Modelling Dynamic Moral Hazard in Individual Outpatient Claims

Yuhao Li LIYUHAO.ECON@OUTLOOK.COM
Economics and Management School
Wuhan University

Rui Cui RCUI@FEM.ECNU.EDU.CN
Faculty of Economics and Management
East China Normal University

Draft Time: January, 2023.

Abstract

Typical health insurance contracts have cost-sharing policies (deductibles, coinsurance rates and caps on out-of-pocket expenditure). These policy designs create dynamic incentives for moral hazard: with each consumption, the gap between the deductible (and the out-of-pocket cap) decreases, and the health care price is effectively cheaper. This paper aims to model the dynamic moral hazard using a Hawkes process framework. A Hawkes process is a counting process with self-exciting and branching features. We build a stylized model to define the dynamic moral hazard in terms of the intensity of a Hawkes process. Based on it, we parameterize the intensity and estimate the reduced form model by a minimum distance estimation method. The dynamic moral hazard in this approach is state-dependent but can be summarized by a single quantity called the branching ratio. We present distributions of the branching ratio for different insurance coverages.

JEL Classification: C41, C51, I12, I13

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

1. Introduction

Well known economic theories suggest that health insurance may increase medical consumption above the socially optimal level (Arrow, 1978; Pauly, 1968). The incentives of this increase are often referred to as the moral hazard (Cutler and Zeckhauser, 2000). Einav et al. (2013) defines the moral hazard as the incremental medical utilization that is due to greater insurance coverage. In this paper, we estimate the *dynamic moral hazard* in outpatient claims using a novel Hawkes process framework. The dynamic feature of moral hazard is introduced by cost-sharing policies of a health insurance plan. These policies often include deductibles, coinsurance rates and out-of-pocket expenditure limits. In a typical

modern health insurance plan, individuals pay the full price of care below the deductible. After exhausting the deductible, individuals pay a portion of the bill equal to the coinsurance rate, and finally, if they reach the limit, there would no cost for further health care for the remainder of the contract year. These cost-sharing characteristics, coupled with the uncertainty of future health care demand, are the primary source of dynamic mechanism behind the moral hazard: with each care consumption, the gap between the accumulated expenditure and the deductible (or the out-of-pocket limit) decreases, making the next consumption effectively cheaper. Thus, an additional benefit of health care utilization today is lower shadow health care prices.

In addition, an outpatient claim itself could also be the source of dynamics. Take the left panel of Figure 1 as an example, where individuals are insured by a fully covered insurance plan [1]. In the figure, the red dots represent outpatient occurrence times in a contract year. Here, we observe that (1) Claims are clustered in time. The cluster is not unique to the individual displayed here, but ubiquitous in the data set. (2) Clustered occurrence times imply the data is unlikely to be stationary. Specifically, durations between claims are not stationary. (3) Since individuals are given a fully covered insurance plan, dynamic incentives arising from the cost-sharing policies are absent. Rather, event-based state-dependency, i.e., previous claims would have effects on future ones, might contribute to this cluster structure.

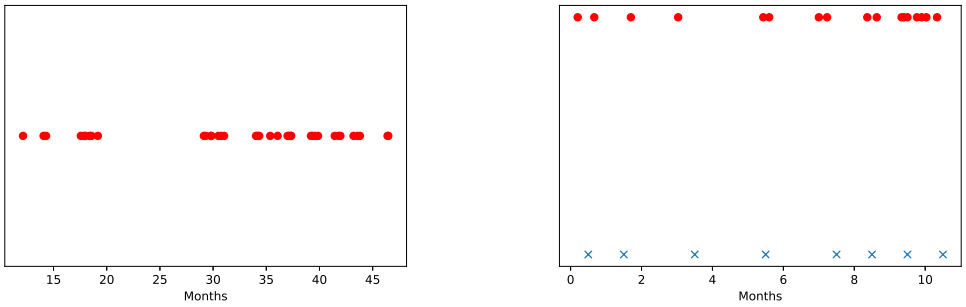
If dynamic moral hazard exists, individuals would react to a shadow price which is lower than the nominal one. Thus, the desired savings thought-to-be-achieved through cost-sharing policies would never be realized, leading to a loss of social welfare. Exploring possible dynamic mechanisms behind the moral hazard is then essential for controlling the size of the health care sector.

In this paper, we focus on the outpatient claim data for two reasons. First, among different types of health care utilization, outpatient activities are arguably most frequent. Existing literature has shown (e.g.,Brot-Goldberg et al. (2017)) that health care utilization reduction (as a consequence of cost-sharing policies) is mainly achieved through quantity rather than price shopping. Thus, it is natural to investigate moral hazard via the number of outpatients. Second, inpatient claims are infrequent and usually come with expenditures that are either close or exceeding the out-of-pocket-expenditure limit. Thus, individuals might lose the primary incentive of dynamic moral hazard once they have experienced an inpatient activity.

Most of empirical studies that tend to estimate moral hazard effects often use aggregate medical care data up to some level (annual level, see, e,g., Cardon and Hendel (2001); Einav et al. (2013); Khwaja (2010), monthly level, see e.g., Cronin (2019), or weekly level, see Diaz-Campo (2022)). However, by aggregating data, part of information is inevitably lost, and the corresponding estimation results may not be reliable. The right panel of Figure 1 helps to illustrate the situation. After aggregating the red dots in a monthly level, we obtain

1. See Section 4 for the details about the data.

the blue crosses, which represent the existence of outpatient activities in the corresponding months. We observe that the data is stationary and non-clustered, thus, useful information about state-dependency or moral hazard may also be altered.



(a) Individual 1, free plan

(b) Individual 2, 95% coinsurance rate plan

Figure 1: Claim times

This paper advocates the Hawkes process as an alternative approach to modelling dynamic outpatient insurance claims. Existing applications of the Hawkes process can be found in finance, see Bacry et al. (2015); Bowsher (2007); Chavez-Demoulin et al. (2005), in seismology, see Zhuang et al. (2002), in criminology, see Mohler et al. (2012) and in insurance, see Cheng and Seol (2020); Dassios and Zhao (2012); Jang and Dassios (2013); Stabile and Torrisi (2010); Swishchuk et al. (2021); Zhu (2013).

The benefits of using the Hawkes process and our contribution to the literature can be summarized as the followings:

- **A non-aggregating data model.** Non-aggregating data permits researchers to perform an in-depth analysis on outpatient activities, however, the irregularity nature of raw data makes hard to construct a meaningful model. The Hawkes process is a counting process consisting of all occurrence times of events, thus, this approach provide a natural framework to analysis non-aggregating data related to times.

- **A dynamic reduced form model.** Dynamic insurance data could help identify moral hazard, see e.g., Abbring et al. (2003a,b). Yet existing literature often adopts the dynamic programming approach or the dynamic discrete choice framework, both are derived from a structural model (hence, researchers have to impose explicit restrictions on individual preferences and utility functions), and are based on the aggregated data, see e.g., Cronin (2019); Diaz-Campo (2022). The Hawkes process is a self-exciting process, i.e., the intensity rate of an occurrence is conditional on a sigma-algebra generated by the process up to current time, this makes the Hawkes process ideal for modelling dynamic mechanisms in a reduced-form approach.

- **A dynamic moral hazard measurement.** In a similar manner to Einav et al. (2013), we define the moral hazard as the increment of the intensity of medical utilization. What differs from the literature is that we focus on the increment of the number of medical utilization events rather than the increment in terms of the monetization. There are two reasons to do so. First, outpatient activities have, on average, lower cost but higher frequency compared to inpatients. The moral hazard risks are more visible in terms of counts than in terms of monetization. Second, it is easier to model different dynamic mechanism when the utilization is presented as outpatient events. In this paper, we consider two sources: one arises from the shadow coinsurance rate, and the other arises from a state dependent effect, i.e., past events would affect the future ones.

