Modelling Dynamic Incentives 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

Abstract

Deductibles in health insurance contracts create dynamic incentives: as costs accumulate, the gap between total spending and the deductible decreases, lowering the effective price of health care. This article uses a self-exciting process framework to identify and quantify these dynamic incentives. We build a model that distinguishes direct and indirect dynamic incentives. Using the Rand Health Insurance Experiment data, we find evidence favoring the existence of dynamic incentives. Furthermore, we find that patients would time their healthcare needs and exhibit retaliatory behaviors after reaching their deductible limits.

*. We thank conference and seminar participants at the EEA-ESEM Lisbon 2017, IAAE Montreal 2018, UC3M, Liaoning University and SUFE. All errors are ours.

JEL Classification: G22, I12, I13, C13

Keywords: Dynamic Incentive, Self-Exciting Process, Health Insurance.

1. Introduction

Deductibles are ubiquitous in the insurance market. In the United States, according to the Kaiser Family Foundation's 2019 Health Benefits Survey, 82% of employer-sponsored health insurance plans feature deductibles. With deductibles, patients are required to pay a proportion of healthcare costs out of pocket before reaching a coverage limit, after which the insurance plan provides comprehensive coverage. Deductibles generate both static and dynamic incentives. Static incentives imply that patients make healthcare decisions in response to the nominal cost below the deductible threshold. Dynamic incentives imply that patients make healthcare decisions in response to a "shadow price" created by the deductible. The shadow price emerges because consumption today reduces the remaining deductible, effectively rendering the next purchase less expensive.

Setting the amount of deductible is an essential element in the design of any health insurance plans. The question "which price" (Einav and Finkelstein, 2018), i.e., whether patients respond to the nominal price set by the deductible or the shadow price induced by the deductible, is of great academic and policy interest. For example, forward-looking patients will react less to a deductible than myopic patients, since the former would respond to the shadow price, while the latter only respond to the much-higher nominal price. Previous studies do not reach a consensus on this issue: Aron-Dine et al. (2015); Einav et al. (2015); Johansson and Palme (2002); Klein et al. (2022) find evidence for substantial dynamic incentives, while Abaluck et al. (2018);

Brot-Goldberg et al. (2017); Dalton et al. (2020); Keeler and Rolph (1988) favor the static incentive hypothesis

Early theoretical literature on patients' responses to cost-sharing incentives concluded that, under certain assumptions, forward-looking individuals only respond to a shadow price called the end-of-year (EOY) price (see, e.g., Ellis (1986); Keeler et al. (1977)). In the literature, this is often defined as the *unconditional* probability of not exceeding the deductible threshold. The most common nonparametric estimator of such a price is the fraction of patients who are unable to surpass the deductible limit by the end of the year (see, e.g., Aron-Dine et al. (2015); Klein et al. (2022)). This definition of shadow price (and its estimator) is unsatisfying, as two otherwise identical patients may have different chances of reaching the threshold depending on their remaining deductibles. Using a biased estimator of the shadow price could lead to flawed conclusions.

Another important yet often ignored question is "whose incentives?" Any data on healthcare utilization results from both patients and doctors. Compelling evidence (see, e.g., Cromwell and Mitchell (1986); Dranove and Wehner (1994); Yip (1998)) shows the existence of so-called "supply-induced-demand" (SID), defined as excess healthcare use beyond what would have occurred if patients were fully informed. It is surprising that vast literature estimating the price sensitivity of *patients* ignores the difference between patients' and doctors' incentives.

The prevailing patient-physician relationship is asymmetric: patients exercise minimal agency over their treatment regimen or the associated costs. Nevertheless, the preponderance of literature, particularly studies utilizing a structural approach, operationalizes healthcare costs as the dependent variable and attempts to identify and quantify patients' reactions to dynamic incentives therein. We call into question this

choice of dependent variable: it is possible that patients demonstrate myopia but physicians exhibit more forward-looking behavior.

The aim of this paper is to identify and quantify patients' responses to dynamic incentives in the context of health insurance with deductibles. We use the Rand Health Insurance Experiment (Rand HIE) data for analysis. The Rand HIE provides a favorable setting for our purpose, as it randomly assigns individuals to different health plans, avoiding the typically confounding adverse selection present in insurance markets. To achieve the goal, we propose an approach that: (1) Focuses on healthcare events rather than healthcare expenditures to identify patients' dynamic incentives. Specifically, we examine outpatient care events, the most frequent type of medical care. Inpatient claims are infrequent and often associate with expenditures that meet or exceed the deductible threshold; thus, we do not focus on them. Accordingly, we choose the individual deductible plan from Rand HIE for analysis. This plan only impose deductibles on outpatient use, while inpatients are fully covered. (2) Uses conditional rather than the unconditional EOY price. For this, we specify a parametric conditional probability of exceeding the limit given an individual's own cumulative costs. (3) Distinguishes direct and indirect dynamic incentives. As defined before, direct incentives arise from changes of the shadow price. Indirect incentives stem from the potential state dependence, i.e., impacts of previous events on subsequent ones. The shadow price would affect one's healthcare records, which then affect future healthcare events through state-dependent effects. (4) Identifies retaliatory healthcare utilization after reaching the deductible limit. Retaliatory use may occur when patients use private information about their realized healthcare needs and time their consumption across coverage periods (e.g., postpone some selective healthcare needs to post-deductible period) so as to reduce out-of-pocket costs or purchase additional coverage.

Related to (1), we hypothesize that patients initiate outpatient care events, while physicians determine subsequent treatments and their scale. We aim to test this hypothesis from both the supply and demand sides. Specifically, we analyze three Rand HIE plans:(i) The free plan that imposes no restrictions on patients or physicians. (ii) The individual deductible plan that places restrictions on patients. (iii) The HMO plan that places restrictions on physicians. The hypothesis holds true if there are no differences in healthcare spending between the free and individual deductible plans, and no differences in outpatient event frequency between the free and HMO plans.

Related to (2), by imposing structure on the shadow price, we gain tractability and interpretability, at the potential cost of misspecification. An advantage is the simplified identification and quantification of (direct) dynamic incentives. Specifically, we need only assess whether the coefficient on remaining deductible is different than zero at a statistically significant level. The magnitude of this coefficient also indicates the degree of price sensitivity among patients. On balance, the benefits of this parametric assumption appear to outweigh the costs for the present analysis.

Related to (3) and (4), focusing on healthcare events rather than healthcare expenditures may reveal that outpatient events tend to cluster temporally. For instance, Figure 1 depicts outpatient incidences times (red dots) of an individual with the free plan. Given the free insurance plan, dynamic incentives from the shadow price are absent. Rather, event-based state-dependence, e.g. potential impacts of previous claims on subsequent ones, may contribute to this clustered structure. In a deductible plan, we call dynamic incentives that carried by the state-dependence *indirect* dynamic incentives. It turns out that indirect dynamic incentives contain important information on retaliatory utilization. We found that the state-dependent effect is stronger (past events affect future ones longer) subsequent to exceeding the deductible than

with a free plan. This result implies that patients would postpone some part of their healthcare needs to post-deductible period.

