exp(-\mu(t-t_j))$ with $\alpha, \mu > 0$. In addition, we specify $h(X(t)) = \exp(\beta_1 X(t))$. A few words on the above specification are in order. Recall that the shadow price is defined as $P_i^s(t) = \mathbb{E}(R_{EOY} \mid X(t)) = R(X(t))$ with $R' < 0$. An increase of $X(t)$ leads to a decreasing of the shadow price, affecting an individual's responsiveness to the medical consumption. The term $\exp(\beta_1 X(t_{ij}))$ aims to measure an individual's reaction to this shadow price change, and we would expect that $\beta_1 > 0$ (increasing the chance to visit a doctor) if individuals would respond to the shadow price and $\beta_1 = 0$ otherwise. The multiplicative structure of $\exp(\beta_1 X(t))\alpha exp(-\mu(t-t_{ij})$ reflects the assumption that individuals are partially influenced by the shadow price (Aron-Dine et al., 2015, Brot-Goldberg et al., 2017). In this specification, individuals would not behave fully rational or forward-looking, and would not consider the shadow price at all times. Rather, the shadow price have different impacts on an individual depending on the elapsed time from the previous episode. For a given occurrence time t_{ij}, we assume that the intensity would jump immediately, and gradually decrease to a plateau until the next event time. Thus, the above multiplicative specification is describing a case where an individual would fully consider the shadow price when she makes the medical utilization decisions, but would become more and more myopic as time goes by until the next medical spending.

The introduction of the time dependent stochastic process $h(X(t))$ creates a challenge for calculating the branching ratio n^* (and consequently, the cluster size N_∞). In the literature, researchers would assume marks are i.i.d and the branching ratio is the expectation over both the mark and the time (Rizoiu et al., 2017):

$$n^* = \int_0^\infty \int_A h(X(t))g(t)dtdF(x)$$

where A is a proper mark domain. This calculation, however, is not applicable to our application as the mark process $X(t)$ is state dependent.

One workaround could be the following. We partition the time line as:

$$[0, T] = \sum_{k=1}^{\kappa} I_k$$

where $I_k = [\tau_{k-1}, \tau_k)$, $\{\tau_k\}_{k=1,\ldots,\kappa}$ is a series of predetermined equispace time points with $\tau_0 = 0$ and $I_k \cap I_j = \emptyset, \forall k \neq j$. Within each interval I_k, we replace $h(X(t))$ with: $h(c_k)$, where $c_k = \min_{t\in I_k} X(t)$. Then, for a cluster that begins in I_k, its branching ratio is:

$$n^* = bh(c_k)\int_0^\infty g(t)dt$$

For example, we could let $h(c_1) = \exp(0) = 1$, i.e., in the initial period, the marks play little role, then the branching ratio is $n_1^* = b\int_0^\infty g(t)dt$. Since both c_k and n_k^* are non-decreasing, the corresponding cluster size $N_{k,\infty}$ is also non-decreasing and approaching to that of a free plan as $X(t)$ approaching to the OPC limit.

For a certain individual i, her intensity under a cost-sharing plan is then:

$$\lambda(t) = b\phi(z_i)\nu\varepsilon_i, \quad t < T_{i1} \tag{12}$$

and

$$\lambda(t) = \phi(z_i)b\left(\nu_i + \int_0^t h(X_i(s))g(t-s)dN_i(s)\right) \tag{13}$$

$$= \phi(z_i)b\left(\nu_i + \sum_{j:t_{ij}<t} h(X_i(t_{ij}))g(t-t_{ij})\right), \quad t \geq T_{i1} \tag{14}$$

As usual, $\nu_i = \nu\varepsilon_i$ with $\mathbb{E}\varepsilon_1 = 1$.

To conclude, our intensity function for the cost-sharing insurance plan is:

$$\lambda^{P95}(t) = b\exp(z_i^\top\gamma)\nu_i, \quad t < T_{i1} \tag{15}$$

$$\lambda^{P95}(t) = b\exp(z_i^\top\gamma)\nu_i + b\exp(z_i^\top\gamma)\sum_{j:t_{ij}<t}\exp(\beta_1 X(t_{ij}))\alpha\exp(-\mu(t-t_{ij})), \quad t \geq T_{i1} \tag{16}$$

4. Estimating and Identifying the Model

4.1 *The Minimum Distance Estimation*

We use a minimum distance method first proposed by Kopperschmidt and Stute (2013) to estimate the model. This method starts from a functional data analysis perspective where each (random) function comes from a counting process with possibly complicated dynamics. The basic idea consists of minimizing the distance between a counting process and its compensator (the Doob-Meyer decomposition). Intuitively, note that $N_i(0) = 0, \forall i$ and a counting process as well as its compensator only takes non negative values, we have:

$$\mathbb{E}(M_i(0)) = 0$$

and

$$\mathbb{E}(N_1(t) \mid \mathcal{F}_i(t-)) = \mathbb{E}(\Lambda_1(t) \mid \mathcal{F}_i(t-))$$

where the expectation is taken over the individual. One advantage of this method is that it does not require the differentiability of the compensator, thus allows unexpected jumps in the intensity function. This is particularly useful in our application, as an individual's expenditure $X(t)$ have two stochastic sources. For the purpose of self-containing, we briefly summarize the results here and a more rigorous discussion of this estimation method can be found in their original paper.