- **A n-observation Hawkes process model.** In Hawkes process literature, researchers usually specify a Hawkes process to model one observation process. For example, in finance, it is bid or ask times for one stock; in seismology, it is earthquake occurrence times in a region; in criminology, it is crime events in one area; and in insurance, it is the ruin of one insurance company. This study differs from the existing literature in the sense that instead of one observation, we have n-observations, where each observation process is a realization of outpatient activities of one individual. As a consequence, we need to model two sources of variations: (1) variations among occurrence times and (2) variations of the number of events for a given period of time. Thus, instead of the commonly used maximum likelihood based estimation method, we use a minium distance estimation method.

Our paper is related to several strands of literature. First, it contributes to the sparse literature that test whether individuals respond to dynamic incentives by nonlinear health insurance contract, see, e.g., Aron-Dine et al. (2015); Einav et al. (2015); Guo and Zhang (2019); Keeler and Rolph (1988). The closest to our paper is Diaz-Campo (2022) which models the dynamic moral hazard using weekly aggregated expenditure data in a dynamic programming framework. In Section 7, we explicitly compare and discuss the strategy adopted in Diaz-Campo (2022), and highlight some advantages of our approach. Our paper is also related to the literature that construct models by optimizing an intensity function. Abbring et al. (2003a) studies adverse selection and moral hazard in car insurance. They optimize a utility model by intensity, and later estimate the intensity model via maximum likelihood methods.

The paper is organized as follows. In section 2, we introduce a stylized model of health care utilization. Section 3 introduces necessary concepts about the Hawkes process. In section 4, We describe the data and the sample construction procedure. Section 5 presents the econometric specification, and section 6 provides interpretations and results. In section

7, We compare our approach with the commonly used dynamic programming framework, and discuss some limitations of the new framework. Section 8 concludes.

2. A Model of Utilization

In this section, we introduce a stylized model of individual utilization. The model allows us to precisely define the object that we focus on, "dynamic moral hazard". It also provides main ingredients in our subsequent econometric specification exercise. The model has a similar specification structure to that of Einav et al. (2013), but the underlying concept is different. Specifically, there are three differences: First, individuals are optimizing different objects. In Einav et al. (2013), a representative individual needs to find an optimal monetized health care utilization to maximize his/her utility; while in this paper, the same individual would find an optimal conditional intensity of health care utilization. Here, given past events, the conditional intensity captures the instantaneous rate of the occurrence of an outpatient event. Second, in our model, an individual is making utilization decisions continuously, while in Einav et al. (2013), an individual makes a one-time utilization decision at the beginning of a period. Third, our model is a dynamic model in the sense that an individual makes decisions based on all the past information (e.g., past doctor visits, accumulated cost,etc), while in Einav et al. (2013), history does not play a role.

2.1 Utilization Choice

Since this is a model of individual behavior, we omit i subscripts to simplify notation throughout this section. We assume that the individual's health care utilization decision is made in order to maximize a trade-off between health gain and (non-monetized) marginal cost. Specifically, we assume that the individual's utility is separable in health $h(\cdot)$ and marginal cost $y(\cdot)$, and can be written as

$$u(a_t, \lambda_0) = h(a_t; \omega_t(c_t)) - y(a_t - \lambda_0)$$

Naturally, $h(\cdot)$ is increasing with a_t, where $a_t \geq 0$ is the intensity, i.e., the instantaneous conditional rate of having an outpatient activity given past events. For a given time interval, the higher the intensity, the greater number of outpatient activities will be. The second argument in the health function, $\omega_t(c_t)$, describes the sensitivity of the outpatient activity at time t, i.e., how responsive health care utilization decisions are to insurance coverage and other dynamic mechanism at time t. Here c_t is the insurance coverage at time t. $y(a_t - \lambda_0)$ is the non-monetized instantaneous marginal cost, we further assume it is convex: it is decreasing for low levels of utilization (when treatment decreases the disutility, and the individual's willingness to consume increases) and is increasing eventually (when there is no further health benefit from treatment and time costs dominate). Finally, we denote λ_0

as the background risk that is identical to all individuals (e.g., public security, food safety, pandemics, etc).

There are several good reasons to focus on optimizing the intensity a_t rather than a monetized health care consumption. The bulk of evidence suggests that introducing cost-sharing tools reduces medical spending. More specifically, the reduction is achieved mainly through quantity whereby individuals purchase fewer medical care services, instead of price shopping whereby individuals search for cheaper providers without compromising the quantity (Brot-Goldberg et al., 2017). In addition, costs of medical services are conventionally assumed to be i.i.d log-normal, see Handel et al. (2015); Keeler and Rolph (1988). Thus, assessing an individual's response to a cost-sharing policy amounts to assessing how this individual adjusts her outpatient quantities.

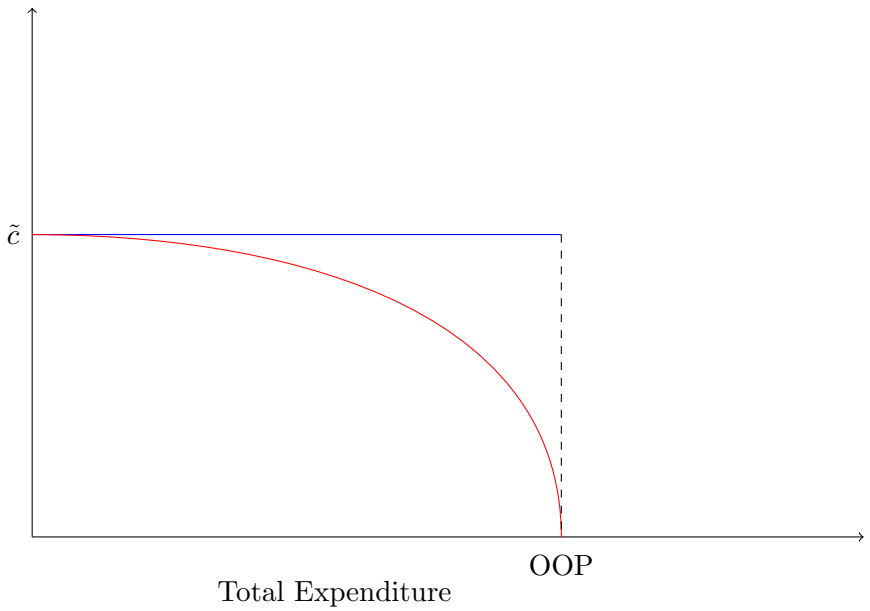
More explanation about the cost function is also needed. First, $c_t = c(x(t))$ represents the insurance coverage, and is a function of the accumulated expenditure $x(t)$. This coverage is often characterized by cost-sharing policies (e.g., deductible, coinsurance-rate and out-of-pocket cap). In this study, we focus on the coinsurance rate. There are two types of coinsurance, as illustrated in Figure 2. The blue straight line represents the nominal coinsurance rate $\tilde{c}$ applicable when total expenditure is below the OOP, after which, this rate reduces to zero. Whereas the red curve is the individual's shadow coinsurance rate, which will represent the insurance coverage here. We further assume that the shadow coinsurance rate c_t is concave with $c' < 0, c'' < 0$. Effectively, we are assuming that individuals would respond more aggressively as the total expenditure approaches to the OOP limit.

Historically, literature studying the price elasticity of demand for health care (i.e., the moral hazard, Cutler and Zeckhauser (2000)) often assumes that individuals only respond to the 'spot' price, see Cutler and Zeckhauser (2000); Keeler and Rolph (1988); Manning et al. (1987). Recent literature deviates from this assumption, and assume that individuals might respond to the 'shadow price' as well, see Aron-Dine et al. (2013); Brot-Goldberg et al. (2017); Einav et al. (2015). We also adopt this deviation, and define the 'shadow coinsurance rate' for the individual as:

$$c_t = c(x(t)) = \mathbb{E}(c_{EOY} \mid x(t))$$

where c_{EOY} is the *nominal* coinsurance rate at the end of a year.

