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

Yuhao Li
liyuhao.econ@whu.edu.cn

School of Economics and Management
Wuhan University

December,2021

Abstract

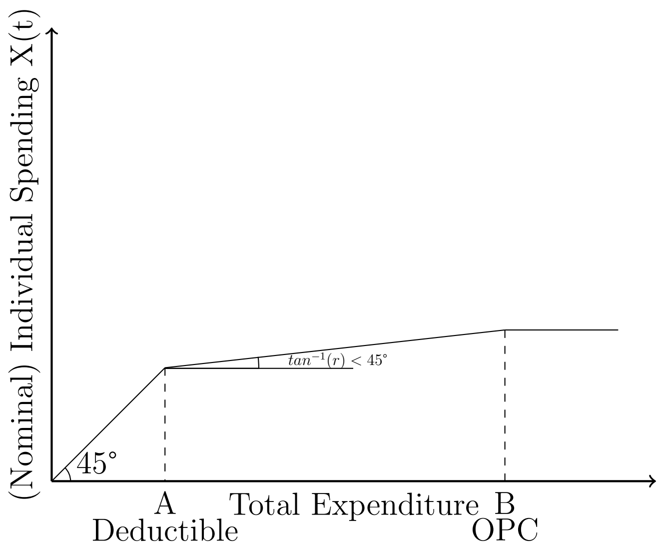
In this paper, a Hawkes process is proposed to study an individual's responsiveness to the outpatient utilization under a nonlinear health insurance contract. For a given individual, each doctor visit record is represented as a point in a Hawkes process. When measuring the responsiveness, we adopt the episode-varying shadow price instead of a constant spot price or the expected end-of-year price. Using the RAND Health Insurance Experiment data, we found that individuals do understand the price incentive of a nonlinear contract. Our model can also describe the cluster patterns of outpatient utilization under different insurance plans. Comparing to a free insurance plan, a canonical individual in the cost sharing insurance plan would have fewer doctor visits in a cluster, and the cluster number also shrinks.

JEL. C41, C13, C51, I13, I12

Keywords. Health Insurance, Shadow Price, Hawkes Process, Dynamic Behavior

1 Introduction

A nonlinear health insurance contract is widely used in practice as a tool to control the moral hazard, i.e., the additional health care that is purchased when an individural became insured. It is characterized by various cost sharing policies. 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 the medical expenditures until 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 bulk of the evidence suggests the introducing of



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

cost sharing tools do reduce spending. More specifically, the reduction is achieved mainly through quantity whereby individuals receive less medical care, 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 an individual makes her medical consumption quantity under a nonlinear price system.

Measuring the consumer's responsiveness to medical care price is a central issue in health economics and a key ingredient in the optimal design of health insurance markets. Historically, literature studying the price elasticity of health insurance contracts often assume that individuals only respond to (out-of-pocket) 'spot' price. Cutler and Zeckhauser (2000) summarize about thirty studies that adopt this assumption. Perhaps, the most famous result in this strand of literature is Keeler and Rolph (1988); Manning et al. (1987), 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. Then trying to summarize an individual's medical spending behavior with single price elasticity is therefore not well-defined. As mentioned in Aron-Dine et al. (2013, 2015), '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 deviated from this assumption, see Aron-Dine et al. (2013); Brot-Goldberg et al. (2017); Einav et al. (2015), but found mixed evidence on individuals' responsiveness to the dynamic incentives created by the health insurance plan.

Among literature that avoid using a single price, most of them depend heavily on specific data structure and strong homogeneous assumptions. For example, Aron-Dine et al. (2015) use 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 strong assumption that 'individual have no private information about their health shocks'. Brot-Goldberg et al. (2017) also use 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, the only characters they use are sextiles based on health status and age.

