The Cost-Sharing, Shadow Price and Cluster in Medical Care Utilization: A Self-Exciting Perspective *

Yuhao LI
yuli@eco.uc3m.es
This Version Jun, 2019, First Version, Aug, 2016

Departamento de Economia, Universidad Carlos III de Madrid

Abstract

In this paper, a self-exciting counting process modelling method is proposed to study the frequency of medical care service utilization under a non-linear budget constraint health insurance policy. This modelling strategy enables researchers to investigate individual's dynamic behavior in a more detailed way. Specifically, for each individual, every doctor visiting record is represented as a point in a self-exciting counting process. Cost associated with such visiting is included in this counting process as a mark. A minimum distance method is employed to find the estimators. Using the Rand Health Insurance Experiment data, we find that individuals respond to a change of shadow price. In addition, using external models, we find that once conditional on a state-dependent structure, there is little unobserved heterogeneity effect. Lastly, we use a matured cluster analysis algorithm to investigate the cluster patterns and discover that compared to free plan, cost-sharing insurance plan with out-of-pocket fees suppress the use of medical services by limiting the number of clusters as well as follow-up visiting within each cluster.

JEL. C41, C13, C51, I13, I12

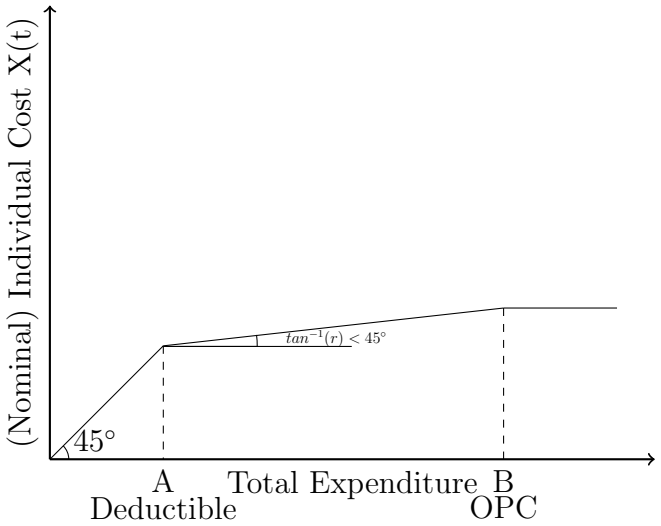
Keywords. Health insurance, non-linear budget constraint, medical service utilization, self-exciting process, history-dependent dynamic, minimum distance estimation

*I am grateful to Prof.Miguel.A.Delgado for supports and guidance throughout this project and to Prof.Winfried Stute for his inspiration and valuable comments. I would also like to thank conference and seminar participants at the EEA-ESEM Lison, the IAAE Montreal, the UC3M Econometrics workshop.

1 Introduction

In this paper, we aim to model individual's doctor-visiting behavior (outpatient only) under a non-linear cost environment using a self-exciting process. Recent studies on health insurance (e.g., Aron-Dine et al. (2012),Einav et al. (2015)) deviate from the classical assumption that individuals only respond to a single linear spot cost[1] and find strong evidence that individuals respond to the dynamic incentives associated with the non-linear nature of a typical health insurance contract. These conclusions suggest that 'it is unlikely that a single elasticity estimate can summarize the spending response to changes in health insurance' and 'such an estimate is not conceptually well defined.' (Aron-Dine et al., 2012).

The driving forces of such a non-linear nature are cost-sharing policies implemented in a health insurance contract. The most common ones are the deductible, the co-insurance rate and the out-of-pocket fee cap (OPC). In a typical setup, individuals need to cover all their medical expenditures below the deductible. Once the threshold is passed, co-insurance is applied, where individuals pay 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 total expenditure is the sum of individual costs and costs paid by the insurance. Points 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 cost for individuals (the slope is 0).

Figure 1: Non-linear Individual Cost (Medical Price)

At the heart of this non-linearity is the stochastic cumulative individual cost $X(t)$. Keeler, Newhouse, and Phelps (1977) is the first theoretical paper that studies

[1]That is, a linear budget constraint.

the consumer's optimal choice under such a non-linear medical price schedule. Using a dynamic programming model, they show that the shadow price of j^{th} episode is a function of demand prior to this episode (hence the cumulative individual cost). One may construct the shadow price (co-insurance rate) as:

$$p^s(t) = 1 - V(X(t))$$

where $0 \leq V(X(t)) \leq 1$ is a bonus that is related to the cumulative individual cost with $V^{'} > 0$. The intuition behind this equation is simple: under the range of deductibles, although individuals need to fully bear the medical cost, each time this person consumes, the remaining deductible is reduced and the next instance consumption is more easily to exceed the deductible. As a result, the shadow price for the next purchase is cheaper than the price of the current one (hence the name 'bonus'). Moreover, as the cumulative individual cost gets closer to the deductible, individuals have greater incentive to consume. That is, there should be a positive (negative) relationship between cumulative individual cost $X(t)$ (remaining deductibles) and the probability of medical utilization.

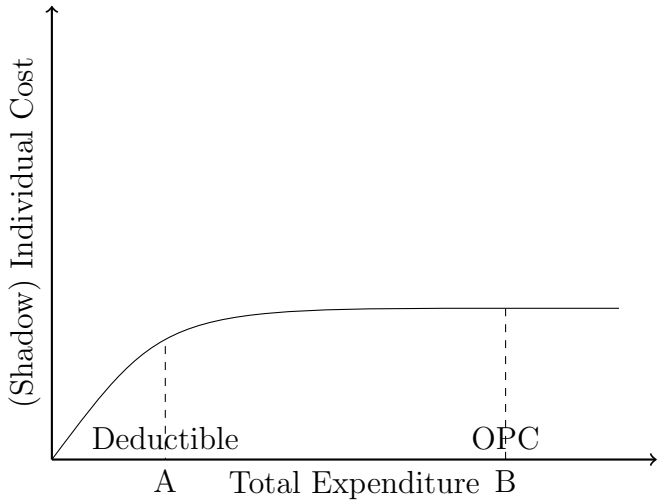
In the literature, Aron-Dine et al. (2012) construct a future price $p^f = 1 - Pr(X(T) \geq \bar{X})$ in order to reject the null hypothesis that individuals only respond to a single spot price. Here $X(T)$ is the cumulative individual cost on the last day of an insurance contract year, and $\bar{X}$ is the deductible. They find a negative relationship between the future price and the initial medical use. Notice that in their construction of future price, only $X(T)$ is used, the rest of $X(t), \forall t < T$ is ignored. In principle, one could construct future price as a function of time using the same method: $p^f(t) = 1 - Pr(X(t) \geq \bar{X})$. But in practice this would lead to a complicated procedure as one needs to use simulated future price to instrument the future price to correct the estimation bias, see Aron-Dine et al. (2012) for details.

Brot-Goldberg et al. (2017) define their shadow expected marginal end-of-year price at month m as a conditional expectation: $p^e_m = \mathbb{E}(r_{EOY}|X(m), Z, H)$ where r_{EOY} is the end-of-year co-insurance rate, Z is a vector of covariates and H is a measurement of health stock. They non-parametrically estimate the probability density function on cells of equivalent consumers using triple $(X(m), Z, H)$. In practice, they only use age as their sole explanatory variable. Their results suggest that shadow price have a limited impact on spending reduction.

Einav et al. (2015) construct and estimate a dynamic economic model to study individual's drug purchase behavior. In each period, the cumulative individual cost is updated by: $X(t) = X(t-1) + x(t)$, where $x(t)$ is the aggregate individual cost in the current period. Thus $X(t)$ here is not 'totally' stochastic: the occurrence time of illness is ignored and $x(t) = \sum_{i:t_i \in \text{current period}} x(t_i)$ is an aggregate random variable. Moreover, one may find difficulties to model the shadow price in a structural model, since the shadow budget constraint is actually unobserved by researchers.