Let $N_1, ..., N_n$ be i.i.d copies of n observed counting processes that are conditional on the increasing filtration $\mathcal{F}_i(t), 1 \leq i \leq n$, which are comprised by the counting process N_i as well as some other external information. Let $\Lambda_{\theta,i}(t|\mathcal{F}_i(t-))$ with $\theta \in \Theta \subset \mathbb{R}^d$ be a given class of parametric compensators. We set,

$$< f, g >_\mu = \int_0^T f(s)g(s)d\mu(s)$$

where T is the terminating time. If f and g are square-integrable functions w.r.t. the measure μ, the corresponding semi-norm is,

$$||f||_\mu = [< f, f >_\mu]^{1/2}$$

Let,

$$\bar{N}_n = \frac{1}{n}\sum_{i=1}^{n} N_i; \bar{\Lambda}_{\theta,n} = \frac{1}{n}\sum_{i=1}^{n} \Lambda_{\theta,i} \tag{17}$$

We call the former the averaged counting process and the later the averaged compensator. Naturally the associated averaged innovation martingale is,

$$\bar{M}_n = \bar{N}_n - \bar{\Lambda}_{\theta_0,n}$$

If we take $\mu = \bar{N}_n$, the quantity $||\bar{N}_n - \bar{\Lambda}_{v,n}||_{\bar{N}_n}$ is then an overall measurement of fitness of $\bar{\Lambda}_{\theta,n}$ to $\bar{N}_n$. The estimator θ_n is computed as,

$$\theta_n = arg \inf_{\theta \in \Theta} ||\bar{N}_n - \bar{\Lambda}_{\theta,n}||_{\bar{N}_n} \tag{18}$$

Kopperschmidt and Stute (2013) has shown that this minimum distance estimator is consistent and asymptotically normal. Specifically, for the consistency result, let $\Theta \in \mathbb{R}^d$ be a bounded open set and for each $\epsilon > 0$. Assume

$$\inf_{||\theta - \theta_0|| \geq \epsilon} ||\mathbb{E}\Lambda_{\theta_0} - \mathbb{E}\Lambda_\theta||_{\mathbb{E}\Lambda_{\theta_0}} > 0 \tag{19}$$

$$\textit{The process}(t, \theta) \rightarrow \Lambda_\theta(t) \textit{ is continuous in } t \in [0, T) \textit{ with probability one} \tag{20}$$

then

$$\lim_{n \to \infty} \theta_n = \theta_0 \textit{ with probability one} \tag{21}$$

The first condition is a weak identification condition, while the second condition guarantees continuity (but not differentiability) of Λ_θ in t and allows for unexpected jumps in the intensity function λ_θ as well.

For the asymptotic normality result, let

$$\Phi_0(\theta) = \frac{\partial}{\partial\theta}\int_E (\mathbb{E}\Lambda_\theta(t) - \mathbb{E}\Lambda_{\theta_0}(t))\mathbb{E}\frac{\partial}{\partial\theta}\Lambda_\theta(t)^T\mathbb{E}\Lambda_{\theta_0}(dt)$$

be a matrix-valued function, where T denotes transposition, $E = [\underline{t}, \bar{t}]$. And suppose (19) and (20) hold, furthermore, assume that

$$\left\| \frac{\partial}{\partial\theta}\left(\mathbb{E}\Lambda_\theta(t) - \mathbb{E}\Lambda_{\theta_0}(t)\right)\mathbb{E}\frac{\partial}{\partial\theta}\Lambda_\theta(t)^T \right\| \leq C(t)$$

for all θ in a neighborhood of θ_0, the function C is integrable w.r.t $\mathbb{E}\Lambda_{\theta_0}$, and

$$\phi(x) = \int_{[x,\bar{t}]}\mathbb{E}\frac{\partial}{\partial\theta}\Lambda_\theta(t)\mathbb{E}\Lambda_{\theta_0}(dt)\mid_{\theta=\theta_0}, \underline{t} \leq x \leq \bar{t}$$

is square integrable w.r.t. $\mathbb{E}\Lambda_{\theta_0}$. Then as $n \to \infty$

$$\sqrt{n}\Phi_0(\theta_0)(\theta_n - \theta_0) \to N(0, C(\theta_0)) \tag{22}$$

where $C(\theta_0)$ is a $d \times d$ matrix with entries

$$C_{ij}(\theta_0) = \int_E \phi_i(x)\phi_j(x)\mathbb{E}\Lambda_{\theta_0}(dx)$$

Let Φ_n be the empirical analogue of Φ_0,

$$\Phi_n(\theta) = \frac{\partial}{\partial\theta}\int_E (\bar{\Lambda}_{\theta,n}(t) - \bar{\Lambda}_{\theta_0,n}(t))\frac{\partial}{\partial\theta}\bar{\Lambda}_{\theta,n}(t)^T\bar{\Lambda}_{\theta_0,n}(dt) \tag{23}$$