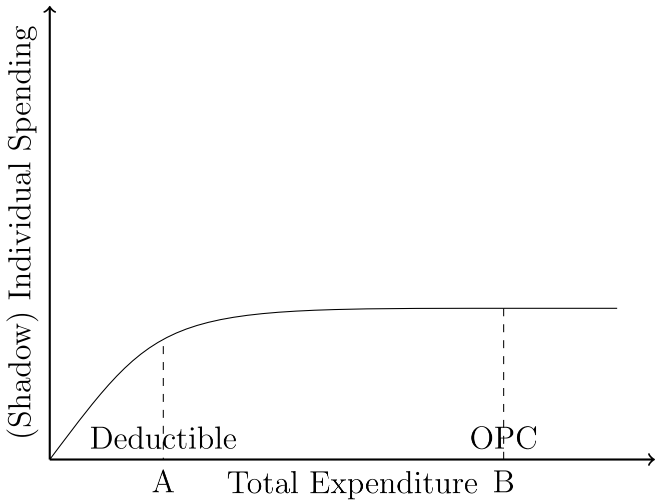
This paper describes a new framework designed for studying the consumer's responsiveness to medical price. We will focus on an individual's outpatient medical consumption quantity, as inpatient consumptions are infrequent and are often associated with large expenditures that exceeds the OPC easily (hence, decreasing future price to near zero). What we contribute to the existing studies is the usage of an episode-variant shadow price instead of a spot price or a fixed end-of-year price as in Aron-Dine et al. (2013). Within this framework, we would model and compare an individual's doctor visits under a free plan and those under a cost sharing plan. Our strategy requires less restrictive data structure and model assumptions. Moreover, unlike previous literature that use static models, we are able to describe an individual's dynamic medical spending based on a shadow price that is conditional on this person's own year-to-date accumulative spending $X(t)$. For a given individual i, the shadow price is defined as:

$$P_i^s(t) = \mathbb{E}(P_{EOY} \mid X_i(t)) = 1 - V(X_i(t))$$

where $P_i^s(t), P_{EOY}$ are the shadow price at time t and the end-of-year spot price respectively. $0 \leq V(X_i(t)) \leq 1$ with $V' > 0$ can be understood as a bonus function. The intuition behind such a definition is simple: If current $X_i(t)$ is under the deductible threshold, every health care consumption will lead to an increase of $X_i(t)$, making an individual to 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 accumulative individual spending $X(t)$ to study an individual's responsiveness to this non-linearity budget constraint.

Keeler, Newhouse, and Phelps (1977) is the first theoretical paper to study 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 accumulative 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 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 the situation.

The proposed framework is based on a counting process called the Hawkes process. In the framework, the observation unit is a counting process that takes a form of the step function: it is piecewise constant with jumps (of size one) occurred at the



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

time when a doctor visit happened. Such a data structure contains rich information. For example, fix arbitrary time t, the counting process takes a non negative integer that indicates the number of events occurred before current time, hence, is a count data. Meanwhile, the distance between two consecutive jumps is the elapsed time or duration. In our application, this duration is the length of a period since the last doctor visit.

The Hawkes process is state dependent (also known as the self-exciting), i.e., some past events would affect the future ones. In our context, there are two self-exciting channels. The first channel is the episode triggering. A doctor visit might be caused by previous ones. Typical example is the recheck examinations. The second channel is the shadow price effect. The shadow price is determined by the cumulation of the past costs, we would expect that with the shadow price decreasing, an individual would respond to the medical consumption more positively. Nevertheless, not all doctor visits are consequences of past experience, some might occur independently. The Hawkes process is well suited for describing 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 the unique self-exciting property of the Hawkes process, 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 (by measuring the intensity of the independent doctor visits) and the size of a cluster.

The paper is organized as follows. Section 2 describes the data and presents a preliminary that suggests the existence of state dependence. Section 3 discusses the specification of the model in detail, and Section 4 would introduce 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 indicating a sign of serial correlation among doctor visits.

2.1 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. 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.

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[1](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 include services provided by non-physicians such as chiropractors and optometrists, and prescription drugs and supplies. There is no deductible in this insurance contract. When one individual has several claims in one day, we would treat all these claims as a single one and sum the non-covered charges.

[1] In 1973 dollars.

The time unit is annual. 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 time information. 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 as 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.

Some 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 enrolment time. Thus all covariates are time-independent.

2.2 Preliminary Results on Cluster

Apart from the individual heterogeneity, doctor visits might be correlated due to some state dependent effects. The shadow price is one channel if individuals do respond to its variation, another channel would be that past episodes might trigger the occurrence of future ones. We would observe doctor visit clusters if the later channel is indeed valid. In this subsection, we provide a descriptive results that indicate such clusters do exist, and their patterns are different by insurance plans. The pattern are characterized by the number of clusters, and the average number of event within each cluster.