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 shadow price is not randomly generated, an incorrectly chosen nominal price would lead to a biased estimation. Second, because the shadow co-insurance rate is a function of cumulative individual cost, it implies that individuals will make medical service utilization decisions in a sequential and contingent way. To sum up, the medical consumption behavior under a non-linear budget constrain is state-dependent. Figure 2 illustrates the situation.



Points A and B are the deductible threshold and OPC, respectively. When the total expenditure is below B, the co-insurance rate (the slope) $0 < r(X) < 1$ is a function of cumulative individual cost with $r^{'} < 0$. Whenever the total expenditure is beyond B, there is no cost for individuals.

Figure 2: Non-linear Individual Shadow Cost (Medical Price)

Notice that at any given time $\bar{t}$ within the insurance year, $X(\bar{t})$ is a random variable satisfying $X(\bar{t}) \geq X(s), \forall s \leq \bar{t}$. This non-decreasing random process is difficult to model directly. However, $X(t)$ is actually a piece-wise constant step function, we may then decompose $X(t)$ as 1) the occurrence time of i^{th} illness episode t_i (the position of i^{th} jump in this step function) and 2) conditional on the occurrence of i^{th} illness, the individual cost $x(t_i)$ for such illness (the size of i^{th} jump). Thus, we could represent the cumulative individual cost as a compound counting process: $X(t) = \sum_{i=1}^{\infty} x(t_i)\mathbb{I}\{t_i \leq t\}$. This structure suggests that we could model the time t_i and the cost $x(t_i)$ separately.

As mentioned before, the primary interest in this paper is the individual's doctor-visiting behavior, however, other sources of medical consumption also contributes to $X(t)$. Typical example is the drug purchase. These random costs serve as external shocks to our interested outpatient costs.

In this paper, we use the self-exciting process to model $X(t)$. Specifically, we assume medical costs are i.i.d. In the literature, the cost distribution is well approxi-

mated by a log-normal (Handel et al., 2015; Keeler and Rolph, 1988). Thus the key to model $X(t)$ is to model its occurrence times $\{t_i\}_{i\in\mathcal{N}^+}$.

What makes the self-exciting process suit for our purpose is that this process is conditional on a filtration that includes a σ-field which is generated by the process itself. This means, all the past information is included and the process is naturally state-dependent.

One key assumption we made is that there is little unobserved heterogeneity effect in our model. We need this assumption because it is difficult to separate the state-dependent effect from the unobserved heterogeneity effect, especially for a recurrent events analysis. This assumption might seem strong at first glance, but later, we will provide some justifications.

We use the RAND Health Insurance Experiment data. Besides it is a random experiment and is widely used in the health insurance literature, one advantage of this dataset is that it includes a detailed episode-level claim-by-claim data. We can then update $X(t)$ whenever an event or external shock occurs.

The main findings are 1) individuals will respond to the shadow price, thus we are rejecting the spot price hypothesis made by most of the literature; 2) conditional on a state-dependent structure, the unobserved heterogeneity plays an insignificant role in individual's outpatient doctor-visiting behavior.

This paper contributes three strands of literatures. First, we enrich the ever-expanding literatures that aim to study individual response to a non-linear budget constraint. Second, we introduce a new econometric tool that can be applied beyond health insurance studies. Potential applications include but not limited to labour economics (studies of multiple unemployment, work absences), industry organization (sequential entry games) and criminology etc. Last, we use a minimum distance estimation method, first introduced by Kopperschmidt and Stute (2013), to obtain the estimators, and we provide a simulation study to exam the performance of this new estimation method.

The paper is constructed as follow. Section 2 introduces some notations and basic concepts about the self-exciting process and its estimation. In addition, a simulation study is performed to study the performance of this estimation method. Section 3 introduces the data. Section 4 presents our model, in which the stochastic property of cumulative individual cost, the effect of cost-sharing policy and the dependence structure of episodes are fully considered. Section 5 presents the results, we also provide some evidences on litter heterogeneity effect to justify our key assumption. Section 6 concludes the paper.

2 Brief Introduction to Self-Exciting Process and its Estimation Method

2.1 Introduction to Self-Exciting Process

The self-exciting process is a counting process whose filtration contains a σ-field that is generated by the process itself. To begin with, we first introduce some basics on the counting process $N(t)$.

$$N(t) = \sum_{i=1}^{\infty} \mathbb{I}\{t_i \leq t\} \tag{1}$$

where $t_i, i \in \mathcal{N}^+$ are occurrence times of realized events, $\mathbb{I}\{\cdot\}$ is the indicator function. An example of such a counting process is illustrated in figure 3

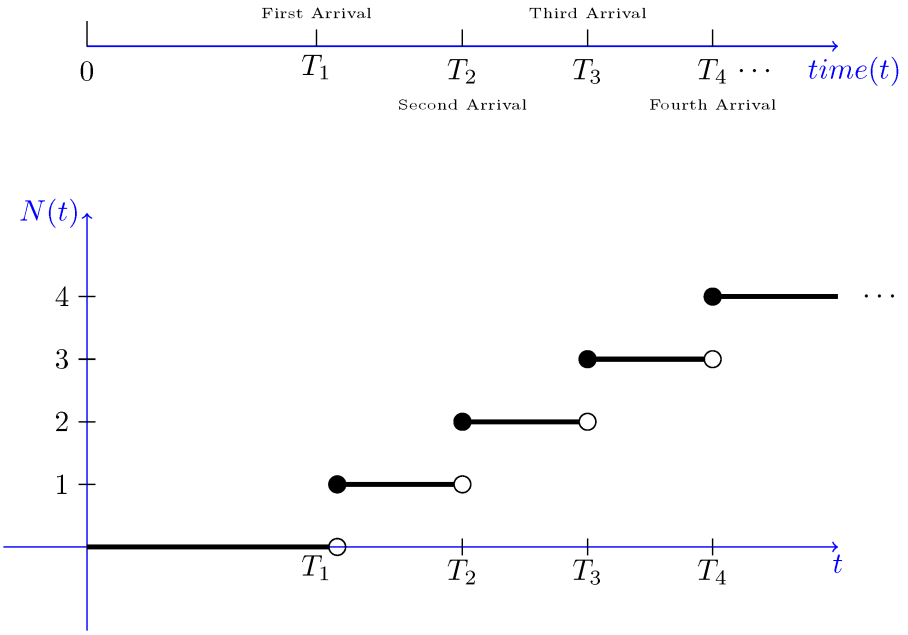


Figure 3: A possible realization of a counting process

The counting process $N(t)$ can be decomposed into a predictable part $\Lambda(t)$ called the compensator and a (local) martingale $M(t)$:

$$N(t) = \Lambda(t) + M(t) \tag{2}$$

This decomposition result is called the Doob-Meyer decomposition theorem, and it is one of the most important workhorse of counting process analysis. Taking expectation on both side and by the property of predictability and martingale we have

$$\mathbb{E}(N(t)) = \Lambda(t)$$

This equation gives us some hints on how to estimate the compensator Λ. The general idea consist of minimizing the distance between the counting process and its compensator. We posture the details of the estimation method. We can interpret the compensator as the mean of the underlying counting process at $t, \forall t$. From an economic perspective, we may say that the cumulative intensity summarizes all the systematic parts of a counting process, while the martingale accounts for the stochastic part.

Conditional on a time dependent filtration $\mathcal{F}_{t-}$, the intensity $\lambda(t)$ is defined as:

$$\lambda(t|\mathcal{F}_{t-}) = \lim_{\Delta t \to 0} \frac{\mathbb{E}(N([t, t+\Delta t])|\mathcal{F}_{t-})}{\Delta t} \tag{3}$$

For this reason, the compensator Λ is also known as the cumulative intensity. The intensity λ of a counting process is a measure of the rate of change of its predictable part.

The intensity also connects to the probability density of the underlying counting process. Let $U_{n+1} = T_{n+1} - T_n$ be the duration between n^{th} and $n + 1^{th}$ arrivals, for each arrival n, let $F_n(du) = Pr\{U_{n+1} \in du\}$ then

$$\Lambda(t) = \Lambda(T_n) + \int_0^{t-T_n} \frac{F_n(dx)}{1 - F_n(x)}, t \in (T_n, T_{n+1}]$$

where T_i is stopping time.