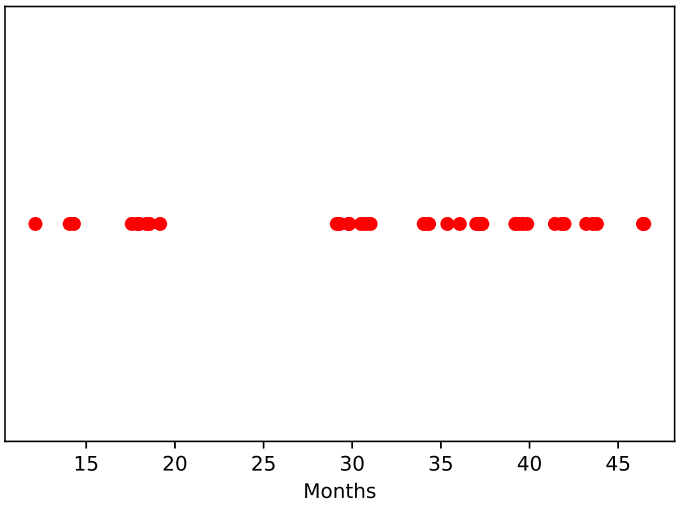


Figure 1: Outpatient claims instances over time under the Rand HIE free plan

Rather than aggregating outpatient events at some temporal resolution (e.g., annually, monthly or weekly) and employing count data regression techniques for analysis, we utilize a novel stochastic process known as the self-exciting process to rigorously analyze the raw line-item outpatient claim data. A self-exciting process is one where the occurrence of an event makes the occurrence of the same event more likely in the near future. In other words, the event itself excites or triggers more of the same event to happen subsequently. Modeling these processes requires taking into account both the long-term rate of occurrence as well as the short-term triggering effect. The self-exciting process allows us to capture the temporal spread of events and the complex dynamic incentives created by cost-sharing policies, and provides a more nuanced understanding of patient behavior and dynamic incentives in health insurance.

Out findings can be summarized as the followings. (1) Patients are more affected by deductibles than physicians. This can be seen in how the intensity of outpatient events decreases when restrictions are placed on patients (the individual plan) but remains unchanged when restrictions are placed on physicians (the HMO plan). (2) Patients are forward-looking, the coefficient associated with the deductible remaining are statistically different than zero. We quantify patients' dynamic incentives (both direct and indirect) as stochastic processes. (3) Patients would time their healthcare needs and exhibit retaliatory behaviors after reaching their deductibles.

Our article relates to several areas of research. First, it adds to the limited research testing whether people respond to dynamic incentives in nonlinear health insurance contracts. Previous studies have taken two different approaches: A reduced-form approach, using quasi-experimental sources of variation to test whether people react to dynamic incentives, and structural modeling, quantifying the response to dynamic incentives using a fully specified structural model. For the first approach, Aron-Dine et al. (2015) studies employees who enroll in health plans in different months. They exploit the fact that annual coverage often resets every January, workers who join a plan later in the year face the same nominal price but a higher shadow price. Using a difference-in-difference framework, they reject the hypothesis of fully myopic behavior and favor the existence of dynamic incentives. A complementary paper, Guo and Zhang (2019), studies individuals who have a large expenditure planned in the future (childbirth). They reject the hypothesis that patients are fully forward-looking, i.e., only respond to the shadow price and do not respond to the nominal price. Klein et al. (2022) uses Dutch health data and exploits two sources of variation in a difference-in-difference-discontinuities design: deductibles reset at the beginning of each year, and deductible limits change over the years. They found strong evidence that individuals are forward-looking. Johansson et al. (2023) exploits a policy in

Sweden where primary out-of-pocket prices were eliminated at age 85, and also finds forward-looking behaviors among the elderly.

Studies on Medicare Part D often adopts the second approach. Some of these studies test the hypothesis of full myopia through the estimation of a discount factor: a discount factor of zero would indicate full myopia. Einav et al. (2015) uses the "donut hole" nonlinear budget in the Medicare Part D to estimate a weekly discount factor of 0.96, rejecting the hypothesis. On the other hand Dalton et al. (2020), estimate a discount factor of zero.

Our article also relates to research constructing models with an intensity function. Abbring et al. (2003) studied adverse selection and moral hazard in car insurance. They optimized a utility model by intensity and later estimated the intensity model using maximum likelihood methods. Finally, our work relates to studies using self-exciting process. The self-exciting process has been widely used in other disciplines. For example, in finance, see Bacry et al. (2015); Bowsher (2007); Chavez-Demoulin et al. (2005), in seismology, see Zhuang et al. (2002), 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), and in criminology, see Mohler et al. (2012).

The article is organized as follows. In section 2, we introduce the data and sample. We also test our behavior hypothesis from provider side. Model specifications are provided in section 3. In section 4, we present a minimum distance method for estimating parameters. Section 5 presents and discusses the results, and finally, Section 6 concludes the paper.

2. Data and Sample

We utilize individual-level, line-item records from the RAND Health Insurance Experiment (hereafter, HIE). The RAND HIE was a randomized field experiment of various insurance plans offered to over 8000 individuals in the U.S. These insured enrollees were assigned to different insurance treatments, and data on their use of health services were collected during their period of participation. The insurance treatments differed primarily in terms of cost-sharing policies, i.e., deductibles, coinsurance rates and out-of-pocket caps. Due to the randomness of the assignments and the nonlinear cost-sharing features, the RAND HIE data is particularly suitable for studying dynamic incentives. Here, we describe the experimental design and the analysis sample.

The Dataset

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 Georgetown 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. The coinsurance rates were set at either 0 (free care), $25, 50$ or 95 percent. 12 plans had a OOPC of $5, 10$ or 15 percent of family income in the previous year. The free plan does not impose OOPC, and a plan (labeled Plan N in the RAND HIE document) imposes a OOPC 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 Construction

We focus on outpatient data from the free plan and the individual deductible plan, which imposes an OOPC of 150 dollars per person or 450 dollars per family. The free plan exhibits the most moral hazard behaviors, so it is a natural choice. We chose the individual deductible plan because it covers 100% of inpatient services but pays 5% (a 95% coinsurance rate) of covered outpatient services until the OOPC is met. Thus, the free and individual deductible plans differ only in outpatient activities.