The function $\omega_t(\cdot) = \omega(c_t, \tau_{[0,t]})$ has two arguments, which represent two sources of dynamic. First, $\tau_{[0,t]} = \sigma(\{t_j\} : t_j < t) = \sigma(N(s) : s < t)$ is a sigma-filed generated by all occurrence times below time t. When fixing a coverage level (e.g., $c = 0$, the full coverage), the individual might have different responsiveness towards health care utilization. For example, due to previous doctor visits, the individual has better understanding of his/her health condition, which might change the attitudes to some preventive health care utilization, resulting an increasing of doctor visit frequency. Second, when fixing $\tau_{[0,t]}$,



The total expenditure is the sum of individual spending and expenditures paid by the insurance. When the total expenditure is below OOP, both types of coinsurance rate are above zero. The nominal coinsurance rate (blue line) remains constant, while the shadow coinsurance rate (red curve) decreases as the total expenditure increases. Whenever the total expenditure is beyond OOP, there is no cost for individuals.

Figure 2: Two Types of Coinsurance Rate

$\omega_t(c_t)$ measures how the individual changes his/her utility with respect to the change of the shadow coinsurance rate. This dynamic is of primary interest (and we use a notation that highlights the role of c_t). We assume that $\omega_t(c_t)$ is decreasing with c_t: the cheaper the coinsurance rate, the greater the incentive of moral hazard. We defer the parametric specification of $\omega_t(c_t)$ in Section 5.

The utility function is parameterized as:

$$u(a_t, \lambda_0) = a_t \cdot \omega_t(c_t) - \frac{1}{2}(a_t - \lambda_0)^2 \tag{1}$$

Given this parametrization, the optimal instantaneous utilization is given by

$$a_t^*(\lambda_0, \omega_t(c_t)) = \arg\max_{a_t \geq 0} u(a_t, \lambda_0) \tag{2}$$

The first order condition implied by Eq. (1) is given by

$$a_t^*(\lambda_0, \omega_t(c_t)) = \max[0, \lambda_0 + \omega_t(c_t)] \tag{3}$$

2.2 Interpreting $\omega_t(c_t)$ as Dynamic Moral Hazard

Abstracting from the potential truncation of utilization at zero, the individual will optimally choose $a_t^* = \lambda_0 + \omega_t(1)$ with no insurance, and $a_t^* = \lambda_0 + \omega_t(0)$ with full insurance coverage. Thus, $\omega_t(\cdot)$ can be thought of as the incremental utilization due to the change in shadow coinsurance rate. Following Einav et al. (2013), we can interpret this model as that $\lambda_0 + \omega_t(1)$ represents non-discretionary instantaneous health care shocks that the individual will pay to treat, regardless of insurance. In addition, there exist discretionary health care utilizations that will not be undertaken without insurance. With insurance, some amount of this instantaneous discretionary care will be consumed, and the increment is $\omega_t(c_t) - \omega_t(1)$. This interpretation also suggests that, after normalizing $\omega_t(1) = 0$, $\omega_t(c_t)$ should be understood as the moral hazard.

Finally, we highlight again that $\omega_t(c_t)$ contains history information, and thus, is dynamic.

3. The Hawkes Process

As mentioned before, we will focus on the quantity of the outpatient consumption. Most literature would adopt some count data regression models to study the quantity. However, these models are based on aggregated data, and hence ignoring the dynamics among events. In this paper, we opt to the counting process approach. Specifically, we assume the optimal intensity derived in the previous section follows a Hawkes process. In this section, we will briefly introduce some related concepts about this process. As before, we will omit i subscripts to simplify notation, but will restore them whenever necessary.

3.1 Backgrounds on Hawkes Process

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

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

where $\mathbb{I}\{A\}$ is an indicator with the value equal to one if an event A occurred, and zero else wise. $N(t)$ is fully characterized by its conditional intensity function, for $t_{j-1} < t \leq t_j$:

$$\begin{aligned} a(t)dt &= a(t \mid \mathcal{F}(t-))dt \\ &= \Pr\left(t_j \in [t, t+dt) \mid \mathcal{F}(t-)\right) \end{aligned}$$

which specifies the probability that an event occurs in the infinitesimal time interval $[t+dt)$ given past events $\mathcal{F}(t-) \supseteq \sigma(N(s) : s < t)$.

$N(t)$ is a Hawkes process if its intensity function is specified as

$$a(t) = \lambda_0 + \int_0^t g(t-s)dN(s) \tag{4}$$

$$= \lambda_0 + \sum_{j:t_j<t} g(t-t_j) \tag{5}$$

where λ_0 is a time-invariant parameter, and $g(\cdot)$ is called the self-exciting kernel. One popular kernel specification is the exponential function (Embrechts et al., 2011; Hawkes, 1971): $g(t) = \alpha \exp(-\mu t), \quad \alpha, \mu > 0$. Note that for $g(t) = 0$ the model reduces to a Poisson process with constant intensity λ_0.

The specification of the Hawkes process fits well with our optimal intensity model derived in the previous section. For the free insurance plan P0, although the insurance coverage is fixed at $c_t = 0, \forall t \in \mathcal{T}$, the moral hazard is still dynamic:

$$\omega_t(0) = \sum_{j:t_j<t} g(t-t_j)$$

This specification highlights the effects of previous outpatient activities, as the individual might update his/her information from past experiences, and transforms some discretionary health care consumptions to non-discretionary consumptions.

As for the cost-sharing plan P95, we use a marked Hawkes process (Daley and Vere-Jones, 2007) to model the dynamic moral hazard, where the shadow coinsurance rate works

as marks:

$$a(t) = \lambda_0 + \omega_t(c_t) = \lambda_0 + \sum_{j:t_j<t} (1 - c(x(t_j)))g(t - t_j)$$

As before, $x(t)$ is the accumulated medical expenditure so far. The exact specification of $g(\cdot)$ and $c(x(t))$ will be deferred to the econometric specification section.

By taking expectation of both sides of Eq. (5) and assuming stationarity (i.e., a finite average event rate $\mathbb{E}a(t) = \kappa$), we can express the average event rate of the process as $\kappa = \lambda_0/(1-n^*)$ where $n^* = \int g(\tau)d\tau$. One can create a direct mapping between the Hawkes process and the well known branching process (Harris et al., 1963) in which exogenous 'immigrant' events occur with an intensity λ_0 and may give rise to m additional endogenous 'offspring' events, where m is drawn from a Poisson distribution with mean n^*. These in turn may themselves give birth to more 'offspring' events.

The value n^* is called the branching ratio, and determines the behavior of the model. If $n^* > 1$, the corresponding process is non-stationary and may explode in finite time. If $n^* < 1$, the process is stationary.

3.2 n observation process

Suppose in a pre-determined interval $(0, T]$, one observes only one process $N(t)$ consisting of occurrence times $t_1, t_2, \ldots, t_n$. Then, we can fit the Hawkes process by maximizing the log-likelihood (Rubin, 1972) over the set of parameters θ:

$$\log L(t_1, \ldots, t_n|\theta) = -\int_0^T a(t|\theta) + \int_0^T \log a(t|\theta) dN(t) \tag{6}$$

In the case of the exponential kernel, $\theta = \{\lambda_0, \alpha, \mu\}$, the branching ratio estimate is then $\hat{n}^* = \hat{\alpha}/\hat{\mu}$.