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 (Ester et al., 1996). 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 A. The ability of this algorithm 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 $minPts = 2$[2]. 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 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

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

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

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 into a Hawkes process and the structure of this process. We then focus on the free insurance plan, where there is no individual expenditure in the model. For outpatient doctor visits, some of them would occur independently, while others might be triggered by previous episodes. Hence one could visualize the doctor visits as clusters, where inside each cluster there is one parent event that occurs independently, and the rest of the events are children and grandchildren of that event. Figure 3 illustrates one possible realization. Here, $\{T_1, T_7, T_9\}$ are the

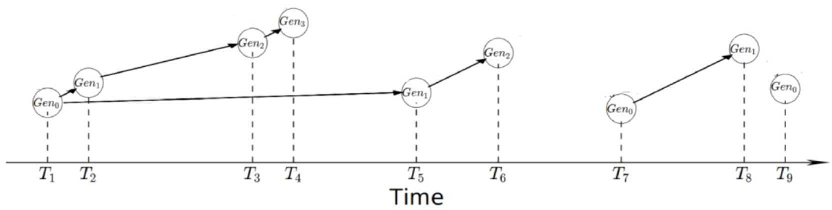


Figure 3: A possible cluster realization

parent events in the first, second and third cluster respectively. Within each cluster, there might be more than one generations of children events. Our model is capable of capturing such a phenomenon. In addition, our model could shed some light on the knowing of the cluster size.

Lastly, we present the model for the cost-sharing plan. From the perspective of econometric modelling, what distinguishes the cost-sharing plan from the free plan is the introduction of individual expenditures 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 one 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 the occurrence times of the events. From following figure one can conclude that the counting process contain rich information: It not only tells how many events has occurred so far, as by fixing a time t, the function $N_i(t)$ is a count number, but also the exact occurrence times of each event, since the occurrence time can be found as

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

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

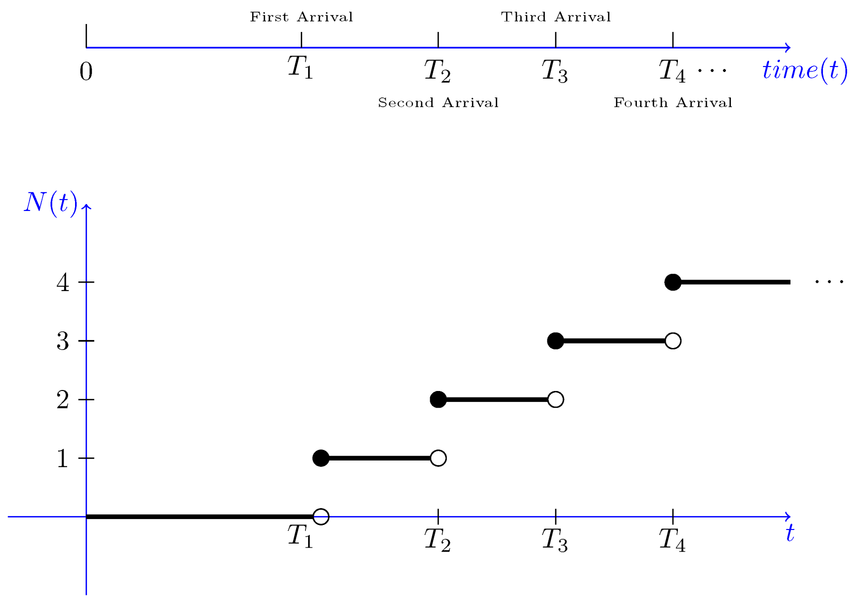


Figure 4: A possible counting process

Doob-Meyer decomposition result:

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

or

$$dN_i(t) = d\Lambda_i(t) + dM_i(t)$$

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 a given filtration $\mathcal{F}_i(t)$, and therefore, trend-free. The $\Lambda_i(t)$ '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 a probability point of view, a intensity function conditional on a filtration $\mathcal{F}_i(t-)$ measures the instantaneous conditional probability of the occurrence of an event:

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

where the filtration $\mathcal{F}_i(t-)$ contains information up to a time just before t. In practice, we observe $N_i(t)$ while $\lambda_i(t)$ may depend on unknown parameters.