The self-exciting process is characterized by its filtration: if the filtration contains the σ-field generated by the counting process itself:

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

we call this counting process a self-exciting process.

In many economic studies it is of interest to introduce some external information. In the example of health insurance, we may investigate the impact of income or education on the usage of medical services. We can, in fact, enrich this filtration to include these covariates. Let $\mathcal{H}_{t-} = \mathcal{H}_0 \vee \mathcal{F}_{t-}$ be the conditioned filtration, where $\mathcal{H}_0$ is the $\sigma - algebra$ generated by some external covariates, such as age, sex, race, income, etc. We interpret this filtration as the 'whole history'. Notice that $\mathcal{H}_0$ can also be time-dependent, i.e., $\mathcal{H}(t)$, for example, in our application, the medical cost $x(t_i)$ associated with each utilization would be the case.

For more technical details about counting process and self-exciting process, we refer readers to Karr (1991).

2.2 Estimation Method

In the counting process literature, likelihood based methods are the most commonly used estimation tools, (e.g.,Ogata and Katsura (1988),Zhuang et al. (2002),Aït-

Sahalia et al. (2015),Bacry and Muzy (2014) and Mohler et al. (2012)). One requirement of using them is the predictability of the cumulative intensity Λ with respect to the filtration $\sigma(N_g(s) : s \leq t)$. That is, conditional on the filtration, the values of all the explanatory variables at time t should be known and observed just before t. However, as pointed out by Kopperschmidt and Stute (2013), in many complicated economic situations, there is little reason to maintain such an assumption. Instead, the cumulative intensity should respect external shocks or impulses. In that case, the model is most likely not dominated and the likelihood methods are difficult to apply.

In our application, a core task is to update the cumulative individual cost whenever an event occurs. Two sources of cost are considered, the first one comes from the main counting process $N^1(t)$ in which an event is a doctor visit and a mark is the associated individual cost. Another one is the drug purchase cost, represented by a mark linked to a drug purchase counting process $N^2(t)$. As a result, the individual cost coming from the drug purchase serves as an external shock to the main counting process. More precisely, the conditional filtration $\mathcal{H}_{t-}$ in our model is generated not only by the main counting process, but also by the external drug purchase counting process, i.e., $\mathcal{H}_{t-} = \mathcal{H}_0 \vee \mathcal{F}_{t-} \vee \mathcal{G}_{t-}$, where $\mathcal{F}_t = \sigma(N^1(s) : s \leq t), \mathcal{G}_{t-} = \sigma(N^2(s) : s \leq t)$.

Figure 4 helps to understand. Here t_i are occurrence times of illness episodes

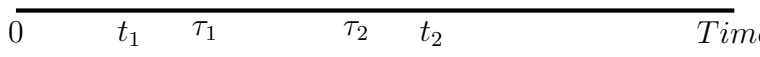


Figure 4: A possible realization of illness episodes and drug purchases

and τ_i are drug purchase times. The interested intensity λ is not predictable with respect to the filtration $\mathcal{F}$ generated only by N^1 since the cumulative individual cost is updated due to drug purchase events. But λ is predictable with respect to $\mathcal{H}$.

To overcome this problem,Kopperschmidt and Stute (2013) develop a parametric minimum distance estimation method. As mentioned earlier, the main estimation idea is to minimize the distance between the counting process and its compensator. This method only requires the observations to be i.i.d. It does not assume the differentiability of the cumulative intensity and allows unexpected jumps in the intensity function.

Formally, let $v_0 \in \Theta \subset \mathbb{R}^d$ be the true parameters, and let $N_{g,1}, ..., N_{g,n}$ be i.i.d copies of n observed ground counting process. For each $1 \leq i \leq n$, let $\mathcal{H}_i(t)$ be an increasing filtration comprising the relevant information about the marked counting process N_i as well as some other external information. Let $\Lambda_{g,v,i}$ with $v \in \Theta \subset \mathbb{R}^d$ be a given class of parametric cumulative ground intensities. Let the true one be $\Lambda_{g,i} = \Lambda_{g,v_0,i}$.

Let,

$$\bar{N}_n = \frac{1}{n}\sum_{i=1}^{n} N_i; \bar{\Lambda}_{v,n} = \frac{1}{n}\sum_{i=1}^{n} \Lambda_{v,i} \tag{4}$$

We call the former averaged point process and the latter averaged cumulative intensity. Naturally the associated averaged innovation martingale is,

$$d\bar{M}_n = d\bar{N}_n - d\bar{\Lambda}_{v_0,n} \tag{5}$$

The optimization object is:

$$||\bar{N}_n - \bar{\Lambda}_{v,n}||_{\bar{N}_n} \tag{6}$$

Where

$$||f||_\mu = [\int_0^T f^2 d\mu]^{1/2}$$

T is a terminating time. This statistic 6 is an overall measurement of fitness of $\bar{\Lambda}_{v,n}$ to $\bar{N}_n$. The estimator v_n is computed as,

$$v_n = arg \inf_{v\in\Theta} ||\bar{N}_n - \bar{\Lambda}_{v,n}||_{\bar{N}_n} \tag{7}$$

Kopperschmidt and Stute (2013) show this estimator is consistency and its asymptotic behaviour is

$$\sqrt{n}\Phi_0(v_0)(v_n - v_0) \to \mathcal{N}_d(0, C(v_0)) \tag{8}$$

where

$$\Phi_0(v) = \frac{\partial}{\partial v}\int_E (\mathbb{E}\Lambda_v(t) - \mathbb{E}\Lambda_{v_0}(t))\mathbb{E}\frac{\partial}{\partial v}\Lambda_v(t)^T \mathbb{E}\Lambda_{v_0}(dt) \tag{9}$$

$C(v_0)$ is a $d \times d$ matrix with entries

$$C_{ij}(v_0) = \int_E \phi_i(x)\phi_j(x)\mathbb{E}\Lambda_{v_0}(dx) \tag{10}$$

and

$$\phi_i(x) = \int_{[x,\bar{t}]} \mathbb{E}\frac{\partial}{\partial v_i}\Lambda_v(t)\mathbb{E}\Lambda_{v_0}(dt) \mid_{v=v_0}, \underline{t} \le x \le \bar{t} \tag{11}$$

Remark Let Φ_n be the empirical analog of Φ_0,

$$\Phi_n(v) = \frac{\partial}{\partial v} \int_E (\bar{\Lambda}_{v,n}(t) - \bar{\Lambda}_{v_0,n}(t)) \frac{\partial}{\partial v} \bar{\Lambda}_{v,n}(t)^T \bar{\Lambda}_{v_0,n}(dt) \tag{12}$$

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

$$\Phi_n(v) = \Phi_0(v) + op(1) \tag{13}$$

Such expansion guarantees that in a finite sample situation, we can replace the unknown matrix $\Phi_0(v_0)$ by $\Phi_n(v_n)$ and $C(v_0)$ by $C^n(v_n)$ without destroying the distributional approximation through $\mathcal{N}_d(0, C(v_0))$, where C^n is the sample analog of C. In practice, one needs to plug and replace the true ones with estimators and replace $\mathbb{E}\Lambda_{v_0}(dt)$ with its empirical counterpart $\bar{N}(dt)$.

In the original Kopperschmidt and Stute (2013) paper, the authors do not provide a numerical simulation study. Here we complete the task by generating a self-exciting process and examining the performance of the minimum distance estimator.

The data generating process we picked is the ETAS (epidemic type aftershock sequence) model. It was first introduced by Ogata and Katsura (1988) and ever since has been widely used in seismology (e.g. Zhuang et al. (2002)). It characterizes earthquake times and magnitudes and belongs to a marked Hawkes process family. The ETAS model has the probabilistic structure we desire: marks are part of the ground intensity and can be separated into ground intensity and conditioned mark density.