Since all $\bar{\Lambda}_{\theta,n}$ are sample means of i.i.d non-decreasing processes, a Glivenko-Cantelli argument yields, with probability one, uniform convergence of $\bar{\Lambda}_{\theta,n} \to \mathbb{E}\Lambda_\theta(t)$ in each t on compact subsets of Θ. We have the expansion,

$$\Phi_n(\theta) = \Phi_0(\theta) + op(1) \tag{24}$$

Such expansion guarantees that in a finite sample situation, we can replace the unknown matrix $\Phi_0(\theta_0)$ by $\Phi_n(\theta_n)$ and $C(\theta_0)$ by $C^n(\theta_n)$ without destroying the distributional approximation through $N(0, C(\theta_0))$, where C^n is the sample analog of C. In practice, one need to plug and replace the true ones with estimators and replace $\mathbb{E}\Lambda_{\theta_0}(dt)$ with its empirical counterpart $\bar{N}(dt)$.

Kopperschmidt and Stute (2013) did not provide a simulation study in their paper. We therefore analyse the finite sample performance of this estimation method, and the results can be found in Appendix A.

4.2 *Identifying the Model*

Since the seminal works of Heckman (1978, 2007), an important part of the analysis is to discover the extent to which dynamic is due to the true state dependence or to the unobserved individual heterogeneity. Such an analysis is obviously relevant to our work. To streamline the presentation, we discuss the model identification based on the free plan's intensity function. A similar argument could apply to the cost-sharing plan's intensity effortlessly.

Recall that our estimation method is based on the Doob-Meyer decomposition result, i.e., the objective function is the distance between a counting process and its compensator, conditional on a time varying filtration:

$$||\bar{N}_n - \bar{\Lambda}_{\theta,n}||_{\bar{N}_n}$$

Note that in our specification

$$\mathbb{E}\bar{\Lambda}_{\theta,n}(t) = \mathbb{E}\Lambda_{1,\theta}(t) = ut + \mathbb{E}\Big[\exp(z_1^\top\gamma)\sum_{j:t_{1j}<t}\Big(1 - \frac{\alpha}{\mu}\exp(-\mu(t-t_{1j}))\Big)\Big]$$

where $u = \mathbb{E}\exp(z_1^\top\gamma)\nu_1 > 0$. We could write down its empirical counterpart as:

$$\tilde{\Lambda}_{\theta,n} = ut + \frac{1}{n}\sum_{i=1}^{n}\exp(z_i^\top\gamma)\sum_{j:t_{ij}<t}\Big(1 - \frac{\alpha}{\mu}\exp(-\mu(t-t_{ij})\Big)$$

Since $\phi(z_i)\nu_i = u + \eta_i$ with $\mathbb{E}\eta_1 = 0$ and η_i is orthogonal to z_i, ν_i, we have

$$\tilde{\Lambda}_{\theta,n} - \bar{\Lambda}_{\theta,n} = \frac{1}{n}\sum_{i=1}^{n}\eta_i = o_p(1) \tag{25}$$

To conclude, because of our model specification, especially the additive structure between the unobserved heterogeneity and the self-exciting function, we are able to identify the expectation of individual covariates, the observed covariates and the self-exciting function. When performing the estimation, one would replace $\bar{\Lambda}_{\theta,n}$ by $\tilde{\Lambda}_{\theta,n}$. In practice, in order to ensure that u is strictly positive, we would write $u = \exp(k)$.

5. Main Results and Robustness Check

5.1 *Main Results*

Like Keeler and Rolph (1988), we assume there are no interactions between the shadow price effect and the effects of other explanatory variables. Thus, we might use data from the free plan to estimate μ, $\phi(z)$ and $\exp(k)$ and plug the estimators into the cost-sharing plan. Table 2 summarizes the results. The shadow price effect is captured by $\exp(\beta_1 X(t))$. We observe that β_1 is positively away from zero and conclude that individuals do understand the design of the insurance policy and take advantage of the shadow price.