The choice of the filtration $\mathcal{F}_i(t-)$ is important in the context of the counting process analysis. If the filtration include a sigma-filed generated by the process itself, i.e.,

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

then the corresponding counting process is called the self-exciting process. The Hawkes process (Hawkes, 1971) is a well known self-exciting process, and is characterized as:

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

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

In some literature, $g : \mathbb{R} \to \mathbb{R}^+$ is called the memory kernel. A popular kernel specification is the exponential function:

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

The Hawkes process is originally proposed to study earthquakes but soon generalizes to other areas such as the finance(Bowsher, 2007) and criminology(Mohler et al., 2012). The construction of this process involves a *branching* mechanism: first, a Poisson process with rate ν independently generates *immigrants*. 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. Such cluster structure is governed by the self-exciting function $g(t)$, it also measures the distance between parent point and child point. 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 ν, while the $N^1(t)$ 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 the existing episodes.

3.2 A Model for the Free Insurance Plan

We begin our description of the model by first focusing on a canonical individual. 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

$$\lambda(t) = \nu, \quad 0 \leq t < T_1 \tag{3}$$

When $t > T_1$, her intensity becomes

$$\lambda(t) = \nu + \int_0^t g(t-s)dN(s) \tag{4}$$

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

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

In our case, we specify $g(t-t_{ij}) = \alpha \exp(-\mu(t-t_{ij}))$ with $\mu > 0$. The exponential exciting function is a popular specification Embrechts et al. (2011); Hawkes (1971). When a Hawkes process is stationary, its cluster pattern can be summarized by a *branching ratio* n^*. Suppose our above specification is stationary, then in the long run, our intensity would eventually reaches a steady value, say, λ^*:

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

Hence,

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

The branching ratio is defined as $n^* = \int_0^\infty g(t)dt$. The clearly, the above result is well defined only if the branching ratio $n^* < 1$. In our canonical Hawkes process, the stationary condition is:

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

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

Specifically, the cluster pattern of a Hawkes process is determined by n^*, and it can be interpreted as the average number of children per event. To see it, suppose the sampling size is normalized to one, then the term $\lambda^*(t)dt$ is the proportion of the events sampled. Among these sampled events, there are νdt parent events generated by the Poisson process with rate ν. Thus, we have νdt families, 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 this cluster N_∞ can also be defined as:

$$N_\infty = \sum_{i \geq 1} A_i \tag{5}$$

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 drive:

$$A_i = A_0 \left(\tilde{n}\right)^i = \left(\tilde{n}\right)^i, i \geq 1 \tag{6}$$

$$N_\infty = \frac{1}{1 - \tilde{n}} \tag{7}$$

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. In our specification, $N_\infty = \mu/(\mu - \alpha)$.

Next, let's consider a specific individual i. When $t < T_1$, 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 we assume $\mathbb{E}\varepsilon_i = 1$, and z_i is a vector of observed individual covariates, and $\phi(z_i) = \exp(z_i^\top \gamma)$. $\nu_i = \nu\varepsilon_i$ is unobserved, and may represent this individual's health status or a summary of the state dependent effect before time 0. When $t > T_1$, 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 separate 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 later, and 2) for a better distinguishing between the unobserved heterogeneity effect and the state dependent effect. Similar additive distinguish can also been found in Kopperschmidt and Stute (2013).

Notice that for the intensity function of individual i, the underlying process, depending on the observed heterogeneity $\phi(z_i)$, may not be stationary.

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

$$\lambda^{P0} = \exp(z_i^\top \gamma)\nu_i + \exp(z_i^\top \gamma) \sum_{j:t_{ij}<t} \alpha \exp(-\mu(t-t_{ij})) \tag{8}$$

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 the 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 $X(t)$, the other major individual expenditure comes from 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 self-exciting 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 an external shocks to the interested process.