The intensity of a ETAS model, for its simplest form, could be:

$$\lambda_g(t|\mathcal{F}_{t-}) = \mu + \sum_{i:t_i<t} e^{\alpha x_i} \left(1 + \frac{t - t_i}{c}\right)^{-1} \tag{14}$$

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:

$$f(x|t, \mathcal{F}_{t-}) = \delta e^{-\delta x} \tag{15}$$

We set the true parameters as $\mu = 0.007$, $\alpha = 1.98$, $c = 0.008$ and $\delta = log(10)$. The simulation method we used is called the *thinning method*, introduced by Ogata (1981), Lewis and Shedler (1979). Briefly, this method first calculates an upper bound for the intensity function in a small time interval, simulating a value for the time to the next possible event using this upper bound, and then calculating the intensity at this simulated point. However these 'events' are known to be simulated

too frequently (Lewis and Shedler, 1979). To overcome this, the method will compare the ratio of the calculated rate with the upper bound to a uniform random number to randomly determine whether the simulated time is treated as an event or not (i.e. thinning). A full description of the algorithm is provided in the Appendix.

We generate $N = 50$, $N = 100$ and $N = 200$ individual counting processes for each repeat of the simulation and in total we have $B = 1000$ repeats. The time-intervals are set to be $[0, 3000]$, $[0, 500]$ and $[0, 100]$.

The estimation results are presented in Tables 1-3. As the number of observations N increase, 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 1: Minimum Distance Estimator Results, with $T = 3000$

$N = 400$	True	MDE	se	CI95	CI90
μ	0.007	0.006957441	0.0006271522	94.9%	92. 5%
α	1.98	1.978269	0.07331051	93.5%	90.8%
c	0.008	0.008130796	0.001724244	93.9%	91.3%
Distance	1.48622	0.715594			
$N = 200$					
μ	0.007	0.006960397	0.0008400477	93.6%	88.3%
α	1.98	1.984108	0.1086315	93.2%	90.9%
c	0.008	0.008042743	0.002413533	92.6%	90.3%
Distance	2.183783	1.226474			
$N = 100$					
μ	0.007	0.00684719	0.0011465	93.4%	90.9%
α	1.98	1.964071	0.1654297	92.1%	90.1%
c	0.008	0.008570634	0.003605053	92.3%	90.5%
Distance	3.169824	2.388298			
$N = 50$					
μ	0.007	0.006809876	0.001541488	89.1%	84.9%
α	1.98	1.974604	0.2765146	87.9%	83.7%
c	0.008	0.008979804	0.005475683	86.9%	83.1%
Distance	4.293006	3.474963			

Note: The distance is calculated using the semi-norm 6 with true parameters and the minimum distance estimators, respectively.se is the mean of the standard error of each simulation. CI95(CI90) is the percentage of the 95%(90%) confidence interval generated by se that covers the true parameter.

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

$N = 400$	True	MDE	se	CI95	CI90
μ	0.007	0.006818704	0.001282338	95.3%	93%
α	1.98	1.985548	0.2580961	96.3%	93.7%
c	0.008	0.008327916	0.005313165	96%	92%
Distance	0.2336264	0.1619424			
$N = 200$					
μ	0.007	0.007056179	0.001783448	92.5%	89.6%
α	1.98	1.977045	0.4486648	91.9%	90.6%
c	0.008	0.009058691	0.008174076	91.5%	89.9%
Distance	0.3579916	0.2119269			
$N = 100$					
μ	0.007	0.00660844	0.002295691	90.1%	86.1%
α	1.98	1.76104	0.85060127	86.6%	83%
c	0.008	0.01662388	0.0174853	86.7%	83.7%
Distance	0.477976	0.4551476			
$N = 50$					
μ	0.007	0.006672302	0.002964079	90.3%	88%
α	1.98	1.761366	2.207844	91.4%	88.9%
c	0.008	0.01808354	0.02508167	90.6%	87.8%
Distance	0.6452985	1.087129			

Note: The distance is calculated using the semi-norm 6 with true parameters and the minimum distance estimators respectively. se is the mean of the standard error of each simulation. CI95(CI90) is the percentage of the 95%(90%) confidence interval generated by se that covers the true parameter.

Table 3: Minimum Distance Estimator Results, with $T = 100$

$N = 400$	True	MDE	se	CI95	CI90
μ	0.007	0.0067466	0.002320197	95.2%	92.9%
α	1.98	1.980313	1.687546	95.1%	94%
c	0.008	0.01027362	0.01646008	95.4%	93.9%
Distance	0.0365799	0.0204637			
$N = 200$					
μ	0.007	0.006614259	0.002845468	93.6%	90.6%
α	1.98	1.91999	2.273823	94.5%	93.3%
c	0.008	0.01357907	0.02549106	93.2%	92.1%
Distance	0.05125482	0.03664879			
$N = 100$					
μ	0.007	0.01317505	0.005716749	81.5%	75.7%
α	1.98	1.719879	2.227818	92.2%	89.6%
c	0.008	0.02089188	0.03664059	89%	86.9%
Distance	0.6294044	0.1808156			
$N = 50$					
μ	0.007	0.01273163	0.006974369	85.9%	82.9%
α	1.98	1.87436	3.961052	95.6%	93.5%
c	0.008	0.02130184	0.04548218	89.2%	87.2%
Distance	0.639077	0.1674467			

Note: The distance is calculated using the semi-norm 6 with true parameters and the minimum distance estimators respectively. se is the mean of the standard error of each simulation. CI95(CI90) is the percentage of the 95%(90%) confidence interval generated by se that covers the true parameter.

3 The Data

The data we used come from the well-known RAND Health Insurance Experiment (RAND HIE), one of the most important health insurance studies ever conducted. It addressed two key questions in health care financing:

1. How much more medical care will people use if it is provided free of charge?
2. What are the consequences for their health?

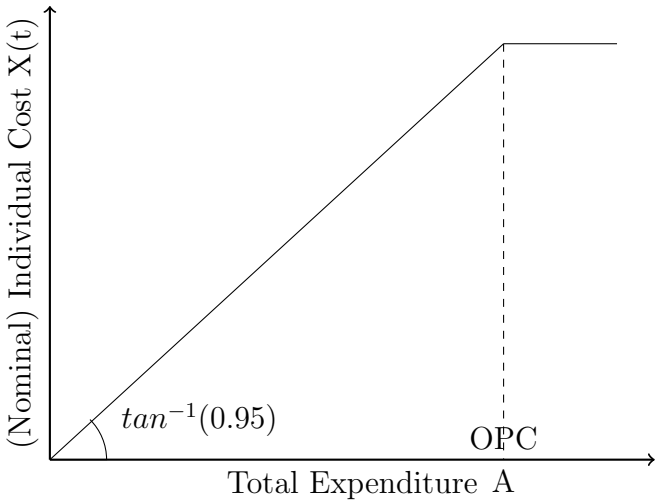
The HIE project was started in 1971 and was funded by the Department of Health, Education, and Welfare. The company randomly assigned 5809 people to insurance plans that either had no cost-sharing, 25%, 50% or 95% coinsurance rates. The out-of-pocket cap varied among different plans. The HIE was conducted 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.

Early literatures that use this data usually avoid the problem of non-linear budget constraint by assuming that individuals only respond to one price system. Typical econometric tools involved are the linear regression (after aggregating the data), the count data regression and the duration analysis. None of them is capable to fully model the stochastic structure of cumulative individual cost $X(t)$.

Because the complicated structure of our self-exciting process, to ease the burden of computation, we only use data from Seattle, which has the largest medical claim records available. We separate the data according to two different insurance plans: zero coinsurance rate plan (free plan, denoted as P0), in which the patient does not pay anything; and a cost-sharing plan (denoted as P95) in which a coinsurance rate of 95% applied and OPC is 150 USD per person or 450 USD per family[2](i.e., before exceeding the OPC, individuals need to pay 95% of the medical care cost, once the OPC is reached, all the cost is paid by the insurance.). The OPC and coinsurance rate in this plan only applied to ambulatory services; inpatient services were free. Both plans covered a wide range of services. Medical expenses included services provided by non-physicians such as chiropractors and optometrists, and prescription drugs and supplies. There is no deductible in this insurance contract. The following figure summarizes the P95 contract design.

