Static and Dynamic Incentives in Individual Outpatient Claims: Identification and Quantification*

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

This paper investigates the impact of static and dynamic incentives on patient behavior in the context of health insurance with deductibles. Using data from the Rand Health Insurance Experiment (Rand HIE) for analysis, we propose a novel approach that focuses on healthcare events rather than healthcare expenditures to identify and quantify patients' incentives from those of physicians. We utilize a conditional shadow price and explicitly specify the state-dependent structure of impacts of previous events on subsequent ones. The study's main findings reveal that patients respond to both nominal and shadow prices, but on average dynamic incentives have roughly four times greater impact compared to static incentives. Furthermore, incentive effects are not uniform across different individuals, with static incentives having a greater impact on "heavy users," and dynamic incentives affecting "light users" more. Lastly, we find that patients time their healthcare needs and exhibit retaliatory behaviors after reaching their deductible limits.

JEL Classification: D12,G22

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

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

1. Introduction

Deductibles are ubiquitous in the insurance market. In the United States, according to the Kaiser Family Foundation's 2022 Health Benefits Survey, 88% 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 may 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 higher nominal price. Previous studies do not reach a consensus on this issue: Aron-Dine et al. (2015); Einav et al. (2015); Johansson et al. (2023); 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

Another important yet often ignored question is "whose incentives?" Any data on healthcare utilization results from both patients and doctors. A number of previous studies (see, e.g., Currie et al. (2011); Einav et al. (2018); Eliason et al. (2018); Gottschalk et al. (2020); Gruber et al. (1999); Gruber and Owings (1996); Jacobson et al. (2010); Nguyen and Derrick (1997); Rice (1983); Rossiter and Wilensky (1984); Yip (1998)) support the existence of physician-side incentives and the so-called "supply-induced-demand" (SID), defined as excess healthcare use beyond what would have occurred if patients were fully informed. Yet, vast literature estimating the price sensitivity of *patients* ignores the difference between patients' and physicians' incentives.

The prevailing patient-physician relationship is asymmetric: patients exercise minimal agency over their treatment regimen or the associated costs, see, e.g., Arrow (1963). 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 both static and dynamic incentives therein. We question this choice of dependent variable, as health states and healthcare costs may not maintain a monotonic correlation (i.e., a healthier patient might have higher healthcare expenditures compared to a sicker patient) due to physicians' moral hazards. Therefore, comparisons among different insurance plans may be invalid.

The aim of this paper is to identify and quantify patients' responses to both the static and 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 isolate patients' incentives from those of physicians, we use healthcare events data rather than healthcare expenditures. 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. We hypothesize that variations in supply-side incentives would not affect individuals' outpatient initiations. This hypothesis aligns with Keeler and Rolph (1988), where they assume that patients initiate outpatient care events, while physicians determine subsequent treatments and their scale. We aim to test this hypothesis. Specifically, we analyze two Rand HIE plans:(1) A free plan that imposes no restrictions on patients or physicians. (2) A Health Maintenance Organization (HMO) plan that places restrictions on physicians. The hypothesis holds true if there are no differences in outpatient event frequencies between the free and HMO plans.

Our strategy for identifying and quantifying static and dynamic incentives depends on variations of nominal and shadow prices in different states within an insurance plan. A state is defined by the relative position of cumulative healthcare spending to the deductible, with two states identified: prior-deductible and post-deductible. Our strategy is also contingent upon variations of cost-sharing policies across different insurance plans. In this study, we examine two insurance plans: a free plan

and an individual deductible (ID) plan. The former covers all healthcare costs while the latter requires individuals to pay a proportion of outpatient costs until reaching the deductible, after which the plan covers all costs. Furthermore, The ID plan only impose deductibles on outpatient use, while inpatients are fully covered. In the free plan, nominal and shadow prices are zero in both states, indicating the absence of both static and dynamic incentives. In contrast, the individual deductible plan features non-zero nominal and shadow prices in the prior-deductible state, indicating the presence of both static and dynamic incentives. A special case arises when cumulative healthcare spending equals the deductible in the prior-deductible state, resulting in a non-zero nominal price but a zero shadow price, indicating the presence of only static incentives. In the post-deductible state, both nominal and shadow prices are zero, but patients may still consume additional healthcare services due to the suppressing effect of the prior-deductible period, leading to retaliatory behaviors. Our model is capable of generating relevant statistics for different plan-states, which can be compared to identify and quantify static and dynamic incentives, as well as retaliatory behaviors, offering a more nuanced understanding of their impact on healthcare behaviors.