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

$$\lambda(t) = b\nu, \quad t < T_1 \tag{9}$$

and

$$\lambda(t) = b\nu + b\int_0^t h(X(s))g(t-s)dN(s) \tag{10}$$

where b is the cost sharing effect. Like before, $g(t-t_i) = \alpha\exp(-\mu(t-t_i))$ with $\mu > 0$. In addition, we specify $h(X(t)) = \exp(\beta_1 X(t))$ with $\beta_1 > 0$. A few words on the above specification is in order. Recall the shadow price is defined as $P_i^s(t) = \mathbb{E}(P_{EOY} \mid X(t)) = 1 - V(X(t))$. An increasing of the bonus function $V(X(t))$ leads to the decreasing of the shadow price, and ultimately increasing the intensity of the doctor visit process. The term $\exp(\beta_1 X(t_i)) \propto V(X(t_i))$ measures an individual's reaction to such a shadow price and we would expect that $\beta_1 > 0$. The multiplicative structure of $\exp(\beta_1 X(t_i))\alpha exp(-\mu(t-t_i)$ reflects the assumption that individuals are partially influenced by the shadow price (Aron-Dine et al., 2015; Brot-Goldberg et al., 2017), i.e., individuals would not behave fully rational, forward-looking and consider the shadow price at all time. Rather, the shadow price would have different impacts on an individual depending on the elapsed time from an episode. For a given occurrence time t_i, we assume that the intensity would jump at this time but will gradually decrease to a plateau until the next event time. 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 will become more and more myopic as time goes by until the next medical spending. The parameter $-\mu < 0$ describes how fast an event is forgotten.

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, often researchers would assume the 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 marks $X(t)$ is state dependent.

One workaround could be the following. Let the partition of the time line be:

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

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 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:

$$\lambda(t) = b\phi(z_i)\nu\varepsilon_i, \quad t < T_1 \tag{11}$$

and

$$\lambda(t) = \phi(z_i)b\left(\nu_i + \int_0^t h(X_i(s))g(t-s)dN_i(s)\right) \tag{12}$$

$$= \phi(z_i)b\left(\nu_i + \sum_{j:t_{ij}<t} h(X_i(t_{ij}))g(t-t_{ij})\right) \tag{13}$$

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:

$$\lambda^{P95} = b\exp(z_i^\top\gamma)\nu_i + b\exp(z_i^\top\gamma)\sum_{j:t_{ij}<t}\exp(\beta_1 X(t_{ij}))\alpha\exp(-\mu(t-t_{ij})) \tag{14}$$

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 the self-exciting 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

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

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 $X_i(t)$ might have several sources, for example, besides the doctor visit fees, the drug purchase accounts for significant part of the individual expenditure. Thus, a drug purchase might happen between two consecutive doctor visits, making the intensity function admits jump. For the purpose of self-contained, 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 process 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,

$$\bar{N}_n = \frac{1}{n}\sum_{i=1}^{n} N_i; \bar{\Lambda}_{v,n} = \frac{1}{n}\sum_{i=1}^{n} \Lambda_{v,i} \tag{15}$$

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

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

If, for μ, 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,

$$\theta_n = arg \inf_{\theta \in \Theta} ||\bar{N}_n - \bar{\Lambda}_{\theta,n}||_{\bar{N}_n} \tag{16}$$

Kopperschmidt and Stute (2013) have shown the consistency and asymptotic normality of this minimum distance estimator. Specifically, for the consistency result, let $\Theta \in \mathbb{R}^d$ be a bounded open set and for each $\epsilon > 0$, we assume,

$$\inf_{||\theta - \theta_0|| \geq \epsilon} ||\mathbb{E}\Lambda_{\theta_0} - \mathbb{E}\Lambda_\theta||_{\mathbb{E}\Lambda_{\theta_0}} > 0 \tag{17}$$

$$\textit{The process}(t, \theta) \to \Lambda_\theta(t) \textit{ is continuous with probability one} \tag{18}$$

Then

$$\lim_{n \to \infty} \theta_n = \theta_0 \textit{ with probability one} \tag{19}$$

The first condition is a weak identification condition, while 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

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

a matrix-valued function, where T denotes transposition, $E = [\underline{t}, \bar{t}]$. And suppose (17) and (18) hold, furthermore, assume that

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

for all θ in a neighborhood of θ_0, 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 $n \to \infty$

$$\sqrt{n}\Phi_0(\theta_0)(\theta_n - \theta_0) \to \mathscr{N}_d(0, C(\theta_0)) \tag{20}$$

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,

$$\Phi_n(\theta) = \frac{\partial}{\partial\theta}\int_E (\bar{\Lambda}_{\theta,n}(t) - \bar{\Lambda}_{\theta_0,n}(t))\frac{\partial}{\partial\theta}\bar{\Lambda}_{\theta,n}(t)^T\bar{\Lambda}_{\theta_0,n}(dt) \tag{21}$$

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,

$$\Phi_n(\theta) = \Phi_0(\theta) + op(1) \tag{22}$$

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 $\mathscr{N}_d(0, C(\theta_0))$, 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) only provided the theoretical results, Monte Carlo simulations that study the performance of this estimation method are conducted by Li and Delgado (2021).