We also include the data of drug purchase records with information such as the purchase dates and the values of non-covered charges. As discussed in the previous section, we may treat the drug purchase as another counting process and as an external shock to our primary one (doctor-visiting counting process). In the original dataset, one individual may have several claims in one day, and we combine all claims

[2]In 1973 dollars.



Point A is the OPC . When the total expenditure is below A, the co-insurance (slope) $r = 0.95$ is applied. Whenever the total expenditure is beyond A, there is no cost for individuals.

Figure 5: Contract design for P95

with an identical date into one and sum the non-covered charges.

The occurrence time stamp is defined as the annual duration between the beginning date of the insurance policy and the date this person visited a health care institution. For example, if the insurance 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 the records with missing duration information. (Hence we exclude the cases of censoring.)

When analyzing the cost-sharing plan, we restrict our dataset within the contract year 1977-1978 since the cost-sharing policies are renewed annually. But such restriction is not needed for the free plan since there is no within-year cost sharing policy. For this plan, the time horizon ranges from 1975 to 1980. When the individual cost information is missing, we replace it with zero. In the end, we have 243 individuals in the free plan with 7638 claims over the years and 131 individuals in the cost-sharing plan and the total number of claims is 1103 within the 1977-1978 contract year.

We also include some demographic covariates: age, sex, education (in terms of schooling years) and log-income. For simplicity, we fixed all ages at the enrolment time. Thus all covariates are time-independent. More covariates can be added, but we are limited by computation capacity.

4 The Model

As discussed before, the focal point of the self-exciting counting process approach is to model the (cumulative) intensity function. We construct the intensities by

explicitly taking different randomness sources of cumulative individual cost $X(t)$ and episode dependence structure into consideration.

We further assume that 1) the initial event (or equivalently, the duration for the first doctor-visiting event under a new insurance contract) is given. Thus, we are not interested in modeling the first event, and the counting processes exclude all the first events. 2) There is no unobserved heterogeneity.

The second assumption seems to be strong at first glance. However, we will provide evidences to justify it in later section.

4.1 Free Insurance Plan

Our intensity $\lambda(t)$ for each individual[3] who belongs to the free insurance plan consists of two parts:$\lambda(t) = \lambda_1\lambda_2(t)$. λ_1 deals with the covariates effect, while λ_2 is the state-dependent (SD) term.

Like many count data regression and duration models, the covariates effect is presented as an exponential function:

$$\lambda_1(Z) = exp(\gamma^T Z) \tag{16}$$

where Z is a vector of individual characteristics including age, sex, education and log-income, etc.

The SD term is specified as:

$$\lambda_2(t) = \sum_{i=1}^{N(t-)} \mu \cdot exp(-\mu(t - t_i)), \mu > 0 \tag{17}$$

The 'kernel' $\mu \cdot exp(\cdot)$ characterizes the episode dependence structure. More specifically, the propensity of a follow-up visit is governed by such a 'kernel': the intensity is high when the elapsed time is short and will gradually decrease as time goes by. We will argue such an assumption is reasonable: the individual is vulnerable when she just receives the treatment and is more likely to be sick again, but she will gradually recover as time goes by and will be less likely to experience sickness. The summation over these 'kernels' means we take all the past episodes into consideration. But the weight for each episode is different. By construction, the effects of far away past experiences will deteriorate, but the latest ones have the most important influences.

The usual method to model such phenomena in a structural form model is to assume health events arrive periodically with a probability $S^{'}$, which is drawn from $F(S^{'}|S)$ where S is the arrival probability from a previous period. Einav et al. (2015) further simplify this assumption by letting S take one of two values, S^L and S^H

[3]Therefore, we ignore the individual subscript.

(with $S^L < S^H$), and that $Pr(S' = S^J | S = S^J) \geq 0.5, J \in \{L, H\}$, so there is weakly positive serial correlation. This exceedingly simplified assumption is made mainly for computational reasons. The above Markov process is most likely inadequate to model the episode cluster structure. We believe our cluster set up is more realistic and is quite difficult, if not impossible, to build within the conventional econometric models.

To sum up, for the free insurance plan, P0, the intensity is expressed as:

$$\lambda_{P0}(t) = \lambda_1(Z)\lambda_2(t) \tag{18}$$

4.2 Cost-Sharing Insurance Plan

As for the cost-sharing plan, λ_1 does not change. The cost sharing policy has two hypothetical effects: 1) The late year effect, that is when the contract year is near the end, individuals, especially those who have already exceeded the OPC may use the medical service more frequently than before (cash-in effect) since the cost-sharing policy will be set to default next year and the shadow co-insurance rate would be expensive once again. 2) The shadow price effect discussed in the introduction section. We update the cumulative individual cost whenever an event occurs. To account for the cost-sharing effects, we modify λ_2 as follows:

$$\lambda_2^*(t) = \beta_1 exp(\beta_1 t) + \sum_{i=1}^{N(t-)} b \exp(\beta_2 X(t_i)) \mu exp(-\mu(t - t_i)) \tag{19}$$

Here $X(t)$ is the cumulative individual cost at time t. It includes the non-covered charge from outpatient medical utilization as well as drug purchase:

$$X(t) = \sum_{i=1}^{N^1(t-)} x_i + \sum_{i=1}^{N^2(t-)} y_i \tag{20}$$

where x_i is the non-covered charge for i^{th} doctor visiting, $N^1(t)$ is the associated ground counting process. y_i is the non-covered charge for i^{th} drug purchase and $N^2(t)$ is the drug purchase ground counting process. The construction of $X(t)$ essentially follows the definition of cumulative individual cost mentioned in the introduction section. Recall the shadow price is defined as $1 - V(X(t))$, where $V(X(t))$ is the bonus which depends on the cumulative individual cost. If $V(X(t)) \propto \exp(\beta_2 X(t))$, then the term $b \exp(\beta_2 X(t))$ can be thought of as a measure of medical utilization bonus. We would expect $\beta_2 > 0$ to be significant if individuals do respond to shadow price.

We use the term $\beta_1 exp(\beta_1 t)$ to model the late year effect: we would observe β_1 significantly greater than zero if such an effect is true.

To summarize, the ground intensity for the cost-sharing plan is:

$$\lambda_{P95}(t) = \lambda_1(Z)\lambda_2^*(t) \tag{21}$$

There are several pieces to put together in order to estimate the parameters of the cost-sharing effects model. As Keeler and Rolph (1988), we assume that there are no interactions between within-year cost sharing effects and the effects of other explanatory variables, so that all the effects of explanatory variables other than cost sharing on frequencies of episodes are summarized in $\lambda_1(Z)$ and all episode dependence structure is captured by $\lambda_2(t)$ ($\lambda_2^{'}(t)$). We first estimate the free plan by minimizing

$$||\bar{N}^{P0} - \lambda_1(Z) \int_0^T \lambda_2(t)dt||_{\bar{N}^{P0}}$$

thus, the individual heterogeneity and the episode dependence structure of the intensity are estimated by $\hat{\lambda}_1(Z)$ and $\hat{\lambda}_2(t)$. When estimating the cost-sharing plan, these two parts are then treated as fixed, which leaves us with only cost-sharing effect parameters (i.e.,β_1, β_2 and b) to be estimated (We exploit the fact that all individuals are assigned to different plans randomly. By plugging the individual specific estimators from the free plan into the cost-sharing plan, we can still have consistent estimators). Thus the minimization object is:

$$||\bar{N}^{P95} - \hat{\lambda}_1(Z) \int_0^T \big(\beta_1 exp(\beta_1 t) + \sum_{i=1}^{N_g(t-)} b \exp(\beta_2 X(t_i))\hat{\mu} exp(-\hat{\mu}(t - t_i))\big)dt||_{\bar{N}^{P95}}$$

5 Main Results

The main results are presented in Table 4.