However, in our application, we have n observational processes $\{N_i(t) : i = 1, \ldots, n\}$, where the subscripts i denote individuals. In practice, for the time interval $\mathcal{T} = (0, T]$, we observe occurrence times $\{t_{i,1}, \ldots t_{i,n_i}; i = 1, 2, \ldots n\}$, where for individual i, n_i is her number of outpatient activities in $\mathcal{T}$. We call this kind of data the *doubly stochastic data*, as for an individual, both the occurrence times and the number of events are stochastic. The fact that n_i is a realization of a random variable complicates the specification of the log-likelihood function. Since we have n-observations, one has to fix the number of events (say, $\bar{n}$) for each log-likelihood contribution function as $\log L_i(t_{i,1}, \ldots, t_{i,\bar{n}}|\theta)$, so that the overall log-likelihood function is $\log L(\theta) = \sum_{i=1}^n \log L_i(t_{i,1}, \ldots, t_{i,\bar{n}}|\theta)$. However, there are two consequences of adopting this strategy. First, for individuals who belong to $\{i = 1, \ldots, n : n_i < \bar{n}\}$, it is simply impossible to write down the corresponding log-likelihood contributions, and researchers have to remove these individuals from the sample. But this might lead to a sample selection problem. Second, for individuals who belong to $\{i =$

$1, \ldots, n : n_i > \bar{n}\}$, although the corresponding log-likelihood contributions are available, a great part of information (i.e., $\{t_{i,j} : j > \bar{n}\}$) has been throw away, leading to a less efficient estimation.

To overcome these challenges, we adopt a minimum distance estimation method, first proposed by Kopperschmidt and Stute (2013). This method is based on the Doob-Meyer decomposition result:

$$N_i(t) = A_i(t) + M_i(t) \tag{7}$$

where $A_i(t) = \int_0^t a_i(s)ds$ is the cumulative intensity function, also known as the compensator, and $M_i(t)$ is a martingale with zero mean: $\mathbb{E}M_i(t|\mathcal{F}_i(t-)) = 0$.

The estimator $\hat{\theta}_n$ is obtained as:

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

where

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

are the averaged Hawkes process and the averaged compensator, respectively. $\bar{M}_n(t|\theta) = \bar{N}_n(t) - \bar{A}_n(t|\theta)$ is the corresponding residual term.

Under suitable assumptions, Kopperschmidt and Stute (2013) have shown that this estimator is consistent and is asymptotically normal:

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

where

$$\Omega = \Phi_0^{-1}(\theta_0) C(\theta_0) \Phi_0^{-1}$$

Notations in the asymptotic variance matrix are:

$$\Phi_0(\theta_0) = \int_0^T \mathbb{E}\frac{\partial}{\partial\theta} A(t|\theta) \mathbb{E}\frac{\partial}{\partial\theta} A(t|\theta)^\top \mathbb{E}A(dt|\theta_0)\Big|_{\theta=\theta_0}$$

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

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

and

$$\psi(s) = \int_s^T \mathbb{E}\frac{\partial}{\partial\theta}A(t|\theta)\mathbb{E}A(dt|\theta_0)\Big|_{\theta=\theta_0}$$

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

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

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

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

The term $\Phi_0(\theta_0)$ can be estimated in the same way and is omitted here.

We also perform a series of simulation studies to examine the finite sample properties of this estimator. In Appendix A, we describe the data generating process as well as the simulation algorithm. In Appendix B, we present the simulation results.

4. Data and Sample

We use individual-level, line-item records from the RAND Health Insurance Experiment (hereafter, HIE). The RAND HIE is a randomized field experiment of different insurance plans offered to more than 8000 individuals in the U.S. These insured enrollees were assigned to different insurance treatment, and data on their use of health services were collected throughout their period of participation. The insurance treatment differs mostly in terms of cost-sharing policies, i.e., deductibles, coinsurance rates and out-of-pocket caps. Because of the randomness of the assignments and the nonlinear cost-sharing features, the RAND HIE data is particularly suitable for the study of dynamic moral hazard. In what follows, we describe the experimental design and the analysis sample.

Experiment Design. The RAND Corporation conducted the HIE from 1974 to 1982 in six sites across the U.S: Dayton, Ohio; Seattle, Washington; Fitchburg and Franklin County, Massachusetts; and Charleston and Georgetow County, South Carolina. Individuals offered enrollment in the experiment represent a random sample from each site, subject to certain eligibility restrictions. 14 different insurance plans were randomly assigned to an individual in a given site and enrollment date. These plans differ in coinsurance rates, delivery systems, and maximum out-of-pocket expenditures (OOP). The coinsurance rates were set at either 0 (free care), $25, 50$ or 95 percent. 12 plans had a OOP of $5, 10$ or 15 percent of family income in the previous year. The free plan does not impose OOP, and a plan (labeled Plan N in the RAND HIE document) impose s a OOP of 150 dollars per person or 450 dollars

per family. All insurance plans feature a zero deductible, a coverage of length of 12 months, and no premiums. The contract year began on the enrollment date and ended on each anniversary of the enrollment date. There are several enrollment dates at each site, and each contract year may span two calendar years.

Sample. Since we are interested in analyzing dynamic moral hazard in outpatient claims, we focus on outpatient data from free plan and the plan N (the one impose OOP of 150 dollars per person or 450 dollars per family). The choice of the free plan is natural: this is the plan the would exhibits most moral hazard behaviors. The reason we choose the plan N is that this plan would cover 100% inpatient services but pays 5% (i.e., a 95% coinsurance rate) of covered outpatient services until deductible is met. The free plan (hereafter, P0 plan) and the plan N (hereafter, P95 plan) thus only differs in outpatient activities.

In this study we use the fee-for-service (FFS) claims line-item to conduct analysis. Each instance of a billed service on a claim form is called a "line item." The RAND HIE use line-item and other relevant data from claim forms to compose the records. The line-item records were organized into 14 files according to the type of medical service involved. For the purpose of this study, we focus on services rendered by physicians or other health professionals (file 06 in the RAND HIE document). Both P0 and P95 plans cover expenses of prescription drugs and supplies, hence, we also include drug purchase information (file 15 in the RAND HIE document).

The RAND HIE relies on the Medical Expense Report (MER) to collect data. On each MER, providers were asked to itemize all service, and for each give the date, the amount charged, and other related information. Some MERs collected information common to other MERs, and each MER collected information unique to itself. Thus, an episode may related to several health care consumptions via different MERs. Specific to our study, we need to merge all related medical consumption to one item.

We make different restrictions to create analysis samples for P0 and P95 plans. For the P0 plan, the main restriction is the age of insured enrollees. We exclude individuals who are younger than 18 and older than 60. The main reason to do so is that the excluded individuals would exhibit different responses to the moral hazard due to health conditions. In the P95 plan, in addition to the age restriction, we also exclude any claims that do not occurred in the contract year 1977-1978. This is because the insurance contract in this plan would set back to default at every anniversary of the enrollment date. While in the free plan, there is no cost sharing restrictions, there is no need to emphasize the 'reset to the default'. Table 1 provides details about the remaining sample size as well as the number of line-item after sequentially applying each exclusion restriction.

The time unit is annual. For example, if an insurance contract begins on Jan-01-1977 and the date of a doctor visit is Oct-01-1977, the time stamp is then 0.748 (years). The demographic covariates included in the model are age, sex, education (in terms of schooling years) and log-income. For simplicity, we fixed all ages at the enrollment time. Thus, all

Table 1: Sample Construction Procedure

P0 Plan		
Steps	Sample size	Line-item size
Outpatient Claims rederned by physicians	6151	172157
Only include individuals with $18 \leq$ age ≤ 60	3442	129760
Select individuals enrolled in the P0 plan	1001	54940
Merge line-item associated with a same episode	1001	28132
P95 Plan		
Steps	Sample size	Line-item size
Outpatient Claims rederned by physicians	6151	172157
Only include individuals with $18 \leq$ age ≤ 60	3442	129760
Focus on the contract year 1977-1978	2422	51306
Select individuals enrolled in the P95 plan	364	4996
Merge line-item associated with a same episode	364	2246

covariates are time-independent. Other data cleaning assumptions include (1) if the value of a doctor visit cost is not available, we replace it with zero; and (2) if information on the education is unknown, we replace it with the average education level.