We also include a health maintenance organization or HMO plan. An HMO is a type of health insurance plan that provides comprehensive health care coverage through a network of doctors, hospitals, and other health care providers. In an HMO, patients generally must stay in-network to receive coverage and typically need a referral from a primary care physician to see a specialist. HMOs typically have lower out-of-pocket costs than other plans since the network of providers and health care services are tightly managed to control costs (i.e., place restrictions on the provider side). In the Rand HMO plan, there are no restrictions on the patients, and we compare it with the free plan to assess the implications on health care utilization.

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 this study, we focus on services rendered by physicians or other

health professionals (file 06 in the RAND HIE document). Both free and ID plans cover expenses of prescription drugs and supplies.

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 provide 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 be related to several health care consumption via different MERs. Specific to our study, we need to merge all related medical consumption to one item.

We apply the same restrictions to create analysis samples for all three plans. First, we exclude individuals younger than 18 and older than 60, primarily because their health conditions would lead to different responses to moral hazard. In addition to the age restriction, we exclude any claims outside the 1978-1979 contract year, since the ID plan resets its terms to default annually on the enrollment date. Table 1 shows the remaining sample sizes and line-item counts after applying each major exclusion.

The time unit is week. For example, if an insurance contract starts on January 1, 1977 and the date of a doctor visit is October 1, 1977, the time stamp is 39 (weeks). The demographic factors in the model are age, sex, education (in years of schooling), and log-income. For simplicity, we assume all ages are fixed at enrollment. So, all factors are time-independent. Other data cleaning assumptions: (1) If a doctor visit cost is unavailable, we replace it with zero. (2) If information on education is unknown, we replace it with the average education level.

As discussed previously, this work focuses on the frequency of outpatient usage rather than healthcare costs as in previous literature. However, cost per outpatient visit remains important despite the change in focus. The hypothesis in this work is that patients determine the initial outpatient activity, while doctors decide on the

Table 1: Sample Construction Procedure for Different Plans

free Plan		
Major Steps	Sample size	Line-item size
Outpatient Claims rederned by physicians	6263	173264
Only include individuals with $18 \leq$ age ≤ 60	3442	129760
Select individuals enrolled in the free plan	1001	45636
Focus on the contract year 1978-1979	723	11182
Merge line-item associated with a same episode	723	6894
ID Plan		
Major Steps	Sample size	Line-item size
Outpatient Claims rederned by physicians	6263	173264
Only include individuals with $18 \leq$ age ≤ 60	3442	129760
Select individuals enrolled in the ID plan	627	19973
Focus on the contract year 1978-1979	403	5123
Merge line-item associated with a same episode	395	2812
HMO Plan		
Major Steps	Sample size	Line-item size
Outpatient Claims rederned by physicians	1928	68532
Only include individuals from the experiment group	1143	38322
Only include individuals with $18 \leq$ age ≤ 60	664	27092
Focus on the contract year 1978-1979	488	5869
Merge line-item associated with a same episode	488	2856

necessary treatment and associated costs. If this is correct, costs should be similar for the free and ID plans from the provider perspective, but differ between the free and HMO plans.

To test this hypothesis from the provider side, we use a state-of-the-art two sample test called the kernel maximum mean discrepancy (MMD). The MMD test is often used to determine if two sets of samples come from the same distribution in kernel methods. The MMD metric calculates the mean values of all pairwise distances between samples from two distributions in a reproducing kernel Hilbert space (RKHS). MMD has been widely used in domain adaptation, transfer learning, outlier detection, and generative models evaluation. Some key features of MMD include: (1) It is a non-parametric test, so it does not assume any specific form of the distributions; (2) It can detect differences in high-order moments, not just means; (3) It has an unbiased empirical estimator and can be easily estimated from samples. We refer readers to the Appendix for the detailed description of this method. Table 2 presents testing results, which clearly favor our behavior hypothesis.

Table 2: Two Sample Test for Cost Distributions

	Free	ID	HMO
Free		True	False
ID	True		False
HMO	False	False	

We test whether the costs of two plans for each outpatient event come from the same distribution. The statement "Do the two cost samples originate from the same distribution?" determines if this is true or false.

Representing the Data

Suppose we observe an increasing series of random outpatient visit times $\{t_1 < t_2 < \dots\}$ for an individual over time. A counting process $N(t)$ for $t \in \mathcal{T} = (0, T]$ records the number of t_j times that occur before time t.:

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

where $\mathbb{I}\{A\}$ is an indicator, equal to 1 if event A occurred and 0 otherwise. The counting process $N(t)$ is fully characterized by its conditional intensity function $a(t)$, 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 $[t + dt)$. If the filtration $\mathcal{F}$ contains history information: $\mathcal{F}(t-) \supseteq \sigma(N(s) : s < t)$, this counting process is called the self-exciting process.

Briefly speaking, a self-exciting process is one where the occurrence of an event influences the occurrence of the same event in the near future. In other words, the event itself excites or triggers more of the same event to happen subsequently. Some examples of self-exciting processes include: (1) Earthquakes: An earthquake makes subsequent earthquakes more likely as the stress on fault lines gets redistributed; (2) Financial market crashes: A market crash increases the likelihood of another crash as panic spreads among investors; (3) Epidemics: An outbreak of a disease makes future outbreaks more probable as the infection spreads among the population; and

(4) Riots: A riot can trigger more rioting as unrest and violence spread from one area to another.

In our context, modeling the health insurance claims as a self-exciting process can determine the impact of cost-sharing policies on incentives. By estimating the excitation strength and decay rate (how lasting one occurrence would affect the future ones), it can evaluate how effective cost-sharing interventions are in reducing additional claims triggered by past claims.

3. The Econometric Specification

Dynamics in the Model

Before presenting our econometric models, we briefly discuss two key dynamic components: cost-sharing policies and state-dependent effects. We omit the individual i subscripts to simplify notation throughout this subsection.