Methodologically, 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 incentives created by cost-sharing policies, and provides a better understanding of outpatient patterns.

Our approach further contributes to the literature in the sense that it: (1) Uses conditional rather than the unconditional shadow price. In our model, an individual's shadow price is defined as a conditional probability of not exceeding the deductible, given this individual's own cumulative healthcare spending up to the current time. (2) Provides a parametric framework to analyze the effects of incentives on the healthcare decisions of individuals when their health status is updated. Notably, we incorporate the updating of individuals' awareness of their health status whenever they conduct

outpatient activities. (3) Quantifies dynamic incentives under two sets of counterfactual analyses: High-Deductible and Copayment with deductible plans.

Related to (1), early theoretical literature on patients' responses to cost-sharing incentives concluded that, under certain assumptions, forward-looking individuals would 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 shadow price 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. Here, we impose a parametric structure on this shadow price where the cumulative healthcare spending is the determinant.

Related to (2), we posit that patients update their awareness of their health status upon visiting a doctor, leading to potential adjustments in their healthcare decisions. Consequently, recent outpatient activity may impact the probability of further outpatients in the near future, signifying state-dependence in outpatient events. This state-dependent structure offers a potential tool for measuring patients' reactions to changes in their health status. It is also worthwhile to investigate the impact of deductibles on this state-dependence. For instance, in the presence of both static and dynamic incentives, patients are likely to reduce their healthcare needs, leading to some needs disappearing while others may be postponed to the post-deductible period. This may result in individuals timing their healthcare needs and exhibiting retaliatory behaviors, where previous outpatients would have a greater impact on future ones placed in the post-deductible region.

Related to (3), high-deductible health plans (HDHPs) and copayment with deductible plans are gaining increasing attention within the health insurance industry. HDHPs are characterized by a higher deductible but lower premiums than traditional insurance plans. In contrast to coinsurance rates, which require patients to pay a fixed percentage of healthcare costs, copayments involve a fixed payment amount from patients regardless of the actual expenses incurred for healthcare services. Our approach employs models that are built on the dynamics of events, with parametric shadow

prices and state-dependence effects as driving forces. This enables us to simulate the properties of these two different health insurance designs.

Our main findings can be summarized as the followings. (1) Patients respond to both nominal and shadow prices. (2) Static and Dynamic incentives differ in ways that affect the temporal dependence structure and scales. On average, dynamic incentives have roughly four times greater impact compared to static incentives. Additionally, we found that static incentives would shrink the cluster size, while dynamic incentives would reduce the overall excitement strength. (3) Incentive effects are not uniform across different individuals. Static incentives have a greater impact on "heavy users," while we found no significant dynamic incentive effects in this group. In contrast, dynamic incentives affect "light users" more, and individuals in this group do not respond to static incentives. (4) Our counterfactual analyses reveal that although a high-deductible plan would increase dynamic incentives, a copyament cost-sharing policy would achieve the same goal while keeping individuals' out-of-pocket fees low.

Our paper 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, while 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 paper 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 paper is organized as follows. In section 2, we discuss data, the selection of dependent variables and sample construction procedures. Model specifications are provided in section 3. In section 4, we present a minimum distance method for estimating parameters. Section 5 presents the estimation results, followed by a quantification of static and dynamic incentives in Section 6. Section 7 conducts two sets of counterfactual analyses: high-deductibles and copayments with deductibles. Lastly, Section 8 concludes the whole paper.

2. Data, Dependent Variable and Sample Construction

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 (OOPCs). Due to the randomness of the assignments and the nonlinear cost-sharing features, the RAND HIE data is particularly suitable for studying dynamic incentives.