5. The Econometric Specification

This section introduces our econometric specification. We will parameterize intensities for both insurance plans. The intensity specification for the free plan P0 will emphasize the dynamics that give rise to the increments of health care consumption. The source of a surge of consumption in a free plan could be the more informed health conditions. In the cost-sharing plan P95, we will specify the shadow coinsurance rate $c_t = c(x(t))$ as well as the dynamic moral hazard $\omega_t(c_t)$. The shadow coinsurance is a function of the accumulated medical expenditure, while the dynamic moral hazard, although depends on the individual's outpatient history, can be summarized by the branching ratio n^*.

5.1 Model for Free Plan P0

For an individual i, we specify his/her intensity as:

$$a_i^{P0}(t) = \exp(\lambda_0) + \alpha \exp(z_i^\top \gamma) \sum_{j:t_{ij}<t} \exp(-\mu(t - t_{ij})), \quad t \in \mathcal{T} \tag{8}$$

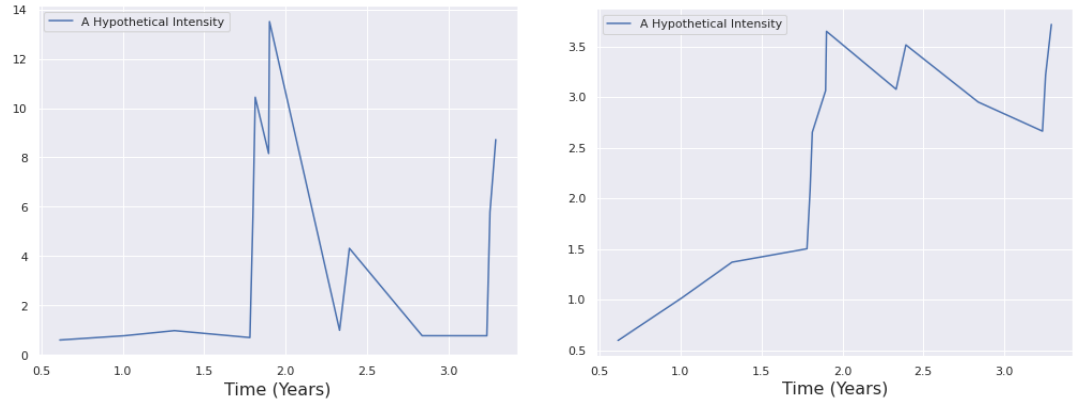
Here, z_i is a vector of individual heterogeneities who enter into the model linearly via an exponential link function. Notice that there is no need to specify a constant term for the γ parameters, since the parameter α captures a base level of the intensity function, i.e., $\alpha = \exp(\gamma_0)$. $\exp(\lambda_0)$ measures background risk, and

$$\omega_{i,t}(0) = \alpha \exp(z_i^\top \gamma) \sum_{j:t_{ij}<t} \exp(-\mu(t - t_{ij}))$$

is the dynamic moral hazard when having the fully covered health insurance. In order to validate the Hawkes process, we further assume that $\alpha, \mu > 0$.

We highlight three points regarding the dynamic moral hazard in the free plan. First, the self-exciting kernel $\exp(-\mu(t - t_{ij}))$ is a decreasing function with the elapsed time $t - t_{ij}$. Since we focus on the outpatient activities, the surge in the intensity function (due to the occurrence of $j-$th event) is unlikely to be permanent, and its impact would disappear eventually. We imagine the source of this surge is the additional awareness of health condition from latest outpatient activity. Examples of this awareness could be doctor's order for a follow-up check, or the nature of the illness.

Second, even the shadow coinsurance rate is fixed at zero, the dynamic moral hazard curve is event-based and non-trivial. Considering a hypothetical individual i such that his/her heterogeneity satisfy $\exp(z_i^\top \gamma) = 0$. Furthermore, we set $\mu = 9.0$, $\alpha = 6.0$, $\exp(\lambda_0) = 0.6$, and has outpatient events occurred at times (0.614, 1.008, 1.318, 1.781, 1.800, 1.814, 1.896, 1.901, 2.332, 2.392,2.838, 3.236, 3.255, 3.290). The left panel of Figure 3 presents the corresponding intensity curve.



(a) $\mu = 9.0$, $\alpha = 6.0$ (b) $\mu = 0.9$, $\alpha = 0.6$

Figure 3: A hypothetical intensity

The spikes in this intensity curve appear in the occurrence times, and occurrence times that are close to each other further amplify the intensity value. Different settings of parameters change the intensity shape dramatically, as illustrated in the right panel of Figure 3, where we set $\mu = 0.9$ and $\alpha = 0.6$, but everything else equal.

Third, although the trajectory of a dynamic moral hazard depends on the individual's outpatient activity times, its branching ratio n_i^* only depends on parameters α, μ and a single indexed element $\exp(z_i^\top \gamma)$:

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

Convenient it is, a branching ratio is not injective to a dynamic moral hazard curve. Take previous illustrations in Figure 3 as an example, where the trajectory differs a lot, but these two dynamic moral hazard have the same branching ratio: $n^* = 6/9 = 0.6/0.9 = 2/3$. Nevertheless, we believe that summarizing the dynamic moral hazard by a branching ratio still provides valuable information, and will report the distribution of them in this study.

5.2 Model for Cost-Sharing Plan P95

The cost-sharing plan we studied here only differs from the free plan by the coinsurance rate of outpatient activities. By the shadow coinsurance rate argument presented before, changes in the dynamic moral hazard should also depend on the variations of the insurance coverage. Since shadow coinsurance rate c is driven by the accumulated health care expenditure $x(t)$, which is piece wise constant between two outpatient activities, it is natural to specify $c(x(t_j))$ entering the intensity along side with each self-exciting kernel $\exp(-\mu(t-t_j))$. Thus, the parametrization for plan P95 is

$$a_i^{P95}(t) = \exp(\lambda_0) + \omega_t(c_t) \tag{9}$$

$$= \exp(\lambda_0) + \alpha \exp(z_i^\top \gamma) \sum_{j:t_{ij}<t} (1 - c_i(x_i(t_{ij}))) \exp(-\mu(t - t_{ij})) \tag{10}$$

Given the accumulated health care expenditure $x_i(t_{ij})$, the shadow coinsurance $c_i(x_i(t_{ij}))$ is the conditional expectation of the nominal coinsurance rate at the end of the contract year, its specification is given by:

$$c_i(x_i(t)) = \mathbb{I}\{x_i(t) \leq OOP\} - (1-b)\exp(\beta_1(x_i(t) - OOP))\mathbb{I}\{x_i(t) \leq OOP\} \tag{11}$$

The meanings of each element are:

- β_1 is the coefficient associated with the gap between current accumulated expenditure and the OOP: $x_i(t) - OOP < 0$. When the accumulated expenditure exceeds the OOP, the shadow coinsurance rate reduces to zero. We restrict $\beta_1 > 0$ such that $c(x(t))$ is decreasing with $x(t)$;

- The coefficient $b \in (0, 1)$ measures the degree of decreasing of health care utilization due to having a cost-sharing policy. The presence of b also implies that we would assume that there is a jump of intensity from below the OOP to above the OOP: below the OOP, the self-exciting kernel is at least discounted by a factor of b (when $x(t) = OOP$), while once above the OOP, there would be no discount.

- The cost-sharing plan P95 we used in this study would impose an OOP of \$150 per individual or \$450 per family. To simplify the analysis, we assume all will only use the individual OOP.

The calculating of the branching ratio in the cost-sharing plan is challenging. Formally, the branching ratio is expressed as:

$$n_i^* = \alpha \exp(z_i^\top \gamma) \int_0^\infty (1 - c_i(x_i(t))) \exp(-\mu t) dt$$

where the shadow coinsurance rate enters into the equation. However, $c_i(x_i(t))$ is stochastic, which makes the above expression impossible to caculate.