The cluster pattern (of the canonical individual) is described by α, μ. We perform a Wald test on the null $H_0 : \alpha = \mu = 0$ against $H_1 : \alpha \neq 0, \mu \neq 0$. The corresponding Wald statistics is 500.015689, clearly rejecting the null. Therefore, we could conclude that for a canonical individual who signs into a free insurance plan, the average number of doctor visits in a cluster (i.e., the cluster size) is approximately $N_{\infty}^{P0} = \hat{\mu}/(\hat{\mu} - \hat{\alpha}) \approx 5.8$. As mentioned before, it is difficult to estimate the cluster size for the same canonical individual who joins a cost-sharing plan. However, by treating the mark process as a piecewise constant, we might approximate the size of a cluster whose parent event occurred in the beginning period as $N_{\infty}^{P95} = \hat{\mu}/(\hat{\mu} - \hat{\alpha}\hat{b}) \approx 2.2$. Thus, the size of this cluster shrinks $(5.8 - 2.2)/5.8 = 62\%$. Since a cluster can only have one independent doctor visit, we could use the intensity of the Poisson process to measure the cluster number. Our result suggests that the cluster number in the cost-sharing plan decreases $(1 - b) = 34\%$ comparing with that of the free plan.

In the explanatory variable vector, we include age, gender, education (in terms of years) and log-income. The interpretation of the corresponding coefficients is not straightforward due to the nonlinear nature. However, we may fix a time period and treat the counting process as count data. The interpretation is then identical to that of the marginal effect at a representative value (MER) of a count data regression model. Let $Y_t = N(t)$ be the number of events occurred before time t. Let scalar z_j denote the $j - th$ covariate. Differentiating

$$\frac{\partial \mathbb{E}(Y_t|Z)}{\partial z_j} = \gamma_j \mathbb{E}(\Lambda(t|Z))$$

by the exponential structure of $\phi(z)$. For example, if $\hat{\gamma_j} = 0.2$, $\bar{\Lambda}_n(t|Z) = 2.5$, then one-unit change in the $j - th$ covariate increases the expectation of Y_t by 0.5 units. With this in mind, we can interpret our results.

- *Age.* At first, intensity values will decrease as age increases. After one passes the age of 41, 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 go the doctor.
- *Education.* The result, by and large, suggests a negative relation between education and the outpatient medical utilization. One explanation is that higher education often associates with a healthier life style, which reduces the hazard rate of visiting a doctor.
- *Income.* Income is positively related to the use of medical services, which is no surprising. A higher income gives individuals the ability to cover the opportunity cost related to the absence from work.

TABLE 2. Basic Results

	Estimator	*Description*
α	17.250964 (24.027315)	
μ	20.861092 (21.182576)	
age	-0.359284 (0.004021)	
age2	0.435267 (0.005364)	$(age)^2/100$
male	-3.599054 (0.053921)	
edu	-1.251602 (0.011095)	
edu2	3.83581 (0.03207)	$(edu)^2/100$
log income	1.694981 (0.014325)	
k	-0.40969 (0.020022)	$\exp(k)$ is the expectation of individual's heterogeneity
b	0.659005 (0.008822)	
β_1	0.002631 (0.00003)	coefficient of X(t)

5.2 *Robustness Check*

5.2.1 *Permanent Shadow Price Setting* Our cost-sharing plan model assumes an individual would react partially to the shadow price. Here in this robustness exercise, we assume the shadow price effect enters the intensity additively. This setting implicitly assumes that an individual would consider the shadow price all the time. Specifically, we assume that:

$$\lambda_i^{P95}(t) = \exp(z_i^\top \gamma)\tilde{a}_i(t), \quad t \geq T_1$$

where

$$\tilde{a}_i(t) = b\left(\nu_i + \sum_{j=1}^{N_i(t-)} \alpha exp(-\mu(t-t_{ij})) + \exp(\beta_1 X_i(t))\right)$$

Table 3 summarizes our estimation result for this specification. In this setting, we still

TABLE 3. Robustness Check Results

	Estimator	*Description*
b	0.670863 (0.022799)	
β_1	0.014421 (0.000121)	coefficient of X(t)

find evidence suggesting that individuals do respond to the shadow price.

5.2.2 *Non-Stationary Intensity Setting* In the main model, we assume that the canonical individual who enters the free insurance plan would have a stationary Hawkes intensity function, i.e., we impose $\alpha < \mu$ or $n^* < 1$. The economical interpretation behind such a restriction is that the cluster size is bounded and the cluster would die out almost surely as time elapsed. Here, we investigate another scenario where $n^* = 1$ or $\alpha = \mu$.

In the literature, such a setting is often referred as the critical regime (Bowsher, 2001). It corresponds to a situation where one cluster lives indefinitely but without exploding. In the context of our application, it implies that all doctor visits belong to one family, and there would be one parent event (possibly before our investigation time) that permanently changed the health status of an individual. In terms of the model specification, this restriction would require $\nu = 0$, i.e., there is no unobserved heterogeneity, and all the heterogeneous outpatient utilization are from the state dependent effect.