2.1 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.

In this study, 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 (ID) 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. To streamline the presentation, we will refer the OOPC as the deductible threshold (DT) hereafter.

2.2 Dependent Variable: Spending or Counts

A natural strategy for identifying and quantifying incentives is to find and measure a statistical difference between the free plan and the ID plan: $Y_{\mathcal{T},Free} - Y_{\mathcal{T},ID}$ within a time interval $\mathcal{T} = [0, T]$. In this subsection, we will discuss the selection of these statistics.

2.2.1 Spending V.S. Counts

Most literature uses healthcare spending to construct dependent variables, see, e.g., Aron-Dine et al. (2015); Brot-Goldberg et al. (2017); Einav et al. (2015); Guo and Zhang (2019); Johansson et al. (2023); Klein et al. (2022),etc. To validate such a choice, certain assumptions are necessary. Let s_t describe the health state of individuals at time t, and let F denote the distribution of that health state. The first assumption reads:

$$F_{Free}\left(s_t|Z_t\right) = F_{ID}\left(s_t|Z_t\right) \tag{1}$$

where Z_t contains individuals' characteristics at time t. This assumption implies that, given ex ante health status and demographics, the dynamic evolution of population health needs throughout the time interval $\mathcal{T}$ is the same in the free plan and the ID plan. Therefore, individuals do not, on average, become sicker throughout the time interval due to the effects of different insurance plans.

Second, to implement analysis based on spending, it is necessary to assume that there are one-to-one monotonic mappings between s_t (which is unobserved) and $S_{t,Free}$ and $S_{t,ID}$ (the cumulative spendings under the free and ID plans). This assumption implies that, for example, if the first 25% of patients have s_t values that place them in the coinsurance region for the ID plan when ranked by cumulative spending, these patients can be directly compared to the first 25% of patients from the free plan.

We believe that the first assumption is justifiable and shall be maintained in this study. While the dynamic evolution of health states plays a crucial role in individual healthcare decision-making, incorporating this element into the model is a complex process that does not yield significant benefits for identifying and quantifying incentives within a relatively short time interval (e.g., one contract year). Consequently, we shall further assume that at any time t, the health state s_t is drawn from the same conditional distribution, given an individual's characteristics.

However, we question the validity of the second assumption. The problem is that due to physicians' moral hazards, mappings from health states to healthcare costs might not be monotonic. Previous literature has identified moral hazards on the supply-side. For instance, Currie et al. (2011) found that when patients have knowledge of appropriate antibiotics use, antibiotic prescription rates and drug expenditures decrease. Gruber and Owings (1996) studied an exogenous change in the financial environment facing obstetrician/gynecologists during the 1970s and found

that an increased income pressure on ob/gyns led them to substitute normal childbirth with a more highly reimbursed alternative, cesarean delivery. Rossiter and Wilensky (1984) discovered that patient health insurance statuses and types are crucial determinants of physician-induced expenditures. Specifically, patients with Medicare and private health insurance have higher physician-induced and ambulatory expenditures on average than those without health insurance. These pieces of evidence suggest that in the presence of supply-side moral hazards, mappings from s_t to S_t might not be monotonic. Depending on various factors, patients with $s_t < s_t'$ might end up spending $S_t > S_t'$. For example, a sicker patient who is under the free plan with enough knowledge about the appropriate use of antibiotics and has a physician with less income pressure might have lower healthcare expenditures than a healthier patient under the ID plan who lacks knowledge on antibiotics and has a physician with heavy financial pressures.

Even if the monotonic assumption holds, quantifying patients' incentives can still present challenges due to moral hazards from physicians. For example, suppose we find evidence that higher (smaller) values of cumulative out-of-pocket fees (shadow prices) correlate with higher expenditures. Can we interpret this evidence as supporting the existence of dynamic incentives on the individual side? It is possible that patients do not respond to shadow prices, but physicians do. This can result in the prescription of much higher-valued treatments, even for the same disease. We have found signs of variations of physician-side incentives in the ID plan. We examined expenditure distributions before and after the deductible threshold. In the absence of supply-side moral hazard, one would expect these two distributions to be identical. However, our findings indicate a significant deviation from this expectation.