4.2 Identifying the Model

Since the seminal works of (Heckman, 1978, 1981), an important part of the analysis is to discover the extent to which dynamic is due to true state dependence or to unobserved individual heterogeneity. Such an analysis is obviously relevant to our work, and this subsection is dedicated to this topic. To streamline the presentation, we discuss the model identification based on free plan's intensity function, similar argument could apply to the cost sharing plan's intensity effortlessly.

Recall our estimation method is based on the Doob-Meyer decomposition result, i.e., the objective function is the distance between the 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

$$\tilde{\Lambda}_{\theta,n} - \bar{\Lambda}_{\theta,n} = \frac{1}{n}\sum_{i=1}^{n} \eta_i = o_p(1) \tag{23}$$

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 the strictly positive of u, we would write $u = \exp(k)$.

5 Main Results and Robustness Check

5.1 Main Results

Like Keeler and Rolph (1988), we assume that there are no interactions between the shadow price effect and the effects of other explanatory variables. Thus, we might use the free plan data 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 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 the average number of total doctor visits in a cluster of a free plan is approximated as $N_{\infty}^{P0} = \hat{\mu}/(\hat{\mu} - \hat{\alpha}) \approx 5.8$. As mentioned before, the cluster size for a cost-sharing plan is hard to estimate. 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, for a canonical individual, 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 comparing with the free plan, the cluster number in the cost sharing plan decreases $(1 - b) = 34\%$.

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 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. Let the count data Y_t be the number of events occurred during before time t. Let scalar z_j denotes 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 $\phi(z)$. That is, for example, if $\hat{\gamma_j} = 0.2$, $\tilde{\Lambda}_n(t|Z) = 2.5$, 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.* The overall effect for age is as follows: at first, the intensity will decrease as age increases, after one passes the age of 41, the intensity and age are positively correlated. It is well-known that the youngsters are more risky compared to their 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.
- *Education.* Education is a significant factor, the result, by and large, suggests

Table 2: Basic Results

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

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

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 not surprising. A higher income gives individuals the ability to cover the opportunity cost related to the absence from work (to visit a doctor).

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 such a shadow price enters the intensity additively. This setting implicitly assumes that an individual would consider the shadow price all the time. Specifically, we assume that:

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

where

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

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

Table 3: Robustness Check Results

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

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

are still able to find evidence that rejects the hypothesis that individuals only respond to the spot price.

5.2.2 Non-Stationary Intensity Setting

In the main model, we assume that a 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 restriction where $n^* = 1$ or $\alpha = \mu$.

In the literature, such a setting is often referred as the critical regime (Brémaud and Massoulié, 2001). It corresponds to a situation where one cluster lives indefinitely without exploding. In the context of our application, it implies that all the doctor visits belong to one family, there would be a 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 want to emphasize that under this restriction, we still find evidence that individuals would respond to the shadow price and do understand to some degree the nature of the non-linear contract.

Table 4: Critical Regime Results

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

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

6 Discussion

In this section, we would discuss the needs to use the counting process framework from an econometric point of view. In addition, we would discuss the 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, the number of doctor visits $N_i(T)$ for each individual also varies significantly. Apart from the time-invariant individual heterogeneities, the state dependent effect that consists of history information is one important reason for such variation. We could 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 $(T_{i1}, \ldots, T_{i(N_i)})$, where $N_i = N_i(T)$ is the random variable representing the number of doctor visits. The joint density of these occurrence times are

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

The conditional p.d.f of T_{ij} can be derived as follows. 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. Therefore,

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

and the conditional C.D.F of T_{ij} is

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

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,

$$f_{T_{i1},\dots,T_{i(N_i)}}(t_{i1},\dots,t_{i(N_i)}) = \exp\left(-\int_0^T \lambda_i(t \mid \mathcal{F}(t-))dt\right)\prod_{j=1}^{N_i}\lambda_i(t_{ij} \mid \mathcal{F}_i(t_{ij}-)) \quad (24)$$

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 due to the way conditional intensity varies with $\mathcal{F}_i(t-)$. Notice here the random variable $N_i = N_i(T)$ is included in the filtration since $\sigma(N_i(s) : s < T) = \sigma(t_{i1}, \dots, t_{i(N_i)}, N_i(T-) = N_i) \subset \mathcal{F}_i(T-)$. The resulting counting process is often called doubly stochastic.