5.1 Interpreting the Covariates

The interpretation of coefficients is not as straightforward as in linear regression. However, we may fix a time period and treat the counting process as count data. The interpretation is then identical to that of a count data regression analysis. Formally, recall the Doob-Meyer decomposition, for a fixed time period $[0, t], \forall t \in [\underline{t}, \bar{t}]$, we have

$$\mathbb{E}(\Lambda(t|Z)) = \mathbb{E}(N(t)|Z) = \mathbb{E}(Y_t|Z)$$

The count data Y_t is the number of events occurring during this time period. 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 $\lambda_1(Z)$. That is, 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 these in mind, we can interpret our results.

Age. The overall effect for age is as follows: at first, the intensity will decrease as age increases, after one passes the age of 41.5, the intensity and age are positively correlated. It is well-known that the youngsters are more risky compared to their

Table 4: Basic Results

	Estimator	*Description*
μ	25.264612***	coefficient of the episode dependent structure
	(4.230528)	
age	-0.230274***	
	(0.084052)	
age2	0.277269***	$(age)^2/100$
	(0.115787)	
male	-0.554904	
	(0.447628)	
edu	-0.190387	
	(0.204672)	
edu2	0.184098	$(edu)^2/100$
	(0.862468)	
log income	0.685771***	
	(0.141580)	
b	0.65898635***	
	(0.0580308)	
β_1	0.1068388	coefficient of late year effect
	(0.27547018)	
β_2	0.00383393***	coefficient of non-covered charge
	(0.00061892885)	
Distance	1.00747	Free Plan
Distance	1.10226	Cost-Sharing Plan

Note: standard errors in brackets, *p<0.1; **p<0.05; ***p<0.01

mid-age counterparts. While as individuals begin to age, they become physically weaker and more prone to sickness.

Sex. Females seem to be more likely to go the doctor, but the result is not significant.

Education. The multi-parameter Wald Test Statistics for $edu = edu2 = 0$ is 13.33125, indicating the education is a significant factor to the doctor-visiting behaviors. The result suggests a negative relation between education and the outpatient medical utilization. One explanation could be that individuals with a higher level of education are positioned in more important jobs and their absence from work may damage not only their output but also that of their peers', thus the potential cost of going to hospital is much higher which leads to a negative correlation.

Income. Income is positively related to the use of medical services, which is not surprising. A higher income gives individuals the ability to cover the opportunity cost related to absence from work (to visit a doctor).

The shadow price effect is captured by $b\exp(\beta_2 R(t))$. The most important parameter here is β_2. If β_2 is close enough to zero, we may observe a flat, almost linear curve, which indicates that individuals only respond to one price system (the spot price system). However, if β_2 is positively away from zero, we can safely claim that individuals do understand the design of the insurance policy and take advantage of the shadow price. Our result provides strong evidence for the shadow price effect and we are confident to reject the null hypothesis: $\beta_2 = 0$.

There is weak evidence supporting the existence of the late year effect.

5.2 Little Unobserved Heterogeneity Effect Evidence

In a variety of contexts, it is often noticed that individuals who have experienced an event in the past are more likely to experience the event again in the future than are individuals who have not experienced the event (Heckman, 1981). One explanation, best known as the unobserved heterogeneity, is that in addition to the observed variables, there are other relevant variables that are unobserved but correlated with the observed ones.

Another explanation of heterogeneity is related to state dependence (SD). This concept says that past experience has a genuine effect on future events in a sense that an otherwise identical individual who did not experience the event would behave differently in the future. The definition of the self-exciting process naturally includes the idea of state dependence.

Compare with the usual unobserved heterogeneity settings, the state dependence differs in several ways. First, the SD term is time dependent, which means that it can be updated, while the typical UH term is time persistent. Second the choice of the SD term is flexible and can be consistent with economic theory. For example, we may

capture the seasonality effect by setting the term as $K(t_i, t) = \alpha sin(\beta(t - t_i) + \gamma) + \delta$, or in our application, we may study the cluster phenomenon of medical care utilization by letting $K(t_i, t) = \mu exp(-\mu(t - t_i)), \mu > 0$.

Before presenting evidences on little unobserved heterogeneity, we would like to demonstrate that a self-exciting process can generate enough heterogeneity without introducing a latent variable to represent the unobserved heterogeneity.

The data generating process (DGP) is the same as in the simulation studies before. Figure 6 presents three quite different individual's event histories simulated by this DGP using the identical parameter settings as stated before.

Individual 1 has the most frequent events experience, the total number of events is 92. Individual 2 is somewhat moderate, with 37 events. Individual 3 has the least frequent events with only 2 during the time interval $[0, 100]$.

We now present some evidences on the little unobserved heterogeneity effect.

5.2.1 Evidence One: Modeling the Initial Duration Using Heckman and Singer's NPMLE

The subject under study here is the initial duration to visit a doctor under the free plan.

In the main model, we assume that the initial duration is given. Here we take the advantage that the initial event under a new free insurance plan can be reasonably assumed to be free from any state dependent effect such that if there exists any unexplained heterogeneity, it should come from the unobserved term.

We specify the hazard rate and its cumulative hazard function for the initial duration as:

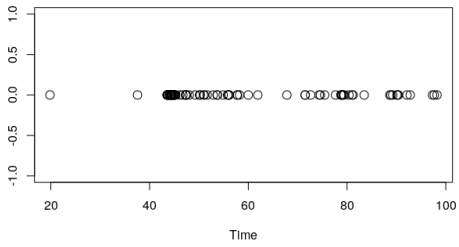
$$h_i(\nu_i, X_i) = \exp(X_i'\beta + \nu_i)$$

$$H_i(t) = h_i(\nu_i, X_i)t$$

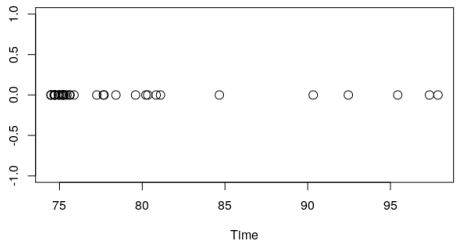
where ν_i is the unobserved heterogeneity and X_i is a vector of observed covariates. The likelihood contribution of each individual is

$$L_i = \exp(-H_i(t))h_i(\nu_i, X_i)$$

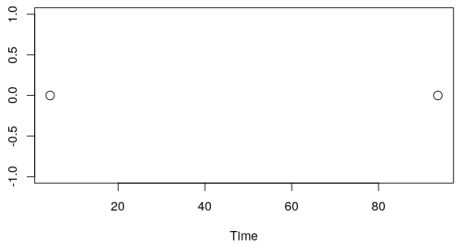
Assume $\nu \sim G$ is independent from X_i, one may use Heckman and Singer (1984)'s non-parametric maximum likelihood estimator (NPMLE) to avoid unjustified assumptions about the distribution G. Instead, one may approximate G in terms of a



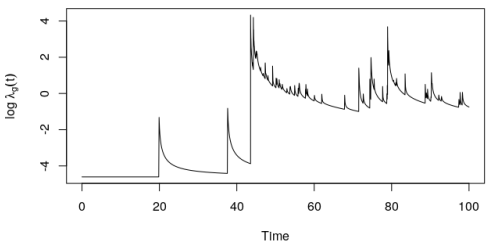
(a) Individual 1, Event Time



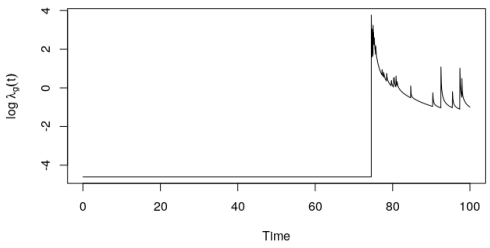
(b) Individual 2, Event Time



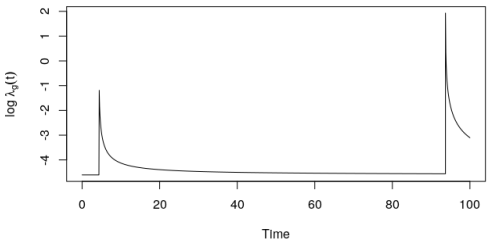
(c) Individual 3, Event Time



(d) Individual 1, log of intenstiy



(e) Individual 2, log of intenstiy



(f) Individual 3, log of intenstiy

Figure 6: Three individual's events histories

discrete distribution.