As stated in the introduction, the logic behind the cost-sharing dynamic mechanism is simple: using a health care service today will effectively decrease the health care cost tomorrow. We refer to this mechanism as the direct dynamic channel, as it measures a patient's reaction to changes in the shadow price. To formally convey this idea, we introduce a shadow coinsurance rate, defined as the expected coinsurance rate at the end of the year given the accumulated spending so far $x(t)$:

$$c_t = c(x(t)) = \mathbb{E}(c_{EOY} \mid x(t)) \tag{1}$$

This shadow rate therefore captures the dynamic, path-dependent nature of health insurance plans with an OOPC. As the nominal coinsurance rate at EOY can be either the nominal rate defined by the insurance plan or zero (in which case, the individual

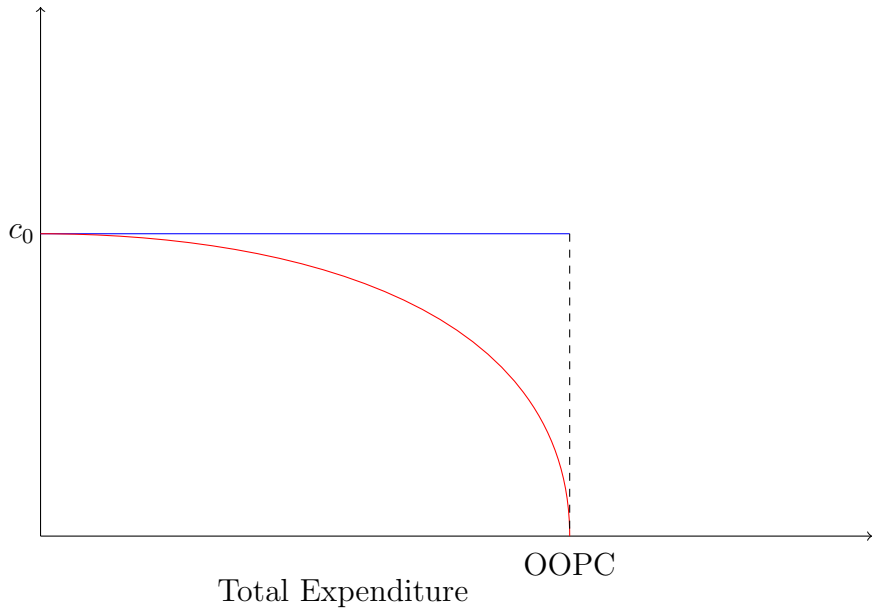
must have exhausted the OOPC), the shadow coinsurance rate can be written as:

$$c(x(t)) = c_0 \Pr\{C_{EOY} = c_0 | x(t)\}$$

This definition implies (1) patients are forward-looking: they will not respond to the spot price (i.e., the nominal rate), but will form an expectation about the coinsurance rate based on their personal experiences (i.e., $x(t)$). (2) The conditional probability (of $C_{EOY} = c_0$) reflects how this individual is "forward-looking": A low probability value implies the individual would have obtained a higher utility by consuming more health care services; whereas a high probability represents the opposite.

We further assume that the shadow coinsurance rate $c(x)$ is concave, with $c' < 0, c'' < 0$. Effectively, we are assuming that individuals would respond more aggressively as the total expenditure approaches the OOPC. The exact form of $c(x)$ will be specified in the next subsection. Figure 2 illustrates a typical scenario: The blue line represents the nominal coinsurance rate c_0 when total spending is under the OOPC, dropping to 0% thereafter. The red curve shows the individual's shadow coinsurance rate, which also reflects insurance coverage.

Another dynamic mechanism is the state dependence. An individual may respond differently to health care even if they are facing the same shadow price. State dependence arises from learning about one's health, changes in attitudes about preventive care, and other personal factors. Overall, state dependence illustrates how individuals' health care decisions depend on their unique experiences and circumstances. For example, more frequent doctor visits could stem from gaining a better understanding of one's health from prior visits. For this reason, we call this dynamic factor the indirect channel. Formally, the term "state dependence" $\tau_{[0,t]}$ is expressed as the



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

Figure 2: Two Types of Coinsurance Rate

sigma-field generated by all occurrence times prior to t:

$$\tau_{[0,t]} = \sigma(\{t_j\} : t_j < t) \tag{2}$$

Specifications for Different Plans

This subsection introduces our econometric model. The free plan has the most generous coverage. It has no restrictions to control either the patient's or the provider's moral hazard. In this plan, both the nominal and shadow coinsurance rates are fixed at zero. Therefore, its intensity model will focus on the state-dependent (the triggering) dynamic mechanism.

The HMO plan requires no cost-sharing from patients. However, it establishes regulations and guidelines for providers to control costs. For this reason, we specify the same intensity model as the free plan, aside from the parameters, to evaluate the effects of the provider-side restrictions.

Finally, the individual deductible plan imposes a fixed coinsurance rate on patients before reaching the out-of-pocket limit (OOPC), but there are no restrictions on the provider's side. In this plan, we will focus on specifying dynamic incentives induced by the shadow coinsurance rate $c_t = c(x(t))$.

Free insurance plan (FREE). For an individual i, we specify his/her intensity as:

$$\begin{aligned} a_i^{FREE}(t) &= \exp(z_i^\top \gamma) \left(\exp(\lambda_0) + \tau_{[0,t]}\right) \\ &= \exp(z_i^\top \gamma) \left(\exp(\lambda_0) + \int_0^t \mu_0^2 \exp(a_0 - \mu_0^2(t-s)) dN_i(s)\right) \\ &= \exp(z_i^\top \gamma) \left(\exp(\lambda_0) + \sum_{j:t_{ij}<t} \mu_0^2 \exp(a_0 - \mu_0^2(t-t_{ij}))\right), \quad t \in \mathcal{T} \end{aligned} \tag{3}$$

We assume the intensity for the free plan takes the form of the Hawkes process. The Hawkes process is a self-exciting point process that represents event occurrences as a stochastic model with time-dependent intensity. The intensity at time t depends on the history of past events. This allows the Hawkes process to capture the self-exciting properties of insurance claims. Specifically, the Hawkes process assumes that the occurrence of an event increases the probability of future events for a certain period of time. The intensity function $a_i^{FREE}(t)$ depends on a background rate $\exp(z_i^\top \gamma)\exp(\lambda_0)$ and a convolution of past events. Modeling outpatient claims using the Hawkes process can help understand the time scale of claim contagion. We refer readers to Appendix A for further details about the Hawkes process.

Elements in the model are:

- The background rate $\exp(z_i^\top \gamma)\exp(\lambda_0)$ represents the long-term average rate of occurrence in the absence of any triggering effect. The parameter λ_0 adjusts the rate for the free plan. This rate captures the intrinsic likelihood of an event happening due purely to exogenous factors.

- The exciting part $\tau_{[0,t]} = \int_0^t \exp(z_i^\top \gamma)\mu_0^2 \exp(a_0 - \mu_0^2(t-s))dN_i(s)$ captures event contagion and temporal dependence in data. The exciting strength $\exp(z_i^\top \gamma + a_0)\mu_0^2$ measures the increase in intensity due to the occurrence of an outpatient event. It determines the magnitude of the triggering impact. The decay rate μ_0^2 determines the time scale over which this triggering impact diminishes. We specify the decay rate as quadratic to ensure it remains non-negative: $\mu_0^2 = 0$ implies no triggering effect, while $\mu_0^2 > 0$ indicates such an effect exists.