Specifically, for the free insurance plan, the intensity function for an individual would be:

$$\lambda_i^{P0}(t) = \exp(z_i^\top \gamma) \sum_{j:t_{ij}<t} \mu \exp(\mu(t-t_{ij}))$$

For the cost-sharing plan, the intensity is:

$$\lambda_i^{P95}(t) = b\exp(z_i^\top \gamma) \sum_{j:t_{ij}<t} \exp(\beta_1 X_i(t_{ij}))\mu\exp(\mu(t-t_{ij}))$$

Table 4 summarizes the results. We still find that individuals would respond to the shadow price, and to some degree, understand the nature of a non-linear contract.

TABLE 4. Critical Regime Results

	Estimator	*Description*
μ	27.871263 (8.398143)	
age	-0.133948 (0.031222)	
age2	0.154709 (0.042255)	$(age)^2/100$
male	-0.717039 (0.473840)	
edu	-0.354950 (0.085860)	
edu2	0.994284 (0.336120)	$(edu)^2/100$
log income	0.592655 (0.036435)	
b	0.658995 (0.041568)	
β_1	0.003889 (0.000320)	coefficient of X(t)

6. DISCUSSION

In this section, we would discuss the need to use the counting process framework from an econometric point of view. In addition, we would discuss reasons of adopting the minimum distance estimation method.

One distinct property of our data is that for a fixed time interval, say $(0, T]$, not only the occurrence times $\{t_{ij}\}$ varies, but also the number of doctor visits $N_i(T)$ varies significantly across individuals. Apart from time-invariant individual heterogeneities, state dependent effects (the triggering effect and the shadow price effect) are important sources for this variation. We investigate the probabilistic structure of our doctor visits data to get a better understanding.

To begin with, we represent a counting process $N_i(t)$ defined in the time interval $(0, T]$ in terms of the occurrence times $(T_{i1}, \dots, T_{i(N_i)})$, where $N_i = N_i(T)$ is the random variable representing the number of doctor visits. The joint density of these occurrence times are

$$f_{T_{i1},\ldots,T_{i(N_i)}}(t_{i1},\ldots,t_{i(N_i)}) = f_{T_{i1}}(t_{i1})f_{T_{i2}}(t_{i2} \mid T_{i1}=t_{i1})\cdots f_{T_{i(N_i)}}(t_{i(N_i)} \mid T_{i(N_i-1)}=t_{i(N_i-1)})$$
$$\times P(N_i(T)-N_i(T_{i(N_i)})=0)$$

To analysis this joint density, let's first study the conditional density function of T_{ij}. Note that $\{T_{ij} > t_{ij} \mid T_{i(j-1)} = t_{i(j-1)}\}$ is equivalent to there being no events in the interval $(t_{i(j-1)}, T_{ij}]$. We could construct a n-th partition of that interval by setting $\Delta t = (t_{ij} - t_{i(j-1)})/n$, and letting $\tau_k = t_{i(j-1)} + k\Delta t$. The probability of observing zero events in the larger interval is equivalent to the probability of observing no events in each of the partition intervals,

$$\begin{aligned}\Pr\left(\Delta N_i(t_{i(j-1)},t_{ij}])=0\right) &= \Pr\left(\Delta N_i\left(\tau_0,\tau_1\right]=0,\ldots,\Delta N_i\left(\tau_{n-1},\tau_n\right]=0\right)\\ &= \Pr\left(\Delta N_i\left(\tau_{n-1},\tau_n\right]=0 \mid \mathcal{F}_{n-1}\right)\cdots\Pr\left(\Delta N_i\left(\tau_0,\tau_1\right]=0 \mid \mathcal{F}_0\right).\end{aligned}$$

where $\Delta N_i((a,b])$ is the counting process increment in the interval $(a,b]$. By the definition of the intensity, it is easy to show that each of these small history dependent increments takes on the value 0 with probability $1-\lambda_i(\tau_k \mid \mathcal{F}_k)\Delta t$. Therefore,

$$\begin{aligned}\Pr\left(\Delta N_i(t_{i(j-1)},t_{ij}])=0\right) &= \lim_{\Delta t\to 0}\prod_k\left(1-\lambda_i\left(\tau_k \mid \mathcal{F}_k\right)\Delta t\right)\\ &= \lim_{\Delta t\to 0}\prod_k\left(\exp\left(-\lambda_i\left(\tau_k \mid \mathcal{F}_k\right)\Delta t\right)+o(\Delta t)\right)\\ &= \lim_{\Delta t\to 0}\exp\left(-\sum_k\lambda_i\left(\tau_k \mid \mathcal{F}_k\right)\Delta t\right)+o(\Delta t)\\ &= \exp\left(-\int_{t_{i(j-1)}}^{t_{ij}}\lambda_i(t \mid \mathcal{F}(t-))dt\right),\end{aligned}$$