Let Q be the (prior unknown) number of support points in this discrete distribution and let $\nu_l, p_l, l = 1, 2, \cdots, Q$ be the associated location scalars and probabilities. The likelihood contribution is:

$$\mathbb{E}[L_i(\nu_i)] = \sum_{l=1}^{Q} p_l L_i(\nu_l), \sum_{l=1}^{Q} p_l = 1$$

where $L_i(\nu_l) = \exp(-H_i(t|\nu_l, X_i))h_i(t|\nu_l, X_i)$.

The likelihood function is

$$L = \prod_{i=1}^{N} \mathbb{E}[L_i(\nu)] = \prod_{i=1}^{N} \sum_{l=1}^{Q} p_l L_i(\nu_l), \sum_{l=1}^{Q} p_l = 1$$

The estimation procedure consists of maximizing the likelihood function with respect to β as well as the heterogeneity parameters ν_l and their probabilities p_l for different values of Q. Starting with $Q = 2$, and then expanding the model with new support points until there is no gain in likelihood function value. Heckman and Singer (1984) has proven that such an estimator is consistent, but its asymptotic distribution has not been discussed yet.

If there is little unobserved heterogeneity, one would expects that when $Q = 2$, one of the mass point probability, say, $p_1 \approx 1$ and the other $p_2 \approx 0$, in addition, the heterogeneity parameters should be similar in value $\nu_1 \approx \nu_2$.

Table 5 shows exactly this situation.

5.2.2 Evidence Two: Group Heterogeneity on the Initial Duration

Instead of assuming individual heterogeneity, one might assume the group heterogeneity and reveal the group affiliation through an external model. Such a typical model is the finite mixture model.

Assume individuals belong to k different groups and the initial duration y_i are governed by a finite mixture reverse Gumbel:

$$p(\mathbf{y}|\Theta) = w_1 f_1(\mathbf{y}|\Theta_1) + w_2 f_2(\mathbf{y}|\Theta_2) + \cdots + w_k f_k(\mathbf{y}|\Theta_k)$$

where $\Theta = (\Theta_1, \Theta_2, \cdots, \Theta_k, \mathbf{w})'$ denotes the vector of all parameters, $\mathbf{w} = (w_1, w_2, \cdots, w_k)'$ is a vector of weight whose elements are restricted to be positive and sum to unity. $f_k(\cdot|\Theta_k)$ is a reverse Gumbel density with the vector of parameters Θ_k.

Additionally, we may equivalently model the finite mixture model in a hierarchical manner using a latent variable l_i, which represents the allocation of each observation

Table 5: Initial Duration (Free Plan) Results

	Estimator	*Description*
age	-0.000572	
age2	-0.001808	$(age)^2/100$
male	-0.294545	
edu	-0.145991	
edu2	0.494294	$(edu)^2/100$
log income	0.092655	
p_1	0.989619	
p_2	0.010380	
ν_1	0.385701	
ν_2	0.385704	

y_i to one of the components:

$$p(y_i|\Theta_k, l_i = k) = f_k(y_i|\Theta_k)$$
$$p(l_i = k) = w_k$$

One may have the group affiliation posterior by the Bayesian rule:

$$\begin{aligned} p(l_i = k|y_i, \Theta_k) &= \frac{p(y_i|l_i = k, \Theta_k) * p(l_i = k)}{p(y_i|\Theta_k)} \\ &= \frac{p(y_i|l_i = k, \Theta_k) * w_k}{\sum_{k=1}^{K} p(y_i|l_i = k, \Theta_k) * w_k} \end{aligned}$$

we may then assign the group affiliation according to the posteriors.

For simplicity, we assume group number $k = 2$.

The typical estimation method for a finite mixture model is the EM algorithm. Once we 'reveal' each individual's group affiliation, we might include this new covariate into our main model. If there is little unobserved heterogeneity effect, we should expect the coefficient for this new covariate to be insignificant from zero.

Table 6 reports the results and further confirm that there is indeed little unobserved heterogeneity effect. The multi-parameter Wald test statistic for $age = age2 = 0$ is 567.885, the statistic for $edu = edu2 = 0$ os 10.719, indicating both age and education is significant. However, the group effect is not significant.

5.2.3 Evidence Three: Group Heterogeneity on counts

Similar to what we did before, here, instead of modeling the initial duration, we model the number of doctor-visiting for five years under a free plan.

Again, we assume the number of group $k = 2$, and the interested parameter would be the coefficient of the group affiliation.

Table 7 reports the results, which further confirm that there is little unobserved heterogeneity.

One possible explanation for this little heterogeneity effect could be this: the subject under this study is the outpatient doctor-visiting, which is mostly caused by minor episodes and injuries. If an individual has not experienced any illness in a relatively long period, then the unobserved heterogeneity might contribute to the next illness, but the effect from the state dependent structure is the dominating cause for the future doctor-visiting. That is this initial illness might cause next episodes and this structure outweighs the individual's unobserved heterogeneity.

To sum up, conditional on a state dependent structure, the unobserved heterogeneity plays little roles (as best demonstrated in the Evidence Two) in individual's doctor-visiting behaviors, at least for outpatient case.

Table 6: Group Heterogeneity on Initial Duration (Free Plan) Results

	Estimator	*Description*
μ	29.035433***	coefficient of the episode dependent structure
	(4.935501)	
age	-0.124143	
	(0.10111)	
age2	0.146278	$(age)^2/100$
	(0.13168)	
male	-0.442570	
	(0.44991)	
edu	-0.134448	
	(0.18627)	
edu2	0.254624	$(edu)^2/100$
	(0.69313)	
log income	0.396356***	
	(0.17695)	
Group	0.561757	Group affiliation
	(0.613766)	
Distance	0.631360	Free Plan

Note: standard errors in brackets, *p<0.1; **p<0.05; ***p<0.01

Table 7: Group Heterogeneity on counts (Free Plan) Results

	Estimator	*Description*
μ	28.270280***	coefficient of the episode dependent structure
	(8.267194)	
age	-0.149708***	
	(0.030242)	
age2	0.174288***	$(age)^2/100$
	(0.045306)	
male	-0.599141	
	(0.588516)	
edu	-0.417321***	
	(0.085243)	
edu2	1.195932***	$(edu)^2/100$
	(0.360884)	
log income	0.64824***	
	(0.042359)	
Group	0.301163	Group affiliation
	(0.377391)	
Distance	0.864397	Free Plan

Note: standard errors in brackets, *p<0.1; **p<0.05; ***p<0.01

5.3 Cluster Analysis

The episodes tend to be clustered or grouped together (i.e., we are rejecting the assumption that episodes are independent). One reason is because of the nature of chronic diseases: regular or frequent treatments are needed to ease or eliminate the pain. Another explanation is because one disease may trigger the occurrence of another one in the short term.

As mentioned before, the dependent structure (or the cluster structure) is governed by $\exp(-\mu t)$. Figure 7 presents such a structure using our estimator $\hat{\mu}$.