HMO Plan. The Rand HMO plan differs from the free plan only in the restrictions placed on providers. If patients are generally responsible for initial contact in outpatient care but providers determine treatment after that, we would expect little

difference in healthcare usage between these two plans. To test this hypothesis, we specify the intensity of the HMO plan as:

$$\begin{aligned} a_i^{HMO}(t) &= \exp(z_i^\top \gamma) \left(\exp(\lambda_1) + \int_0^t \mu_1^2 \exp(a_1 - \mu_1^2 (t-s)) dN_i(s) \right) \\ &= \exp(z_i^\top \gamma) \left(\exp(\lambda_1) + \sum_{j: t_{ij} < t} \mu_1^2 \exp(a_1 - \mu_1^2 (t - t_{ij})) \right), \quad t \in \mathcal{T} \end{aligned} \tag{4}$$

Under the null, we have $\lambda_0 - \lambda_1 = a_0 - a_1 = \mu_0^2 - \mu_1^2 = 0$.

Individual Deductible Plan (ID). As discussed previously, if assigned to the individual deductible insurance plan, a person may react to both direct and indirect incentives. The direct effect measures increased intensity from changing expenditures $x_i(t)$ over time. The indirect effect stems from the state-dependent mechanism and represents a triggering effect. When this person exceeds the OOPC, the direct effect would disappear, and the intensity model would be identical to that of the free plan, aside from parameters. It is of considerable interest to test for and compare differences in health care consumption between the free plan and the individual deductible plan (after exhausting the OOPC). An important question in this regard is whether individuals exhibit "retaliatory spending" upon exceeding the OOPC. To this end, we will parameterize two intensities, one for the case where the cumulative spending is below the OOPC and one for the case where OOPC has been reached.

Suppose that the current time t is such that $x_i(t) < OOPC$, the intensity is specified as:

$$a_i^{ID}(t) = \exp(z_i^\top \gamma) \left(a_i^{direct}(t) + a_i^{indirect}(t) \right) \tag{5}$$

where the direct and indirect effects are specified respectively as:

$$a_i^{direct}(t) = \exp(\lambda_2)\exp(b^2(x_i(t) - OOPC)) \tag{6}$$

$$\begin{aligned} a_i^{indirect}(t) &= \exp(b^2(x_i(t) - OOPC)) \int_0^t \mu_2^2 \exp(a_2 - \mu_2^2(t-s)) dN_i(s) \\ &= \exp(b^2(x_i(t) - OOPC)) \sum_{j:t_{ij}<t} \mu_2^2 \exp(a_2 - \mu_2^2(t - t_{ij})) \end{aligned} \tag{7}$$

Elements in the model are:

- λ_2 adjusts the cost-sharing plan.

- The driving force for changes in the intensity level is the shadow coinsurance rate $c(x_i(t))$,

$$c(x_i(t)) = c_0(1 - \exp(b^2(x_i(t) - OOPC))) \tag{8}$$

  where c_0, as defined before, is the nominal coinsurance rate. The term $\exp(b^2(x_i(t) - OOPC))$ is the probability of exceeding the OOPC (Keeler and Rolph, 1988).

- $b^2 = 0$ implies that individuals are myopic and solely react to the spot coinsurance rate. In contrast, $b^2 > 0$ indicates that individuals are forward-looking and comprehend the dynamic nature of the cost-sharing policy.

- The cumulative expenditure also affects the triggering effect as described by $\exp(b^2(x_i(t) - OOPC))\mu_2^2 \exp(a_2 - \mu_2^2(t-s))$: The strength starts low when $x_i(t)$ is low, but continues growing as $x_i(t)$ increases.

Next, we specify the intensity when $t : x_i(t) > OOPC$. In this case, the patient has reached the OOPC and would have no out-of-pocket costs. The nominal and shadow coinsurance rates remain zero, and the plan is the same as the free plan in

terms of restrictions on patients.

$$\begin{aligned} a_i^{ID}(t) &= \exp(z_i^\top \gamma) \left(\exp(\lambda_3) + \int_0^t \mu_3^2 \exp(a_3 - \mu_3^2(t-s)) dN_i(s) \right) \\ &= \exp(z_i^\top \gamma) \left(\exp(\lambda_3) + \sum_{j:t_{ij}<t} \mu_3^2 \exp(a_3 - \mu_3^2(t-t_{ij})) \right) \end{aligned} \tag{9}$$

We use $\{\lambda_3, \mu_3^2, a_3\}$ to test for and compare differences in health care consumption between the free plan and post-OOPC individual deductible plan.

Dynamic Incentives

The dynamic incentive $\Delta(t)$ we considered here can be seen as increased healthcare utilization due to lower *shadow* healthcare prices. In this subsection, we analyze this dynamic incentive in the ID plan.

Recall that the shadow price at time t is defined as the conditional probability of not exceeding the OOPC given an individual i's cumulative expenditure: $c(x_i(t)) = 1 - \exp(b^2(x_i(t) - OOPC))$. Thus, we have:

$$\Delta_i(t) = \frac{\partial a_i^{ID}(t)}{\partial c(x_i(t))} = -\exp(z_i^\top \gamma) \left(\exp(\lambda_2) + \sum_{j:t_{ij}<t} \mu_2^2 \exp(a_2 - \mu_2^2(t-t_{ij})) \right) \tag{10}$$

The dynamic incentive itself is a stochastic process and varies randomly based on a person's attributes and outpatient history.This formula gives researchers a better understanding of an individual's health care decisions. A decrease in the shadow price would create the biggest incentive if someone recently had an outpatient visit. In contrast, if such a decrease occurs long after their last outpatient care, its impact is limited.

4. Estimation Method and Calculation Details

Estimation Method

Recall that we represent an individual's event times $\{t_j\}_{j\in[n]}$ as a counting process

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

Given this set of event times, we can estimate the parameters θ by maximizing the log-likelihood (Rubin, 1972):

$$\log L(t_1, \ldots, t_n|\theta) = -\int_0^T a(t|\theta) + \int_0^T \log a(t|\theta) dN(t). \tag{11}$$

where $a(t|\theta)$ is its intensity.

In our application, we have n observational processes $\{N_i(t)\}_{i\in[n]}$, where for each individual, there are $\{n_i\}_{i\in[n]}$ random occurrences of outpatient events over a time interval $\mathcal{T}$. We refer this kind of data as doubly stochastic, since for each person, both the event times and number of events are random variables. The fact that n_i is random complicates specifying the log-likelihood function. To calculate each log-likelihood contribution $\log L_i(t_{i1}, \ldots, t_{i\bar{n}}|\theta)$, we must fix the number of events ($\bar{n}$) for each individual. Thus, the overall log-likelihood function is $\log L(\theta) = \sum_{i=1}^{n} \log L_i(t_{i1}, \ldots, t_{i\bar{n}}|\theta)$.