We can, nevertheless, find a branching ratio where the shadow coinsurance rate 'freezes' at one particular point v . In which case, the 'branching ratio' is

$$n_i^*(c_i(v)) = \alpha \cdot (1 - c_i(v)) \cdot \exp(z_i^\top \gamma) \int_0^\infty \exp(-\mu t) dt$$

For example, if we set $c_i(x_i(t)) = c_i(0)$, then the branching ratio is

$$\begin{aligned} n_i^*(c_i(0)) &= \frac{\alpha \exp(z_i^\top \gamma)(1 - c_i(0))}{\mu} \\ &= \frac{\alpha \exp(z_i^\top \gamma)(1 - b) \exp(-\beta_1 \cdot OOP)}{\mu} \end{aligned}$$

6. Results

This section presents the empirical results. We begin with the interpretation of individual heterogeneity effects. This interpretation is identical to that of the marginal effect at a representative value (MER) of a count data model when we fix a period and treat the counting process as a count data. Specifically, Let $Y_{it} = N_i(t)$ be the number of events that occurred before time t. Let scalar z_{ij} denote the $j - th$ covariate. Differentiating

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

by the exponential structure of $\exp(z_i^\top \gamma)$.

The estimation results is presented in Table 2. First, let's focus on time-invariant explanatory variables, in which we include age, sex, education (in terms of schooling years) and log-income as individual covariates. We introduce age2 and edu2 to model the nonlinear effect of age and education, and they are defined as $age^2/100$ and $edu^2/100$, respectively.

For free plan, we observe:

- *Age.* At first, intensity values will decrease as age increases. After one passes the age of 48, intensity values and age are positively correlated. It is well-known that youngsters are more risky compared to their mid-age counterparts. While, as individuals begin to age, they become physically weaker and are more prone to sickness.

Table 2: Main Results

	Free Plan	*Cost-Sharing Plan*
α	4.40	3.68
	(5.08)	(0.76)
μ	7.77	7.04
	(7.73)	(1.61)
age	-0.53	-0.47
	(0.06)	(1.36)
age2	0.55	0.64
	(0.09)	(1.65)
male	-0.45	-0.36
	(0.16)	(3.65)
edu	-1.06	-1.52
	(0.04)	(3.11)
edu2	4.20	4.79
	(0.17)	(4.42)
log income	1.76	1.93
	(0.11)	(3.52)
λ_0	-0.40	-0.17
	(0.01)	(1.34)
b		0.67
		(0.09)
β_1		0.04
		(0.01)

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

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

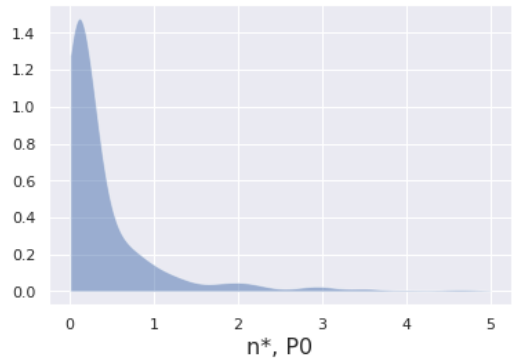
- *Education and Income.* Income is positively correlated with the use of medical service. The result on education, by and large, suggests an u-shape relation between education and the outpatient utilization. The better education one obtains, the fewer outpatient activities one would conduct. The turning point is around 12 schooling years, roughly one got a high school diploma. After which, better education implies higher chance to visit a doctor. One explanation is that higher education often associates with a healthier lifestyle, which reduces the hazard rate of visiting a doctor. While above high school degrees are highly correlated with high income, which gives individuals the ability to cover the opportunity cost related to the absence from work.

Similar observations also appear in the cost-sharing plan. The point estimators are similar to those of free plan, which is no surprising given the random assignment experiment design. However, the standard errors are large. There are two possible explanations, first, the sample size in P95 plan is relatively small ($n = 364$), and the finite sample properties are compromized. Second, since the model is highly nonlinear, the complexity level is high, so that the bias is small but the vairance is large. A natural solution to the second explanation would be some kind of regularization estimation, but this is beyond the scope of this papaer, and will be left for future research. We perform a Wald test where the null hypothesis is that all the coefficients assocaited with time-invariant variables are zero, and the althernative hypothesis is just its negation. The test statistic ends up with a value of 1002.5, clearly rejecting the null.

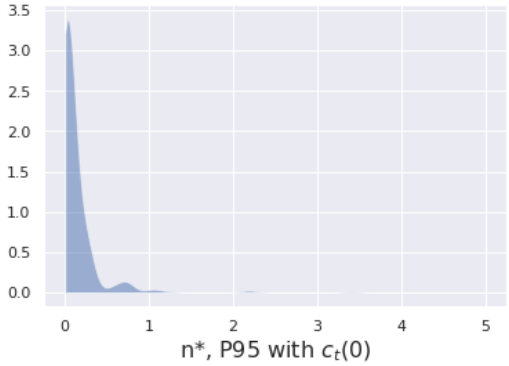
Next, we analyze the dynamic moral hazard results. As discussed before, we will use the branching ratios $\{n_i^* : i = 1, \ldots, n\}$ to summarize the dynamic moral hazard trajectories. Table 3 summarizes its descriptive statistics. On average, the introducing of a 95% coinsurance rate reduces the branching ratio (and more or less the dynamic moral hazard) compared to the free plan. Figure 4 summarizes the branching ratio distributions for P0 plan and P95 plans with different shadow costs c_t. It is clear that branching ratios are most concentrated in smal values in the least covered situation (i.e., P95 with $c_t(0)$), with the increasement of the shadow coinsurance, branching ratios began to expand, but most of them are still in the region of stationary, (i.e., $n^* < 1$). Branching ratios that correspond to non-stationary Hawkes process (i.e., the number of events is infinite when the time goes to infinite.) also begin to show up as the shadow coinsurance rate increases. In our context, these processes are associated with sick individuals who have to consume a great amount of health care services. In figure 5, we present the non-stationary percentage against the OOP gap. It appears that when the accumulated expenditure reaches the half way, there exists a sharp jump of the nonstationary processes.

Table 3: Descriptive Statistics of n^*

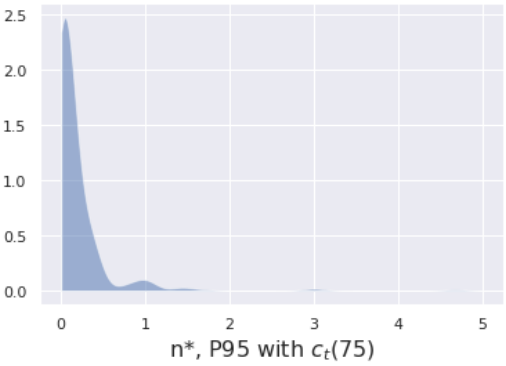
	P0	P95, $c_t(0)$	P95, $c_t(75)$	P95, $c_t(125)$
mean	0.604	0.205	0.281	0.347
variance	2.984	2.462	4.629	7.052
1st quantile	0.066	0.016	0.022	0.027
median	0.142	0.033	0.045	0.056
3rd quantile	0.45	0.107	0.147	0.182
% of $n^* > 1$	12.9%	2.2%	4.4%	5.5%