where the limit of the sum in the exponential term is the Riemann integral of the conditional intensity function. Hence,

$$P\{T_{ij} > t_{ij} \mid T_{i(j-1)} = t_{i(j-1)}\} = \exp\left(-\int_{t_{i(j-1)}}^{t_{ij}}\lambda_i(t \mid \mathcal{F}(t-))dt\right)$$

and its conditional p.d.f is then:

$$f_{T_{ij}}\left(t_{ij} \mid T_{i(j-1)} = t_{i(j-1)}\right) = \lambda_i\left(t_{ij} \mid \mathcal{F}_i(t_{ij}-)\right)\exp\left\{-\int_{t_{i(j-1)}}^{t_{ij}}\lambda_i(t \mid \mathcal{F}_i(t-))dt\right\}$$

Also note that

$$P(N_i(T)-N_i(T_{i(N_i)})=0) = \exp\left(-\int_{t_{i(N(T)-1)}}^{T}\lambda_i(t \mid \mathcal{F}(t-))dt\right)$$

thus,

$$f_{T_{i1},\dots,T_{i(N_i)}}(t_{i1},\dots,t_{i(N_i)}) = \exp\left(-\int_0^T \lambda_i(t \mid \mathcal{F}(t-))dt\right)\prod_{j=1}^{N_i}\lambda_i(t_{ij} \mid \mathcal{F}_i(t_{ij}-)) \qquad (26)$$

In this situation, the joint p.d.f includes two sources of randomness: one due to the variability described by the p.d.f, and the second comes the conditional intensity process $\lambda_i(t \mid \mathcal{F}(t-))$ (or equivalently, the randomness of $\mathcal{F}_i(t-)$). Notice here the random variable $N_i = N_i(T)$ is included in the filtration since $\sigma(N_i(s) : s < T) = \sigma(t_{i1},\dots,t_{i(N_i)}, N_i(T-) = N_i) \subseteq \mathcal{F}_i(T-)$. As mentioned before, if the conditional intensity function depends on history or on any stochastic process, the resulting counting process is called a doubly stochastic process.

In our application, the filtration $\mathcal{F}_i(t-)$ not only include an individual's doctor-visit history, but also contains this individual's cumulative spending $X_i(t-)$, which is in fact a stochastic process. Thus, the underlying data generating process is a doubly stochastic process, and a proper econometric model should fully reflect this probabilistic structure. Conventional dynamic econometric frameworks, however, often fail to recognize the randomness of $N_i(T)$. For example, although we could represent the doctor visit records in terms of a panel of durations $d_{ij} = t_{ij} - t_{i(j-1)}$, the classical dynamic panel data models can only be applied to a balanced panel data. This balanced data structure discards the randomness of $N_i(T)$ and implicitly imposes a sample selection mechanism. The dynamic duration model of Heckman and Walker (1990) does not consider this randomness neither. They constructed a likelihood function based on the joint density of k complete durations $(D_{i1},\dots,D_{ik})$ and a $k+1$st incomplete duration $\tilde{D}_{i(k+1)}$.

The counting process framework, on the other hand, captures the variation of $N_i(T)$. Note that the conditional intensity function $\lambda_i(t \mid \mathcal{F}_i(t-))$ has already appeared in the joint density function of the process. While the expectation of the random variable N_i is nothing but $\mathbb{E}N_i(T) = \Lambda_i(T \mid \mathcal{F}_i(T-))$. In fact, the conditional intensity function would uniquely characterize the probability structure of a counting process, see Proposition 7.2.IV Daley and Vere-Jones (2003).

Lastly, we discuss the estimation method. The doubly stochastic property also brings challenges to building a likelihood function. To see it, we need to represent the probabilistic structure of the counting process in terms of its finite dimensional distributions (or fidis). The basic idea consists of 'projecting' the process onto a finite-dimensional vector space, and study its variation. This amounts to first fix the number of event n in the time interval $[0, T)$, and study the corresponding joint density given n. Essentially, by projection, we eliminate one source of randomness, and to fully describe the probability structure of the process, we need enumerate $n = 1, 2, \dots,$.

Denote by $f_T^{(j)}(t)$ the joint probability density for the first j event times. The joint density of a counting process is then:

$$f_{T_{i1}}^{(1)}(t_{i1}) = \lambda_i(t_{i1}) \exp\left[-\int_0^{t_{i1}} \lambda_i(t)dt\right]$$

for $j = 1$ and

$$\begin{aligned} f_T^{(j)}(t) = \lambda_i(t_{i1}) &\left[\prod_{k=2}^{j} \lambda_i(t_{ik} \mid N_i(t_{ik}-) = k-1, t_{i1}, \ldots, t_{i(k-1)})\right] \\ &\times \exp\left[-\int_0^{t_{i1}} \lambda_i(t)dt - \sum_{k=2}^{j} \int_{t_{i(k-1)}}^{t_{ik}} \lambda_i(t \mid \tilde{N}_i(t_{ik}-) = k-1, t_{i1}, \ldots, t_{i(k-1)})dt\right] \end{aligned}$$

for $j \geq 2$ and $0 \leq t_{i1} < \ldots = t_{ij}$. See Snyder and Miller (2012) for detailed proof.