However, adopting this strategy has two consequences. First, for individuals with $\{i : n_i < \bar{n}\}$, it is impossible to specify their likelihood contributions, and removing these individuals would introduce sample selection bias. Second, for individuals with $\{i : n_i > \bar{n}\}$, although the corresponding log-likelihood contributions can be specified, much information (i.e., $\{t_{ij} : j > \bar{n}\}$) is discarded, reducing estimation efficiency.

To overcome these challenges, we adopt a minimum distance estimation method, first proposed by Kopperschmidt and Stute (2013). This method relies on the Doob-

Meyer decomposition:

$$N_i(t) = A_i(t) + M_i(t)$$

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 counting 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) showed that this estimator is consistent and asymptotically normal. We briefly summarize their asymptotic and inference results in Appendix B.

Calculation Details

The application of minimum distance estimation is straightforward for the free plan and HMO plan. For example, in the free plan, an individual's compensator is calculated as

$$A_i^{FREE}(t) = \exp(z_i^\top \gamma)\left(\exp(\lambda_0)t + \sum_{j:t_{ij}<t} \exp(a_0)\left(1 - \exp(-\mu_0^2(t - t_{ij}))\right)\right)$$

However, the OOPC in the individual deductible plan introduces complications for estimation. As discussed previously, before and after surpassing the OOPC, individuals face different incentives. Thus, we should conceptualize the individual counting process $N_i(t)$ as the summation of two distinct counting processes:

$$N_i(t) = N_i^{before}(t) + N_i^{after}(t) \tag{12}$$

When time t is such that $t : x_i(t) < OOPC$, $N_i^{after}(t) = 0$; while when $t : x_i(t) \geq OOPC$, $N_i^{before}(t)$ remains unchanged. In practice, we estimate the parameters as follows. For an individual i, let $\tilde{t}_i$ and k_i be the last outpatient time and its position in the time set when cumulative spending is below the OOPC, i.e., $\tilde{t}_i = t_{ik_i}$. For $t < \tilde{t}_i$, construct the counting process $N_i^{before}(t)$ and write its compensator as:

$$A_i^{before}(t) = \exp(z_i^\top \gamma) \left(A_i^{direct}(t) + A_i^{indirect}(t) \right)$$

where

$$A_i^{direct}(t) = \sum_{j:t_{ij}<t} \exp(\lambda_2) \exp(b^2(x_i(t_{i(j-1)}) - OOPC))(t_{ij} - t_{i(j-1)})$$
$$+ \exp(\lambda_2) \exp(b^2(x_i(t_{ij}) - OOPC))(t - t_{ij}), \quad \text{with } t_{i0} = 0, x_i(0) = 0$$

and

$$A_i^{indirect}(t) = \sum_{j:t_{ij}<t} \left(\sum_{k=j}^{k_i-1} \exp(a_2 + b^2(x_i(t_{ik}) - OOPC)) \left(\exp(-\mu_2^2(t_{ik} - t_{ij})) - \exp(-\mu_2^2(t_{i(k-1)} - t_{ij})) \right) \right.$$
$$\left. + \exp(a_2 + b^2(x_i(\tilde{t}_i) - OOPC)) \left(\exp(-\mu_2^2(t_{i(k_i-1)} - t_{ij})) - \exp(-\mu_2^2(t - t_{ij})) \right) \right)$$

When $t \geq \tilde{t}_i$, we will use the whole counting process $N_i(t)$ to estimate parameters but modify the compensator as:

$$A_i^{after}(t) = \begin{cases} N_i(t), & t < \tilde{t}_i \\ \exp(z_i^\top \gamma) \left(A^{background}(t) + A_i^{triggering}(t) \right), & t \geq \tilde{t}_i \end{cases}$$

where

$$A^{background}(t) = \exp(\lambda_3)(t - \tilde{t}_i)$$

and

$$\begin{aligned} A_i^{triggering}(t) = & \sum_{j: t_{ij} < \tilde{t}_i} \exp(a_3) \left(\exp(-\mu_3^2(\tilde{t}_i - t_{ij})) - \exp(-\mu_3^2(t - t_{ij})) \right) \\ & + \sum_{j: \tilde{t}_i \leq t_{ij} < t} \exp(a_3) \left(1 - \exp(-\mu_3^2(t - t_{ij})) \right) \end{aligned}$$

5. Results

This section presents and compares the results of different insurance plans. First, we compare the free plan and the HMO plan, which differ only in provider restrictions. Since Rand HIE only offers the HMO plan in Seattle, we also present Seattle-area results for the free plan. Next, we compare the free plan and the individual deductible plan, which differ only in patient restrictions. Of particular interest are differences in patient behavior between the free plan and the post-OOPC individual deductible plan.

The goals of these comparisons are:

- *Free V.S. HMO:* To test the hypothesis that patients typically initiate contact for outpatient care but providers determine treatment afterward. If correct, we anticipate few differences in estimated parameters.

- *Free V.S. Individual Deductible:* To analyze how demand-side restrictions affect overall healthcare use. We focus on both direct and indirect impacts of cost-sharing policies.

- *Free V.S. Post-OOPC Individual Deductible:* To test for "retaliatory spending" after reaching the OOPC. Before surpassing the OOPC, patients may postpone some elective care, causing retaliatory spending.

These comparisons are based on dynamic parameter estimation results presented in Table 3. Lastly, we also present and discuss impacts from individual heterogeneities, presenting in Table 4.

Comparisons among Different Plans

Free V.S. HMO. Since Rand HIE only offers an HMO plan in Seattle, we estimate the free plan using data only from Seattle for comparison. The corresponding results are presented in columns 2 and 3 in Table 3. There are two key points: First, the decay rates in both plans are significantly different from zero. The Walt test statistics of $\hat{\mu}_0^2 = 6.961$ and $\hat{\mu}_1^2 = 16.518$ are greater than the critical value at significant level of $\alpha = 0.05$, so both plans show self-exciting properties. Second, all parameters (dynamic and individual heterogeneity parameters) are insignificantly different in both plans despite HMO restrictions. This strongly supports our hypothesis that patients drive initial contacts in outpatient care. Also, in Section 2, we rejected the hypothesis that free and HMO plan costs come from the same distribution. This implies that doctors determine treatment after initial contacts.