To begin, we summarize various statistics regarding expenditures before and after the deductible in the contract year of 1978 (Table 1). We use the standard wild bootstrap procedure to draw inferences[1]. The results reveal that expenditures are higher in the post-deductible period than in the prior-deductible period in almost every quantile. We also plot the estimated densities of the two distributions (Figure 1). The estimation is done using the standard kernel density procedure.

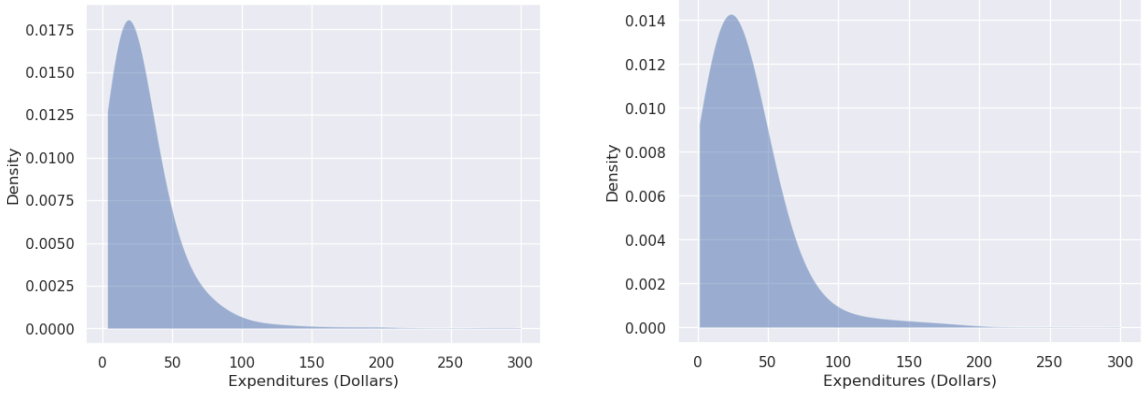
There are multiple factors that can contribute to inconsistencies in healthcare spending. For example, individuals may strategically time their healthcare needs and

1. Specifically, we resample expenditures with replacement. On each iteration of resample, the sample size is $n = 1000$, and we repeat $b = 1000$ iterations.

Table 1: Descriptive Statistics on Expenditure Distributions

Statistics	Prior-Deductible	Post-Deductible
Mean	34.755 (1.97)	43.682 (2.741)
Maximum	848.084 (142.332)	1134.724 (244.162)
Quantile 75	36.221 (1.498)	41.377 (1.552)
Quantile 50	19.741 (0.541)	25.441 (1.579)
Quantile 25	12.78 (0.461)	14.968 (0.159)

Unit: U.S. dollars. Standard wild bootstrap procedure is used to draw inferences. Numbers in the parentheses are empirical standard errors.



(a) Density of Prior-Deductible Expenditure

(b) Density of Post-Deductible Expenditure

Figure 1: Expenditure Densities for Prior and Post-Deductible

postpone discretionary yet costly expenditures until after reaching their deductible. While these explanations are plausible, they do not necessarily preclude the presence of supply-side moral hazards that may compromise the use of spending as a dependent variable.

Although healthcare spending may not be an unbiased proxy for an individual's health state s_t, it is still significantly correlated with s_t. A closer examination reveals that the total spending S_t can be expressed as a product of the number of healthcare consumptions C_t up to time t, and the average spending $\bar{S}$:

$$S_t = C_t \cdot \bar{S} \tag{2}$$

We believe that supply-side moral hazards will not significantly impact C_t. Formally, we make two key assumptions. Firstly, we assume that there exist one-to-one, monotonic mappings between s_t and $C_{t,Free}$ and $C_{t,ID}$, which represent outpatient counts under the free and ID plans, respectively. Secondly, we assume that variations in supply-side moral hazards would not affect these monotonic mappings. These assumptions align with the one made in Keeler and Rolph (1988), which states that patients trigger healthcare events while physicians determine subsequent treatments and their scale. We posit that the first assumption is both reasonable and easily justifiable, and our focus is on providing evidence for the second assumption.