A straightforward replacement of the stochastic filtration $\mathcal{F}_i(T-)$ by its realizations $\{t_{i1}, \ldots, t_{i(n_i)}, N_i = n_i\}$ in Equation 26 yields

$$f_{T_{i1},\ldots,T_{i(n_i)}}\left(t_{i1}, \ldots, t_{i(n_i)}\right) = P\left(N_i(T) = n_i \mid t_{i1}, \ldots, t_{i(n_i)}\right) f_T^{(n_i)}(t) \tag{27}$$

This joint density function consists of 1) the projected distribution of the process onto a n-dimensional vector space and 2) the probability of observing exactly n events in the time interval $[0, T)$. Heckman and Walker (1990)[3] use this likelihood contributor to study the relationship between household income and the timing and spacing of births. Implicitly, they have assumed that the unobserved fecundity is a sufficient statistic for the variation of n_i.This assumption might be reasonable for fertility applications, but it certainly can not be extended to other applications without justification. One direct consequence of using this likelihood contributor is the problem of sample selection, as all the individuals who has $N_i \neq n_i$ would be deleted from the dataset.

7. Conclusion

In this paper, we provide a model that could describe the dynamic behavior of an individual's outpatient consumption under different health insurance plans. In our framework, the unit of an observation is a Hawkes process, which is a counting process whose filtration is generated by the process itself. It allows researchers to take historical information into the model. In addition, its filtration could include external shocks. The Hawkes process can also describe an individual's doctor visits cluster patterns. A typical cluster consists of an independent doctor visit and follow up visits that are offsprings to this visit.

[3]Their model also considered the unobserved heterogeneity term, and use nonparametric maximum likelihood estimation method to obtain the estimators.

We specify the Hawkes process via its intensity function, and apply the model to the classical RAND Health Insurance Experiment data. The estimation results suggest that individuals do take the shadow price into consideration when making their spending decisions. Our finding is consistent with recent literatures on the spending effects of nonlinear health insurance contracts (Aron-Dine et al., 2015, Einav et al., 2015). Studying the cluster patterns, we found that for a canonical individual, if she joins a cost-sharing plan (instead of a free plan), during the initial period, the number of clusters would decrease 34%, and the cluster size would shrink 62%.

Lastly, we justify the need to use our framework from an econometric perspective. A path dependent dynamic model of doctor visits has the feature of doubly stochastic, i.e., the variation amongst the data has two sources. One comes from the specified density, and the other is the randomness of the filtration. Importantly, the second randomness includes the variation of the number of doctor visits in a fixed time interval. Conventional econometric tools often discard the second variation and lead to a sample selection problem. Our framework, however, depends on the intensity function, which uniquely determines the probability structure of a counting process, and thus avoids the sample selection problem.

Appendix A: Simulation Studies

A.1 *Data Generating Process and Simulation Method*

We demonstrate the performance of the minimum distance method by using 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 both the earthquake times and magnitudes as the marks. The intensity of a ETAS model, for its simplest form, could be:

$$\lambda_g(t|\mathcal{F}_{t-}) = \mu + \sum_{i:t_i<t} e^{\alpha x_i}\left(1+\frac{t-t_i}{c}\right)^{-1} \tag{28}$$

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

$$f(x|t,\mathcal{F}_{t-}) = \delta e^{-\delta x} \tag{29}$$

The above data generating process can be simulated using the R package 'PtProcess' (Harte, 2010).[4] The *thinning method* is used to generate the data. This method was first introduced by Lewis and Gerald (1979), Ogata (1981). The procedure consists of

[4] https://cran.r-project.org/package=PtProcess

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 $\lambda_g(t|\mathcal{F}_{t-})$ in the interval as
$$\lambda_{max} = \max_{t\in(\tau,\tau+\delta)} \lambda_g(t|\mathcal{F}_{t-})$$
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
$$U \le \frac{\lambda_g(\tau+\xi|\mathcal{F}_{t-})}{\lambda_{max}}$$
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

A.2 *Simulation Results*

We 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 $[0, 3000]$, $[0, 500]$ and $[0, 100]$. For each simulation setting, we run $B = 1000$ repeats. We 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 rate gets 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.)