A proper econometric model should fully reflect this doubly stochastic nature of our data. Conventional dynamic econometric frameworks, however, often fail to recognize the randomness of $N_i(T)$. For example, although we could represent the doctor visits in terms of a panel of durations $d_{ij} = t_{ij} - t_{i(j-1)}$, without a proper sample selection mechanism, 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) also does not consider this randomness, as 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. The random variable N_i is nothing but the value of the corresponding counting process at the terminal time: $N_i = N_i(T)$, whose expectation is described by the cumulative intensity function: $\mathbb{E}N_i(T) = \Lambda_i(T \mid \mathcal{F}_i(T-))$. 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 (2007).

Lastly, we discuss the estimation method. The first reason we use an estimation method based on the conditional intensity is that the doubly stochastic property brings challenges to building a likelihood based estimation method. To see it, we need to represent the probabilistic structure of the counting process in terms of its finite dimensional distrbutions (or fidis). 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:

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

for $j = 1$ and

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

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 24 yields

$$f_{T_{i1},\ldots,T_{i(n_i)}}\left(t_{i1}, \ldots, t_{i(n_i)}\right) = P\left(N_i(T) = n_i \mid t_{i1}, \ldots, t_{i(n_i)}\right) f_T^{(n_i)}(t) \tag{25}$$

This is the likelihood contributor described in Heckman and Walker (1990). It is conditional on a fixed number of events n_i and hence ignores the doubly stochastic property. A direct consequence of using likelihood contributor like Equation 25 is sample selection: 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. We specify and estimate a dynamic model using the Hawkes process framework, and have shown that an individual do respond to the shadow price introduced by the nonlinear health insurance. The nonlinear plans are norm in health insurance, yet most existing literature on the medical consumption tends to assume that individuals only respond to the spot price.

In our framework, the unit of an observation is a Hawkes process, which, by definition, is conditional on a filtration that is generated by the process itself. It allows researchers to take historical information into the model. In addition, its filtration could include external shocks such as an individual's expenditure over time. One key feature of Hawkes process is its ability to describe cluster patterns. A cluster consists of an independent doctor visit and follow up visits that are offsprings of this visit.

Using the classical RAND Health Insurance Experiment data, we first provide some descriptive evidence on the existence of cluster structures of an individual's doctor visit records under both free plan and cost sharing plan. The results suggest that the outpatient consumption is subject to the state dependent effect. We then specify and estimate a Hawkes process of doctor visit process. 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) and we do not view our results as particularly surprising. In addition, for a canonical individual, we found that comparing to the free plan, the number of doctor visit clusters would decrease 34% in a 95% co-insurance rate plan, and during the initial period of the contract, the cluster size under the cost sharing plan shrinks 62%.

References

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

Aron-Dine, A., L. Einav, A. Finkelstein, and M. Cullen (2015): "Moral hazard in health insurance: do dynamic incentives matter?" *Review of Economics and Statistics*, 97, 725–741.

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

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

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

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

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

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

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

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. (1996): “A density-based algorithm for discovering clusters in large spatial databases with noise.” in *Kdd*, vol. 96, 226–231.

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

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*, JSTOR, 227–269.

——— (1981): “Heterogeneity and state dependence,” in *Studies in labor markets*, University of Chicago Press, 91–140.

Heckman, J. J. and J. R. Walker (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.

Hu, J., F. Wang, J. Sun, R. Sorrentino, and S. Ebadollahi (2012): “A healthcare utilization analysis framework for hot spotting and contextual anomaly detection,” in *AMIA annual symposium proceedings*, American Medical Informatics Association, vol. 2012, 360.

Keeler, E. B., J. P. Newhouse, and C. E. Phelps (1977): “Deductibles and the demand for medical care services: The theory of a consumer facing a variable price schedule under uncertainty,” *Econometrica*, 641–655.

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

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

Li, Y. and M. A. Delgado (2021): "The Specification of Dynamic Two State Models in Continuous Time," *Wuhan University Working Paper*.

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

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

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

Snyder, D. L. and M. I. Miller (2012): *Random point processes in time and space*, Springer Science & Business Media.

A 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.