The Rand HIE data includes an HMO plan, which helps us verifying the second assumption. A Health Maintenance Organization (HMO) is a type of health insurance plan that provides comprehensive healthcare coverage through a network of doctors, hospitals, and other healthcare 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 usually have lower out-of-pocket costs than other plans since the network of providers and health care services is tightly managed to control costs (i.e., place restrictions on the provider side). As a result, expenditure distributions of an HMO plan and the free plan are different. The Rand HMO plan does not impose any financial constraints, such as deductibles, coinsurance rates, or out-of-pocket caps, on patients. We conduct a comparative analysis between this plan and the free plan (in the third contract year) to investigate the impact of supply-side moral hazards on healthcare counts. It is important to note that the outpatients

covered by the HMO plan are a subset of those covered by the free plan. We only consider outpatients covered by both plans when constructing count statistics.

Table 2 presents the descriptive statistics for outpatient data from the free plan and the HMO plan. The statistics demonstrate that the outpatient events between the two plans are statistically similar at conventional levels. This finding provides evidence that supports our claim: variations in supply-side incentives would not affect individuals' outpatient initiations.

Table 2: Descriptive Statistics on Count Distributions

Statistics	HMO	Free Plan
Mean	5.027	6.079
	(0.645)	(0.835)
Maximum	54.345	70.281
	(37.287)	(26.951)
Quantile 75	6.099	6.527
	(0.709)	(0.98)
Quantile 50	3.022	3.033
	(0.389)	(0.339)
Quantile 25	1.064	1.013
	(0.266)	(0.144)

The Rand HIE only offers the HMO plan in the Seattle area. As a result, we only use participants from the same area who are enrolled in the free plan to construct the free plan sample. Within the HMO plan, there are both experimental and control groups. Only the experimental group consists of randomly assigned participants. Therefore, we exclude individuals from the control group to form the HMO sample. Standard wild bootstrap procedure is used to draw inferences. Numbers in the parentheses are empirical standard errors.

2.2.2 Data Representation

In the literature, the conventional method for constructing count data involves aggregating the number of outpatients within a given time interval $\mathcal{T}$. For example, Keeler and Rolph (1988) partition the time into high deductible-remaining, low deductible-remaining, and post-deductible intervals and count the number of outpatient events within each interval. They also utilize a negative binomial model to characterize the distribution of counts in each of these three intervals. This methodology entails the implicit assumption that outpatient events within each interval are

independent, allowing them to be aggregated, and that the occurrence rate of each event within intervals is constant.

However, outpatient events might be state-dependent as evidenced in Figure 2, where incidence times of an individual with the free plan are observed to be temporally correlated. In the absence of static and dynamic incentives due to the free insurance plan, event-based state-dependence may contribute to this clustered structure. In

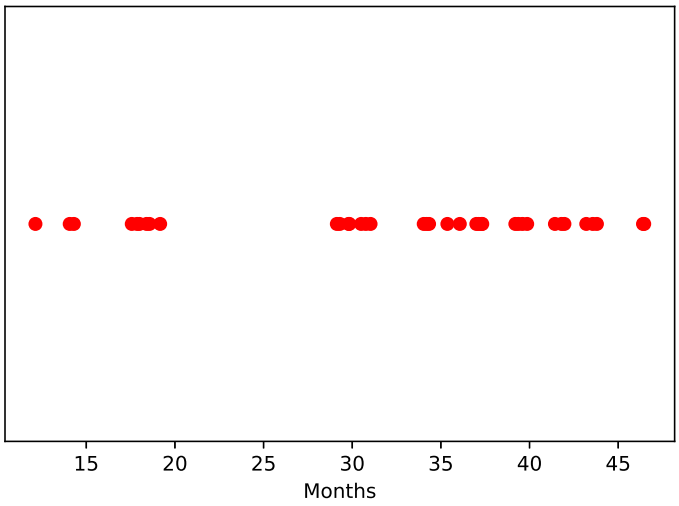


Figure 2: Outpatient claims instances over time under the Rand HIE free plan

order to develop a comprehensive understanding of the impact of deductibles on outpatient utilization, particularly with regards to their temporal dependence structure, we have adopted an innovative self-exciting process approach instead of traditional aggregation methods. The self-exciting process is a specialized counting process that provides researchers with the ability to investigate state-dependent effects in a more nuanced manner.