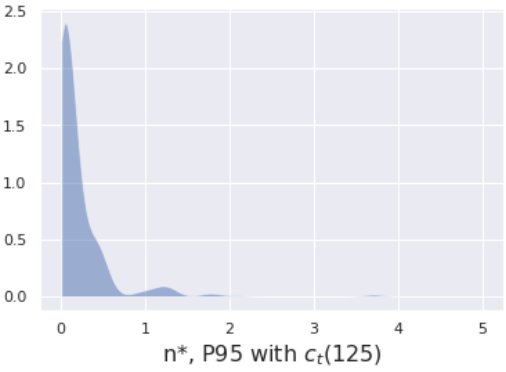
(a) Free Plan P0



(b) Cost-sharing Plan P95 with $c_t(0)$



(c) Cost-sharing Plan P95 with $c_t(75)$



(d) Cost-sharing Plan P95 with $c_t(125)$

Figure 4: Branching ratio distribution

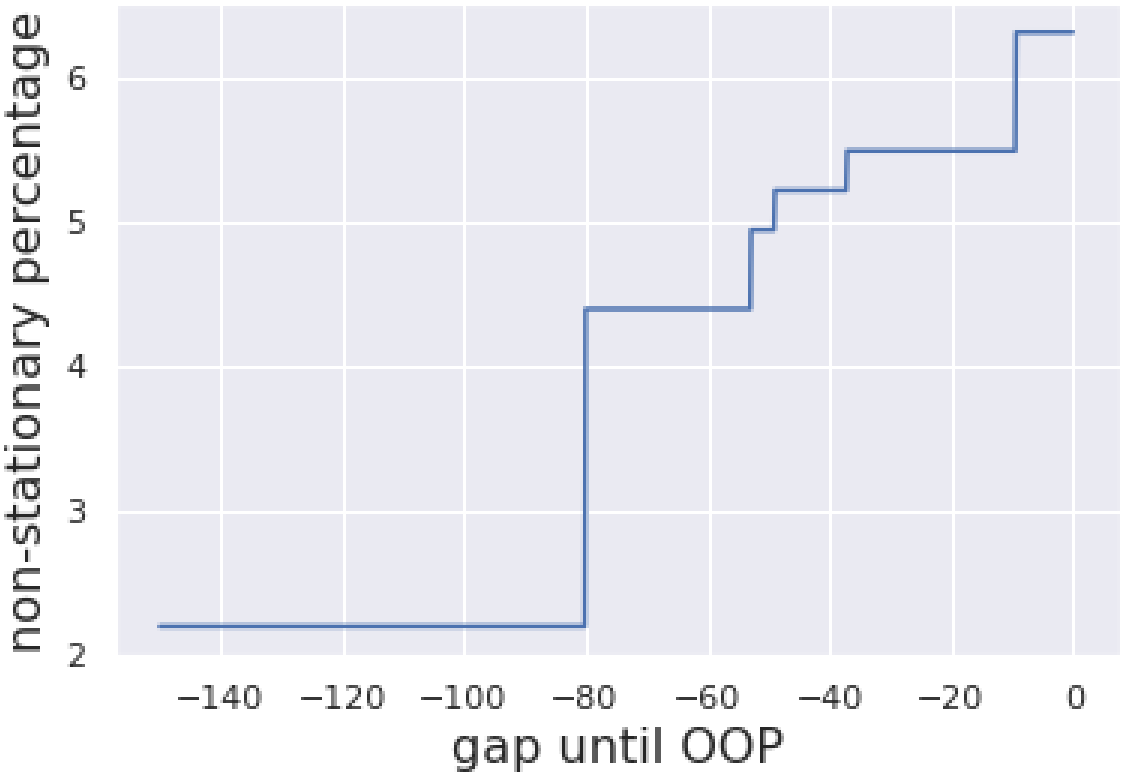


Figure 5: evolving of non-stationary branching ratio

7. Discussion

Dynamic programming is a popular framework when constructing a dynamic model. In this section, we first compare our Hawkes process approach to this approach. We discuss the differences between these two approaches and highlight some advantages of the Hawkes process approach. Next, we also discuss some limitations of the Hawkes process. Specifically, we discuss the issue of the unobserved heterogeneity.

7.1 Dynamic Programming Model of Moral Hazard

Literature on modelling the dynamic moral hazard is scarce, to the best of our knowledge, the only other paper that aims to model the moral hazard dynamically is Diaz-Campo (2022), where the author uses a dynamic programming approach to construct the model. We first summarize her model and then make a comparison. As usual, we omit i to simplify notations.

The author builds a single-agent, finite-horizon, dynamic and stochastic model. Begin with the utility function, which is a trade-off between health h and the residual income y:

$$\begin{aligned} u(c_t, C_{t-1}, \omega) &= h(c_t - \lambda_t; \omega) + y(y_t, c_t, C_{t-1}) \\ &= \left[(c_t - \lambda_t) - \frac{1}{2\omega}(c_t - \lambda_t)^2\right] + [y_t - K(c_t, C_{t-1})] \end{aligned}$$

where c_t is the monetized utilization in period t, C_{t-1} is the accumulated monetized utilization, λ_t is a health risk realzied in period t, y_t is a period income, ω is a price elasticity, and $K(\cdot)$ is the out-of-pocket expenditure function.

Let T be the number of periods, the individual solves

$$\max_{\{c_1,\ldots,c_T\}} \sum_{t=1}^{T} \delta^{t-1} \mathbb{E}\left[u(c_t, \lambda_t, \omega)\right] \text{ s.t. } K(C_T, 0) + Y_t \leq \text{Income}, C_T = \sum_{t=1}^{T} c_t, Y_T = \sum_{t=1}^{T} y_t$$

The author has shown the optimal choice of monetized utilization c_t^* is

$$c_t^*(C_{t-1}, \lambda_t) = \max\left[0, \omega\left(1 - \frac{\partial K(c_t, C_{t-1})}{\partial c_t} + \delta \frac{\partial \mathbb{E}V_{t+1}(C_{t-1} + c_t, \lambda_{t+1})}{\partial c_t}\right)\right]$$

where the term

$$\frac{\partial K(c_t, C_{t-1})}{\partial c_t} - \delta \frac{\partial \mathbb{E}V_{t+1}(C_{t-1} + c_t, \lambda_{t+1})}{\partial c_t}$$

is defined as the shadow coinsurance rate, and the value function satisfies

$$V_{t+1}(C_{t-1}, \lambda_{t+1}) = \max_{c_t}[u(c_t, \lambda_t, \omega) + \delta]\mathbb{E}V_{t+1}(C_t, \lambda_{t+1})$$

The whole model is estimated by GMM framework, see Diaz-Campo (2022) for detailed explanation.

Compared to the dynamic programming framework, our approach have advantages over three differences. First, the optimal object is different. in Diaz-Campo (2022), the author aims to optimize a monetized health care utilization in each period; while in our paper, the optimal object is the intensity of the utilization. The dynamic programming is a discrete strategy, it involves a choice of the period frequency. In Diaz-Campo (2022), the author chooses a weekly frequency but provides limited reasoning. The same model should also work for a monthly or daily frequency, but the results from these model are different. Furthermore, there is no general theory providing principles on frequency choices. In contrast, our approach is a continuous strategy: given history, the intensity measures the instantaneous conditional rate of the occurrence of an outpatient activity. The frequency is set as the infinitesimal time interval.

Second, the dynamic programming approach uses aggregated data, while our approach uses non-aggregated data. Aggregating data leads to a loss of information as illustrated in Figure 1, where important features of the data, e.g., clusters, can not be detected. Furthermore, by using monetized data, certain dynamic features are also lost, and this point leads to our third argument.

Third, the dynamic mechanisms are different. In the dynamic programming approach, the only source for dynamics is the accumulated expenditure. We make two comments on this issue. (1) The shadow price is the difference between the spot price and the term $\partial \mathbb{E} V_{t+1}(\cdot)/\partial c_t$, where the value function V_{t+1} is driven dynamically solely by the accumulated expenditure C_{t-1}. (2) The health risk $\{\lambda_t : t = 1, \dots T\}$ are drawn independently from a distribution $F_\lambda(\cdot)$, whereas in practice, health risks are often state-dependent, i.e., past illness would alter the distribution of health risks. In contrast, we include two sources of dynamic in the Hawkes process approach. Besides the accumulated expenditure, we include a summation of self-exciting kernels $\sum_{j:t_j<t} \alpha \exp(-\mu(t-t_j))$ to represent the state-dependent effect, which describes how past episodes affect the future ones. The best example to illustrate our advantage is by looking at the dynamic moral hazard of the free plan. In the dynamic programming approach, the moral hazard should be a constant, which would generate stationary observations. However, as shown in the left panel of Figure 1, the occurrence times in P0 are still clustered and highly non-stationary, which indicates the existence of a second source of dynamics.