TABLE A.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 A.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 A.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%

References

Aron-Dine, A., Einav, L., & Finkelstein, A. (2013). The RAND health insurance experiment, three decades later. *Journal of Economic Perspectives*, 27(1), 197-222. [2, 3]

Aron-Dine, A., Einav, L., Finkelstein, A., & Cullen, M. (2015). Moral hazard in health insurance: do dynamic incentives matter?. *Review of Economics and Statistics*, 97(4), 725-741. [2, 3, 14, 26]

Bowsher, C. G. (2007). "Modelling security market events in continuous time: Intensity based, multivariate point process models." Journal of Econometrics, 141(2), 876-912. [10]

Brémaud, P., & Massoulié, L. (2001). "Hawkes branching point processes without ancestors." *Journal of applied probability*, 38(1), 122-135. [21]

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. [2, 3, 14]

Cutler, David M and Zeckhauser, Richard J (2000), "The anatomy of health insurance." *Handbook of health economics*, 563-643, Elsevier. [2]

Daley, D. J., & Vere-Jones, D. (2003). "An introduction to the theory of point processes: volume I: elementary theory and methods." Springer New York. [24]

Einav, L., Finkelstein, A., & Schrimpf, P. (2015). The response of drug expenditure to nonlinear contract design: Evidence from Medicare Part D. *The quarterly journal of economics*, 130(2), 841-899. [3, 26]

Embrechts, P., Liniger, T., & Lin, L. (2011). "Multivariate Hawkes processes: an application to financial data."*Journal of Applied Probability*, 48(A), 367-378. [10]

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). "A density-based algorithm for discovering clusters in large spatial databases with noise." In *kdd* (Vol. 96, No. 34, pp. 226-231). [7]

Hawkes, A. G. (1971). "Spectra of some self-exciting and mutually exciting point processes." *Biometrika*, 58(1), 83-90. [10]

Heckman, J. J. (1978). "Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence." In *Annales de l'INSEE* (pp. 227-269). Institut national de la statistique et des études économiques. [18]

Heckman, J. J. (2007). "3. Heterogeneity and State Dependence." In *Studies in labor markets* (pp. 91-140). University of Chicago Press. [18]

Heckman, J. J., & Walker, J. R. (1990). "The relationship between wages and income and the timing and spacing of births: Evidence from Swedish longitudinal data." *Econometrica: journal of the Econometric Society*, 1411-1441. [24, 25]

Harte, David (2010). "PtProcess: An R package for modelling marked point processes indexed by time." In *Journal of Statistical Software* Vol.35, 1-32. American Medical Informatics Association. [26]

Hu, J., Wang, F., Sun, J., Sorrentino, R., & Ebadollahi, S. (2012). "A healthcare utilization analysis framework for hot spotting and contextual anomaly detection." In *AMIA annual symposium proceedings* (Vol. 2012, p. 360). American Medical Informatics Association. [5]

Kass, R. E., Eden, U. T., & Brown, E. N. (2014). "Analysis of neural data." Springer, New York. [10]

Keeler, E. B., Newhouse, J. P., & Phelps, C. E. (1977). "Deductibles and the demand for medical care services: The theory of a consumer facing a variable price schedule under uncertainty." *Econometrica*: Journal of the Econometric Society, 641-655. [4]

Keeler, E. B., & Rolph, J. E. (1988). "The demand for episodes of treatment in the health insurance experiment." *Journal of health economics*, 7(4), 337-367. [2, 18]

Kopperschmidt, K., & Stute, W. (2013). "The statistical analysis of self-exciting point processes." *Statistica Sinica*, 1273-1298. [13, 15, 16, 17]

Lewis, Peter A and Shedler, Gerald S (1979). "Simulation of nonhomogeneous Poisson processes by thinning." *Naval Research Logistics Quarterly*, Vol. 26, 403-413. [26]

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., & Leibowitz, A. (1987). "Health insurance and the demand for medical care: evidence from a randomized experiment." *The American economic review*, 251-277. [2]

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). "Self-exciting point process modeling of crime." *Journal of the American Statistical Association*, 106(493), 100-108. [10]

Ogata, Yosihiko and Katsura, Koichi (1988). "Likelihood analysis of spatial inhomogeneity for marked point patterns." *Annals of the Institute of Statistical Mathematics*, 40(1), 29-39. [26]

Ogata, Yosihiko (1981). "On Lewis' simulation method for point processes." *IEEE Transactions on Information Theory*, 27(1), 23-31. [26]

Rizoiu, M. A., Lee, Y., Mishra, S., & Xie, L. (2017). "A tutorial on hawkes processes for events in social media." *arXiv preprint* arXiv:1708.06401. [14]

Snyder, D. L., & Miller, M. I. (2012). "Random point processes in time and space." Springer Science & Business Media. [25]

Zhuang, Jiancang and Ogata, Yosihiko and Vere-Jones, David (2002). "Stochastic declustering of space-time earthquake occurrences." *Journal of the American Statistical Association*, 97(458), 369-380 [26]