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, equals 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 conditional 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, leading to the state-dependent structure. 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 the context of health insurance claims, a self-exciting process model is appropriate due to the presence of state-dependent patterns in the data. Additionally, the state-dependent structure presents a potential tool for identifying and quantifying individuals' retaliatory behaviors. This is particularly pertinent in cases where such behaviors exist, as previous events could have greater impacts on future ones occurring in the post-deductible region.

2.3 Sample Construction

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 3 shows the remaining sample sizes and line-item counts after applying each major exclusion.

Table 3: 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

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.

3. The Econometric Specification

3.1 Dynamics in the Model

Before presenting our econometric models, we briefly discuss two key dynamic components: deductibles 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 price, defined as the expected coinsurance rate at the end of the year given the accumulated spending so far $x(t)$:

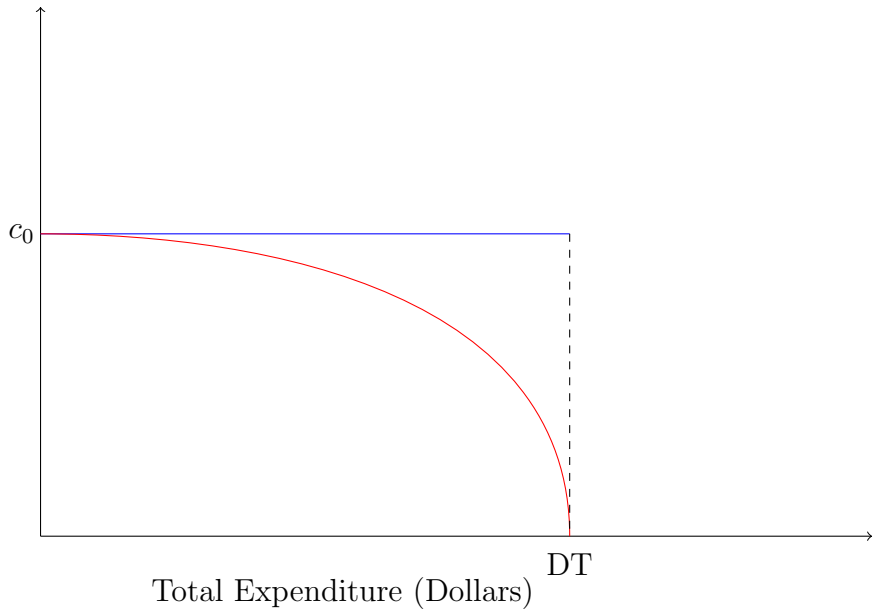
$$c_t = c(x(t)) = \mathbb{E}(c_{EOY} \mid x(t)) \tag{3}$$

This shadow rate therefore captures the dynamic, path-dependent nature of health insurance plans with deductibles. As the nominal coinsurance rate at EOY can be either the nominal rate c_0 defined by the insurance plan or zero (in which case, the individual must have exhausted the DT), the shadow price is:

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

This definition implies (1) patients are forward-looking: they 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 price $c(x)$ is concave, with $c' < 0, c'' < 0$. Effectively, we are assuming that individuals would respond more aggressively as the total expenditure approaches the deductible threshold. The exact form of $c(x)$ will be specified in the next subsection. Figure 3 illustrates a typical scenario: The blue line represents the nominal coinsurance rate c_0 when total spending is under the DT, dropping to 0% thereafter. The red curve shows the individual's shadow price, which also reflects insurance coverage.