Free V.S. Individual Deductible. The results for the free plan (full sample) and the individual deductible plan (before exceeding the OOPC) appear in columns 1 and 4 of Table 3, respectively. Among the dynamic parameters, the decay rates

Table 3: Estimation Results for Dynamic Parameters

	FREE	FREE(Seattle)	HMO(Seattle)	ID-Before OOPC	ID-After OOPC
λ_0	-2.003	-1.874			
	(4.655)	(6.244)			
λ_1			-1.717		
			(9.853)		
λ_2				-1.997	
				(6.486)	
λ_3					-1.352
					(1.346)
μ_0	1.948	1.464			
	(0.627)	(0.331)			
μ_1			2.037		
			(0.536)		
μ_2				2.062	
				(0.456)	
μ_3					0.832
					(0.182)
a_0	0.848	1.128			
	(4.557)	(5.399)			
a_1			1.076		
			(9.312)		
a_2				0.670	
				(6.643)	
a_3					2.388
					(0.848)
b				1.320	
				(0.265)	
Individual Heterogeneities	YES	YES	YES	YES	YES

These parameters jointly determine dynamic properties of different models. $\{\lambda_k\}_{k=0,1,2,3}$ determine the background rates, $\{\mu_k\}_{k=0,1,2,3}$ determine the decay rates of outpatient events, $\{a_k\}_{k=0,1,2,3}$ together with μ's determine the exciting strength, and b determines the direct impact of the shadow coinsurance rate. We replace the cumulative cost $x(t)$ with $x(t)/100$ in the model to avoid overflow in computing. Consequently, the OOPC threshold is replaced by 1.5. One should use Chi-square to test whether $\{\mu_0^2, \mu_1^2, \mu_2^2, \mu_3^2, b^2\}$ are different from zero. Numbers in the parentheses are estimated standard errors.

Table 4: Estimation Results for Individual Heterogeneity Parameters

	FREE	FREE(Seattle)	HMO(Seattle)	ID-Before OOPC	ID-After OOPC
age	-0.535	-0.592	-0.526	-0.532	-0.517
	(0.140)	(0.221)	(0.263)	(0.237)	(0.313)
age2	0.576	0.550	0.601	0.571	0.530
	(0.176)	(0.405)	(0.348)	(0.332)	(0.487)
sex	-0.554	-0.605	-0.488	-0.370	-0.705
	(0.529)	(0.354)	(0.674)	(1.145)	(0.576)
edu	-1.065	-0.958	-0.953	-1.211	-1.156
	(0.221)	(0.602)	(1.106)	(0.248)	(0.413)
edu2	3.911	3.932	4.128	5.082	4.665
	(0.796)	(2.293)	(3.982)	(0.931)	(1.294)
income	1.716	1.725	1.413	1.910	1.592
	(0.473)	(0.604)	(0.512)	(0.601)	(0.431)

$age2$ and $edu2$ are defined as $age^2/100$ and $edu^2/100$, respectively to avoid overflow computing. *edu* is measured in terms of schooling years, *income* is defined as the logarithm of annual income (unadjusted inflation). Numbers in the parentheses are estimated standard errors.

between these two plans seem identical. Furthermore, the decay rate in the ID plan is significantly different from zero, indicating the indirect channel exists. However, the exciting strength in the ID plan is discounted by the shadow price effect. The OOPC remaining coefficient is also significantly different from zero, implying the direct channel also exists.

Free V.S. Post-OOPC Individual Deductible. Column 5 in Table 3 shows the results for the individual deductible plan (after exceeding the OOPC). The only difference between these two plans is the decay rate. The difference $\hat{\mu}_0 - \hat{\mu}_3 = 1.281$ has a standard error of $(0.67^2 + 0.116^2)^{1/2} = 0.680$ (since patients from the free and individual deductible plans are independent), indicating this difference is significantly larger than zero.

A slower decay rate in the exciting part under the post-OOPC ID plan indicates higher, longer-lasting self-excitement from past claims. Specifically, the function driving the triggering effect of past claims on future claims decays more gradually. This suggests that outpatient claims raise the risk of subsequent claims for a longer time, indicating stronger incentive.

Individual Heterogeneities

The interpretation of individual heterogeneity effects here resembles that of the marginal effect at a representative value (MER) in count data models when we fix a period and treat the counting process as count data. Specifically, let $Y_{it} = N_i(t)$ denote the number of events that occurred before time t. Let the scalar z_{ij} represent 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 $\exp(z_i^\top \gamma)$.

As individual heterogeneities across plans are statistically the same, we focus on the free plan (full sample) for interpretation.

Time-invariant explanatory variables include age, sex, education (in schooling years), and log-income as individual factors. We introduce age2 and edu2 to model the nonlinear impact of age and education; they are defined as $age^2/100$ and $edu^2/100$, respectively.

For free plan, we observe:

- *Age*. Initially, intensity values decrease as age increases. After age 47, intensity values and age are positively correlated. It is well established that younger individuals have higher risk compared to middle-aged counterparts. However, as individuals age, they become physically more vulnerable and prone to illness.

- *Gender*. Gender seems insignificant to outpatient activities.

- *Education and Income*.Income is positively associated with medical service utilization. Educational attainment results suggest a U-shaped relationship between education and outpatient utilization. Greater educational attainment is associated with lower outpatient activity, until approximately 13 years of schooling (roughly equivalent to a high school diploma). Thereafter, higher education is associated with greater likelihood of physician visits. One potential explanation is that higher education is often correlated with healthier lifestyles, reducing demand for outpatient care. However, above high school levels of education are also associated with higher income, enabling individuals to overcome opportunity costs related to work absence.

True or Spurious State-Dependence

Ever since Heckman (1981), unobserved heterogeneity has posed considerable challenges for empirical analyses of state-dependence. Failure to account for unobserved heterogeneity can yield spurious state-dependence: conditional on unobserved heterogeneity and other covariates, events may in fact be independent. Events may appear contagious in models that do not properly control for unobserved heterogeneity, as this confounding factor gives rise to the illusion of state-dependence.

It is legitimate and important to ask: Is the state-dependent effect (i.e., the triggering effect) observed in our model true or spurious? We believe the triggering effect in our model is genuine. The reasoning is as follows. For now, suppose the model has spurious state-dependence. Then, exciting functions $\sum_{j:t_{ij}<t} \exp(a_k - \mu_k^2(t-t_{ij})), k = 0, 1, 2, 3$ are purely results of an unobserved heterogeneity η_i, and are approximations of this unobserved heterogeneity.

We have shown that patients may spend more after exceeding their OOPC. This should affect parameters controlling the background rate or exciting strength if state-dependence is spurious. However, the results in Column 5 of Table 3 indicate that only the decay rate differs from the free plan (full sample). Since individuals' unobserved heterogeneities are unlikely to change just from exceeding the OOPC, the only explanation for the changed decay rate is genuine state-dependence. Thus, the exciting part in the intensity model should be understood as an approximation for both the unobserved heterogeneity and the state-dependence effect.

Comparing with Keeler and Rolph (1988)

Keeler and Rolph (1988) pioneered the use of event-based data to analyze moral hazard in health insurance. Their analysis of the Rand data yielded different results than ours. Here we discuss these differences and why they occurred. For the sake of

clarity, we briefly restate their findings: (1) Almost all of the reduction in medical use due to coinsurance comes from reduction on the number of events; (2) Event occurrence rates differ more between free care and cost-sharing plans than between high and low levels of coinsurance; (3) Outpatient events appear to arrive independently over time; (4) There is no retaliatory spending after patients reached their OOPC; and (5) Individuals are economic myopia, i.e., patients do not respond to dynamic incentives.