7.2 On the Issue of the Unobserved Heterogeneity

Since Heckman (1981), distinguish between the unobserved heterogeneity and the state dependence has been well recognized. Nevertheless, making such distinguish is anything but trivial. In the dynamic programming approach, this is done by imposing strong assumption. For example, one has to parametrize the distribution of the unobserved health risk λ_t,

usually, $F_\lambda(\cdot)$ is log-normal. Furthermore, λ is uncorrelated with other variables. In the counting process literature, Li (2022) proposes a first ratio operation to cancel out the fixed effect. This method uses the fact that one can build a counting process from a sequence of durations, and works in a generalized accelerated failure time duration:

$$t_{ij} = G(x_{ij}; \beta_0)\nu_i u_{ij}$$

where $G(\cdot)$ are known functions up to parameters β_0 , t_{ij} is the $j-$th duration of individual i, x_{ij} is a vector of state-dependent variables, ν_i is fixed effect individual heterogeneity, and $\{u_{ij}\}$ are i.i.d error terms. With some mild assumptions, the first ratio operation works on this duration model as:

$$\tilde{t}_{ij} = t_{ij}/t_{i(j-1)} = (G(x_{ij}; \beta_0)/G(x_{i(j-1)}; \beta_0)) \cdot (u_{ij}/u_{i(j-1)})$$

Let $s_{ij} = \sum_{k=1}^{j} \tilde{t}_{ik}$, one could use $\{s_{ij}\}$ to construct a counting process,

$$N_i(s) = \sum_{j=1}^{\infty} \mathbb{I}\{s_{ij} \leq s\}$$

whose compensator has shown to be

$$A_i(s) = A_i(s_{i(j-1)}) + \int_{s_{i(j-1)}}^{s} h_{ij}(z - s_{i(j-1)})\mathbb{I}\{S_{ij} > z > s_{i(j-1)}\}dz$$

where $h_{ij}(\cdot)$ is the hazard rate of $\tilde{t}_{ij}$.

Although there exists method to separate the unobserved heterogeneity and the state-dependence, we believe that the unobserved heterogeneity in this paper is not a serious concern for three reasons. First, the RAND HIE is a randomized field trial of different insurance plans. Insurance plans were randomly assigned to individuals. This prevents the possibility that less-healthy individuals (an unobserved covariate), anticipating large health care utilization, opt to more generous insurance coverage (i.e., the adverse selection problem). Second, in addition to having observed individual heterogeneities in the model, we also leverage the self-exciting part (i.e., $\sum_{j:t_{ij}<t} \alpha \exp(-\mu(t-t_{ij}))$) of the intensity as an approximation to the unobserved heterogeneity. Finally, the first-ratio operation mentioned before requires a particular counting process specification, although such counting process is capable of modelling cluster data, it is not clear under which mechanism a cluster is generated, nor do we know the structure of these clusters.

8. Conclusion

In this paper we model the dynamic moral hazard via a Hawkes process framework. The primary source of dynamic is cost-sharing policies in standard health insurance contracts. These policy designs generate a nonlinear price scheme where current health care utilization lowers future expected prices. We focus in outpatient claims as they are frequent and often less expensive compare to inpatient services. The outpatient claims provide rich history of individual health care utilization, making estimating the dynamic moral hazard much easier. Nevertheless, outpatient claims highly nonlinear and clustered, even in a fully covered insurance plan, where price incentives are absent. Furthermore, outpatient claims are doubly stochastic in a sense that for an individual, both the claim times and the number of claims in a given time period are stochastic.

To overcome these challenges, we adopt the Hawkes process framework. The Hawkes process is a counting process with the self-exciting property, i.e., past claims will affect the occurrence of future claims. It also has a branching structure, which naturally leads to a cluster interpretation. These properties make the Hawkes process an ideal tool to model the outpatient claim data.

We build a stylized model to explain and define the dynamic moral hazard. This model consists of a utility function that takes the intensity as argument. We derive the optimal intensity and show that the self-exciting kernel of the Hawkes process can be understood as the dynamic moral hazard. We specify two sources of dynamic: one arises from the nonlinear price scheme generated by the cost-sharing policies, and the other arises from a genuine state dependent effect.

The Doob-Meyer decomposition of a Hawkes process is used to construct moment condition, based on which, a minimum distance principle is employed to estimate the model. We further summary the estimated dynamic moral hazard by a single quantity called the branching ratio. Distributions of the branching ratio for different insurance coverage is presented, and we show that individuals respond to there dynamic mechanisms.

References

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

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

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

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.

Arrow, K. J. (1978): "Uncertainty and the welfare economics of medical care," in *Uncertainty in economics*, Elsevier, 345–375.

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

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

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.

Cardon, J. H. and I. Hendel (2001): "Asymmetric information in health insurance: evidence from the National Medical Expenditure Survey," *RAND Journal of Economics*, 408–427.

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

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

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

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

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

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

Diaz-Campo, C. (2022): "Dynamic Moral Hazard in Nonlinear Health Insurance Contracts," *Working paper, Washington University in St. Louis*.

Einav, L., A. Finkelstein, S. P. Ryan, P. Schrimpf, and M. R. Cullen (2013): "Selection on moral hazard in health insurance," *American Economic Review*, 103, 178–219.

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.

Guo, A. and J. Zhang (2019): "What to expect when you are expecting: Are health care consumers forward-looking?" *Journal of Health Economics*, 67, 102216.

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

Harris, T. E. et al. (1963): *The theory of branching processes*, vol. 6, Springer Berlin.

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

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

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

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

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

Khwaja, A. (2010): "Estimating willingness to pay for medicare using a dynamic life-cycle model of demand for health insurance," *Journal of Econometrics*, 156, 130–147.

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

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

Li, Y. (2022): "Analyzing Dynamic Multiple Spell Durations Using Counting Processes," *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*.

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

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

Pauly, M. V. (1968): "The economics of moral hazard: comment," *The american economic review*, 58, 531–537.

Rubin, I. (1972): "Regular point processes and their detection," *IEEE Transactions on Information Theory*, 18, 547–557.

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

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

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

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

A. Data Generating Process and Simulation Details

The data generating process for simulation studies is the epidemic type aftershock sequence (ETAS) model. The ETAS model was first introduced by Ogata and Katsura (1988) and ever since has been widely used in seismology literature (Zhuang et al., 2002). The model extends the classical Hawkes model and includes the marks, it characterizes both the earthquake times and magnitudes. The intensity of a ETAS model, for its simplest form, could be:

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

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

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

The above data generating process can be simulated using the R package 'PtProcess' (Harte, 2010).[2]. We set the true parameters as $\mu = 0.007$, $\alpha = 1.98$, $c = 0.008$ and $\delta = log(10)$ and generate $N = 50, N = 100, N = 200$ and $N = 400$ individual counting processes. The time-intervals are set to be $(0, 100], (0, 500]$ and $(0, 3000]$. For each simulation setting, we run $B = 1000$ repeats.

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

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

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

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

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

go to step 6.

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

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. Simulation Results

We report standard deviation (SD), median of absolute deviation (MAD), 95% confidence interval coverage rate (CI95) and 90% confidence interval coverage rate (CI90). The results are presented below. As the number of observations N increases, the estimators become more stable and their empirical coverage rates get closer to the theoretical ones.

Table 4: Minimum Distance Estimator Results, with $T = 100$

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

Table 5: Minimum Distance Estimator Results, with $T = 500$

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

Table 6: Minimum Distance Estimator Results, with $T = 3000$

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