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

Figure 3: Nominal and Shadow Prices

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

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

$$\tau_{[0,t]} = \sigma(\{t_j\} : t_j < t) \tag{4}$$

3.2 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 prices are fixed at zero. Therefore, its intensity model will focus on the state-dependent (the triggering) dynamic mechanism.

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

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

$$\begin{aligned} a_i^{FREE}(t) &= \exp(z_i^\top \gamma + \lambda_1)\left(1 + \tau_{[0,t]}\right) \\ &= \exp(z_i^\top \gamma + \lambda_1)\left(1 + \int_0^t \mu_1^2 \exp(a_1 - \mu_1^2(t-s)) dN_i(s)\right) \\ &= \exp(z_i^\top \gamma + \lambda_1)\left(1 + \sum_{j:t_{ij}<t} \mu_1^2 \exp(a_1 - \mu_1^2(t - t_{ij}))\right), \quad t \in \mathcal{T} \end{aligned} \tag{5}$$

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 $\exp(z_i^\top \gamma)\exp(\lambda_1)$ 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 $\exp(z_i^\top \gamma + \lambda_1)$ represents the long-term average rate of occurrence in the absence of any triggering effect. The parameter λ_1 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 $\int_0^t \exp(z_i^\top \gamma)\mu_1^2 \exp(\lambda_1 + a_1 - \mu_1^2(t-s))dN_i(s)$ captures event contagion and temporal dependence in data. The exciting strength $\exp(z_i^\top \gamma + a_1 + \lambda_1)\mu_1^2$ 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: $\mu_1^2 = 0$ implies no triggering effect, while $\mu_1^2 > 0$ indicates such an effect exists.

ID Plan. 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 DT, the direct effect would disappear, and the intensity model would be identical to that of the free plan, aside from parameters. It is of significant interest to test for and compare differences in health care utilization between the free plan and the individual deductible plan (both prior to and after reaching the DT). When comparing the free plan with the individual deductible plan before reaching the DT, we focus on the situation where deductible remaining is zero, where both plans have the same shadow price (zero) but differ in nominal prices. We can infer whether individuals respond to the nominal price by comparing outpatient intensities of these two plans. When comparing the free plan with the individual deductible plan after reaching the DT, we aim to determine whether individuals exhibit "retaliatory spending" upon exceeding the DT. To this end, we will parameterize two intensities, one for the case where cumulative spending is below the DT and one for the case where the DT has been reached.

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

$$a_i^{ID}(t) = \exp(z_i^\top \gamma + \lambda_2)\left(a_i^{direct}(t) + a_i^{indirect}(t)\right) \tag{6}$$

where the direct and indirect effects are specified respectively as:

$$a_i^{direct}(t) = \exp(b^2(x_i(t) - DT)) \tag{7}$$

$$
\begin{aligned}
a_i^{indirect}(t) &= \exp(b^2(x_i(t) - DT)) \int_0^t \mu_2^2 \exp(a_2 - \mu_2^2(t-s)) dN_i(s) \\
&= \exp(b^2(x_i(t) - DT)) \sum_{j:t_{ij}<t} \mu_2^2 \exp(a_2 - \mu_2^2(t - t_{ij}))
\end{aligned} \tag{8}
$$

Elements in the model are:

- λ_2 adjusts the cost-sharing plan.

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

$$
c(x_i(t)) = (1 - \exp(b^2(x_i(t) - DT))) \tag{9}
$$

  The term $\exp(b^2(x_i(t) - DT))$ is the probability of exceeding the DT (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 $\exp(b^2(x_i(t) - DT))\mu_2^2 \exp(a_2 - \mu_2^2(t-s))$: The strength starts low when $x_i(t)$ is low, but continues growing as $x_i(t)$ increases.

Next, we specify the intensity when $t : x_i(t) > DT$. In this case, the patient has reached the DT 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.

$$
\begin{aligned}
a_i^{ID}(t) &= \exp(z_i^\top \gamma + \lambda_3) \left(1 + \int_0^t \mu_3^2 \exp(a_3 - \mu_3^2(t-s)) dN_i(s)\right) \\
&= \exp(z_i^\top \gamma + \lambda_3) \left(1 + \sum_{j:t_{ij}<t} \mu_3^2 \exp(a_3 - \mu_3^2(t - t_{ij}))\right)
\end{aligned} \tag{10}
$$

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-deductible plan.