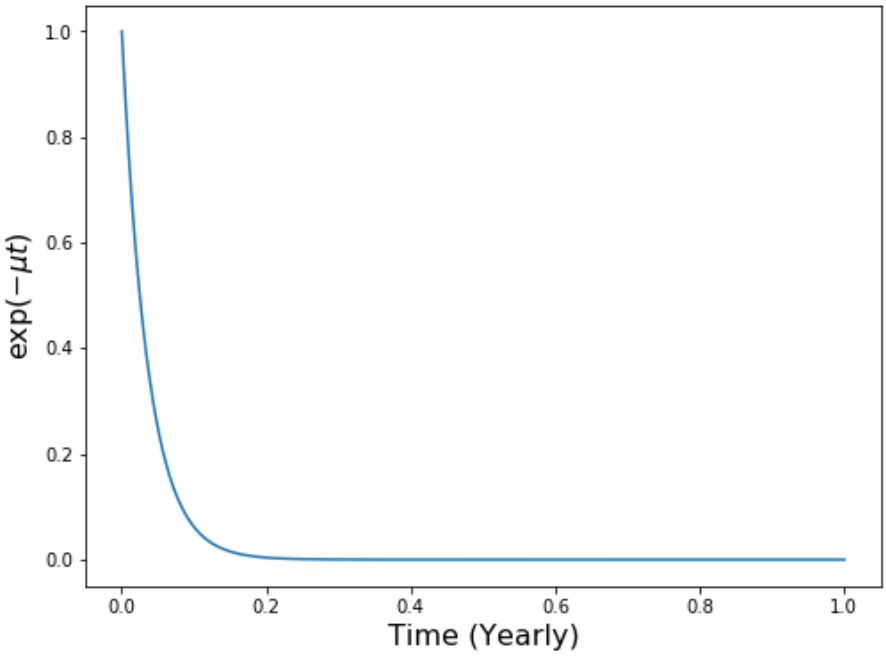


Figure 7: The cluster structure among episodes

The likelihood of follow-up visiting is high at the beginning and then decreases as time goes by. After roughly 3 weeks to one month, the likelihood is small enough to ignore.

Literatures have documented that cost-sharing policies reduce the frequency of medical utilization (e.g. Keeler and Rolph (1988); Aron-Dine et al. (2015)). Our question is how do these policies affect the cluster structure among doctor visiting? Will they reduce the average number of clusters per person? Will they reduce the average follow-up visits inside a cluster? To the best of our knowledge, few literature has touched on these issues, since most of them use the hypothesis that episodes are independent.

Here we use a cluster analysis algorithm called DBSCAN (Density-based spatial clustering of applications with noise) that is widely used in computer science and

statistical learning.

For this algorithm, there are two inputs: Eps, the radius of one density region, and minPts, the minimum number of points required to form a dense region. For the purpose of DBSCAN clustering, all points are classified as core points, border points and 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 Appendix.

The ability of this algorithm to identify 'noise' points is particularly appealing to us. This is because some acute episodes are small in scale and only need one doctor visit to fully recover. They are not linked to the rest of episodes.

Based on the estimation of the SD term, we set $Eps = 21$ days and as a rule of thumb $minPts = 2^4$. For the purpose of comparison, we restrict the time horizon in both plans to 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 then 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 8 summarizes the results.

Table 8: 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

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.

6 Conclusion

In this paper, we provide a methodology to construct a behavioral model of medical care utilization. At the core of this method is the self-exciting counting process. It allows researchers to take historical information into the model. A minimum distance estimation is employed. By doing so, one may introduce external shocks to the self-exciting process. This enables researchers to use more realistic model settings. We use such a methodology to build a decision making process model of medical care utilization and find that individuals are responsive to shadow

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

price and take into account the dynamic incentives. Furthermore, using different external models (Duration analysis on initial duration, finite mixture models), we find that once taking the state-dependent structure into the model, the unobserved heterogeneity plays an insignificant role. Lastly, using a matured statistical learning algorithm, we analyze the cluster structure of doctor visiting behaviors. We find cost-sharing policies do affect the clusters in numerous ways.

References

Yacine Aït-Sahalia, Julio Cacho-Diaz, and Roger JA Laeven. Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, 117(3):585–606, 2015.

Aviva Aron-Dine, Liran Einav, and Amy Finkelstein. The rand health insurance experiment, three decades later. Technical report, National Bureau of Economic Research, 2012.

Aviva Aron-Dine, Liran Einav, Amy Finkelstein, and Mark Cullen. Moral hazard in health insurance: do dynamic incentives matter? *Review of Economics and Statistics*, 97(4):725–741, 2015.

Emmanuel Bacry and Jean-François Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7):1147–1166, 2014.

Zarek C Brot-Goldberg, Amitabh Chandra, Benjamin R Handel, and Jonathan T Kolstad. What does a deductible do? the impact of cost-sharing on health care prices, quantities, and spending dynamics. *The Quarterly Journal of Economics*, 132(3):1261–1318, 2017.

Liran Einav, Amy Finkelstein, and Paul Schrimpf. The response of drug expenditure to nonlinear contract design: Evidence from medicare part d. *The quarterly journal of economics*, 130(2):841–899, 2015.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

Benjamin R Handel, Jonathan T Kolstad, and Johannes Spinnewijn. Information frictions and adverse selection: Policy interventions in health insurance markets. Technical report, National Bureau of Economic Research, 2015.

James Heckman and Burton Singer. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, pages 271–320, 1984.

James J Heckman. Heterogeneity and state dependence. In *Studies in labor markets*, pages 91–140. University of Chicago Press, 1981.

Alan Karr. *Point processes and their statistical inference*, volume 7. CRC press, 1991.

Emmett B Keeler and John E Rolph. The demand for episodes of treatment in the health insurance experiment. *Journal of Health Economics*, 7(4):337–367, 1988.

Emmett B Keeler, Joseph P Newhouse, and Charles E Phelps. Deductibles and the demand for medical care services: The theory of a consumer facing a variable price schedule under uncertainty. *Econometrica*, pages 641–655, 1977.

Kai Kopperschmidt and Winfried Stute. The statistical analysis of self-exciting point processes. *Stast. Sinica*, 23:1273–1298, 2013.

Peter A Lewis and Gerald S Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 2012.

Yosihiko Ogata. On lewis' simulation method for point processes. *Information Theory, IEEE Transactions on*, 27(1):23–31, 1981.

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

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

A The Thinning Method for Simulation

The detailed thinning method steps can be summarised as:

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

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

   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

B DBSCAN Cluster Analysis

The DBSCAN algorithm classified all points into three: core points, border points and noise points. We start by defining these points. For a set of points $X = \{x_1, x_2, \cdots, x_N\}$.

Definition ϵ neighbourhood of a point x, denoted by $N_\epsilon(x)$ is defined by $N_\epsilon(x) = \{y \in X : d(y, x) \leq \epsilon\}$. Where $d()$ is a metric.

Definition Density is defined as $\rho(x) = |N_\epsilon(x)|$, the number of points in a ϵ neighbourhood.

Definition Core point: let $x \in X$, if $\rho(x) \geq minPts$, then we call x a core point. The set of all core points is denoted as X_c, let $X_{nc} = X \setminus X_c$ be the set of all non-core points.

Definition Border point: if $x \in X_{nc}$ and $\exists y \in X$ such that $y \in N_\epsilon(x) \cap X_c$, then x is called a border point. Let X_{bd} be the set of all border points.

Definition Noise point: let $X_{noise} = X \setminus (X_c \cup X_{bd})$, if $x \in X_{noise}$, then we call x is a noise point.

To define what is a cluster under the DBSCAN setting, we need a few more definitions about 'reachable'.

Definition Directly density-reachable: if $x \in X_c$ and $y \in N_\epsilon(x)$, we may say y is directly reachable from x.

Definition Density-reachable: let $x_1, x_2, \cdots, x_m \in X, m \geq 2$. If x_{i+1} is directly density-reachable from x_i, $i = 1, 2, \cdots, m-1$. We call x_m is density-reachable from x_1.

Definition Density-connected: a point x is density connected to a point y if there exists another point $z \in X$ such that both y and x are density-reachable from z.

Definition Cluster: a non-empty subset C of X is called cluster if it satisfies:

- (Maximality) $\forall x, y$: if $x \in C$ and y is density-reachable from x, then $y \in C$.
- (Connectivity) $\forall x, y \in C$: x is density-reachable to y.

For a detailed algorithm description, we refer to the original Ester et al.(1996)Ester et al. (1996) paper.