Our results differ most from the last three conclusions of Keeler and Rolph (1988): (1) Outpatient events are state-dependent and clustered over time; (2) Patients would consume more healthcare services after exhausting their OOPC; and (3) Individuals are forward-looking and understand dynamic incentives.

To analyze the differences, we need to comprehend their modelling strategy. First, they assume that several outpatient activities are actually part of a single "event" called the episode, and organize the data into episodes of treatment that contain all the spending associated with a given bout of illness, chronic condition, or well-care procedure. Next, they assume arrivals of episodes are Poisson distributed with an unobserved individual heterogeneity that follows a Gamma distribution, i.e., they fit the episode data with a negative binomial model.

By grouping related outpatient events into a single episode, they failed to capture the dynamic nature of the data. In contrast, our work models the temporal spread of events using a self-exciting process without aggregating the data. In fact, their compression method essentially "eliminated" the self-exciting aspect of our model, leaving behind a simple Poisson process. Different ways in modelling explain the first difference: seeing events as independent or influenced by previous occurrences.

The second difference can also be explained by differences in modelling strategies. As shown in Table 3, the major difference between the post-OOPC ID plan and the

free plan (full sample) is the decay rate. The background rate between these two plans remains the same, and in this sense, our result is consistent with that of Keeler and Rolph (1988). The retaliatory behaviors manifest exclusively within the exciting component and reflect solely in the event dependence structure.

To investigate at which price (shadow or nominal) patients respond, Keeler and Rolph (1988) split the before-OOPC period into high OOPC-remaining and low OOPC-remaining periods, and compare the occurrence rates between these two periods. They found no significant differences between these two rates and concluded that patients are economically myopic. In contrast, we find evidence that patients understand the dynamic incentives created by the shadow price. One explanation for this disparity is individuals' response to shadow price is nonlinear. Specifically, the response is small when OOPC-remaining is high but large when OOPC-remaining is low. For example, if we divide pre-OOPC situations into those with more or less than \$50 of OOPC remaining, a patient's response to a change at this cutoff (using our results) is just $\exp(1.32(0.5-1.5)) = 0.267$. If the division is more or less than \$100, the response would be would be $\exp(1.32(1-1.5)) = 0.517$.

6. Conclusion

This study provides evidence that patients respond to dynamic incentives in health insurance with deductibles. We propose a novel self-exciting approach that focuses on healthcare events, distinguishes direct and indirect dynamic incentives, and identifies retaliatory behaviors. Specifically, we analyze three health plans in the Rand HIE and find that patients are dynamically forward-looking in response to deductibles, while physicians do not change their prescribing behaviors. We quantify patients' dynamic

incentives using a self-exciting model and detect retaliatory behaviors subsequent to reaching the deductible limit.

References

Abaluck, J., J. Gruber, and A. Swanson (2018): "Prescription drug use under Medicare Part D: A linear model of nonlinear budget sets," *Journal of public economics*, 164, 106–138.

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

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.

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.

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.

Cromwell, J. and J. B. Mitchell (1986): "Physician-induced demand for surgery," *Journal of health economics*, 5, 293–313.

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.

Dalton, C. M., G. Gowrisankaran, and R. J. Town (2020): "Salience, myopia, and complex dynamic incentives: Evidence from Medicare Part D," *The Review of Economic Studies*, 87, 822–869.

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

Dranove, D. and P. Wehner (1994): "Physician-induced demand for childbirths," *Journal of health economics*, 13, 61–73.

Einav, L. and A. Finkelstein (2018): "Moral hazard in health insurance: what we know and how we know it," *Journal of the European Economic Association*, 16, 957–982.

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.

Ellis, R. P. (1986): "Rational behavior in the presence of coverage ceilings and deductibles," *The RAND Journal of Economics*, 158–175.

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.

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.

Johansson, N., C. Sonja, J. S. Kunz, D. Petrie, and M. Svensson (2023): "Reductions in out-of-pocket prices and forward-looking moral hazard in health care demand," *Journal of health economics*, 87, 102710.

Johansson, P. and M. Palme (2002): "Assessing the effect of public policy on worker absenteeism," *Journal of Human Resources*, 381–409.

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.

Klein, T. J., M. Salm, and S. Upadhyay (2022): "The response to dynamic incentives in insurance contracts with a deductible: Evidence from a differences-in-regression-discontinuities design," *Journal of Public Economics*, 210, 104660.

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.

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.

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.

Yip, W. C. (1998): "Physician response to Medicare fee reductions: changes in the volume of coronary artery bypass graft (CABG) surgeries in the Medicare and private sectors," *Journal of health economics*, 17, 675–699.

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.

Appendix A. Backgrounds on Hawkes Process

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

$$a(t) = \lambda_0 + \int_0^t g(t-s)dN(s) \tag{13}$$

$$= \lambda_0 + \sum_{j:t_j<t} g(t-t_j) \tag{14}$$

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. To see this, recall the free insurance plan, although the insurance coverage is fixed at $c_t = 0, \forall t \in \mathcal{T}$, the moral hazard is still dynamic:

$$\omega(0, \tau_{[0,t]}) = \sum_{j:t_j<t} g(t-t_j).$$

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

As for the cost-sharing plan (ID), we use a marked Hawkes process (Daley and Vere-Jones, 2007) to model dynamic incentives, where the shadow coinsurance rate works as marks:

$$a(t) = \lambda_0 + \omega(c_t, \tau_{[0,t]}) = \lambda_0 + \sum_{j:t_j<t} (1 - c(x(t_j)))g(t-t_j).$$

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. (14) 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 $n^* = \int g(s)ds$. 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 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. In case of the exponential kernel, the branching ratio is $n^* = \alpha/\mu$.

Appendix B. Asymptotic and Inference results of the Estimator

Let $\hat{\theta}_n$ be the minimum distance estimator, we have:

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

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

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

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

Appendix C. Data Generating Process and Simulation Studies

DGP and Simulation Algorithm

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).[1]. 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.

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

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

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

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

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

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

   go to step 6.

   Else no events occurred in interval $(\tau, \tau+\delta)$, and set the start point at $\tau \leftarrow \tau+\delta$ and return to step 2
6. Simulate a uniform random number U on the interval $(0, 1)$
7. If

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

   then a new 'event' occurs at time $t_i = \tau + \xi$. Simulate the associated marks for this new event.
8. Increase $\tau \leftarrow \tau + \xi$ for the next event simulation
9. Return to step 2

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 5: 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 6: 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 7: 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%