4. Estimation Method and Calculation Details

4.1 Estimation Method

Denote $[n] = \{1, 2, \ldots, n\}$ and 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. To examine the performance of the minimum distance estimators, we conduct simulation studies in Appendix C.

4.2 Calculation Details

The application of minimum distance estimation is straightforward for the free plan. In this plan, an individual's compensator is calculated as

$$A_i^{FREE}(t) = \exp(z_i^\top \gamma + \lambda_1)\left(t + \sum_{j:t_{ij}<t} \exp(a_1)\left(1 - \exp(-\mu_1^2(t - t_{ij}))\right)\right)$$

However, the DT in the individual deductible plan introduces complications for estimation. As discussed previously, before and after surpassing the DT, 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 $t : x_i(t) < DT$, $N_i^{after}(t) = 0$; while when $t : x_i(t) \geq DT$, $N_i^{before}(t)$ 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 DT, 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:

$$A_i^{before}(t) = \exp(z_i^\top \gamma + \lambda_2) \left(A_i^{direct}(t) + A_i^{indirect}(t) \right)$$

where

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

and

$$\begin{aligned} A_i^{indirect}(t) = \sum_{j:t_{ij}<t} \Bigg(& \sum_{k=j}^{k_i-1} \exp(a_2 + b^2(x_i(t_{ik}) - DT)) \\ & \times \left(\exp(-\mu_2^2(t_{ik} - t_{ij})) - \exp(-\mu_2^2(t_{i(k-1)} - t_{ij})) \right) \\ & + \exp(a_2 + b^2(x_i(\tilde{t}_i) - DT)) \left(\exp(-\mu_2^2(t_{i(k_i-1)} - t_{ij})) - \exp(-\mu_2^2(t - t_{ij})) \right) \Bigg) \end{aligned}$$

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

$$A_i^{after}(t) = \begin{cases} N_i(t), & t < \tilde{t}_i \\ \exp(z_i^\top \gamma + \lambda_3) \left(A^{background}(t) + A_i^{triggering}(t) \right), & t \geq \tilde{t}_i \end{cases}$$

where

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

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 the free plan and the individual deductible plan, which differ only in patient restrictions. Goals of these comparisons are:

- *Free V.S. Prior-Deductible:* First, to analyze how demand-side restrictions affect overall healthcare use. We focus on both direct and indirect impacts of cost-sharing policies. Second, to test whether patients respond to the nominal price by focusing on a situation where the deductible-remaining is zero.
- *Free V.S. Post-Deductible:* To test for "retaliatory spending" after reaching the DT. Before surpassing the DT, patients may postpone some non-essential healthcare needs, causing retaliatory spending in post-DT period.

These comparisons are based on dynamic parameter estimation results presented in Table 4. Lastly, we also present and discuss impacts from individual heterogeneities, presenting in Table 5.

5.1 Comparisons among Different Plans

Free V.S. Prior-Deductible. The results for the free plan and the individual deductible plan (before exceeding the DT) appear in columns (1) and (2) of Table 4, respectively. Among the dynamic parameters estimated, the decay rate in the ID plan is significantly different from zero at conventional levels, indicating the presence of an indirect channel. However, the exciting strength in the ID plan is attenuated by the shadow price effect. The deductible-remaining coefficient is also significantly different from zero at conventional levels, implying the existence of a direct channel as well.

The parametric specification of remaining deductibles facilitates an analysis of a situation where the nominal price remains unchanged in the prior-deductible plan yet the shadow price between these two plans is equivalent (i.e., zero): $x_i(t) = DT$. Under this circumstance, the effect of the deductible reduces to a unit. We ascertained that decay rates in these two plans are statistically significantly different. Specifically, in the prior-deductible plan, previous outpatient events decay at a faster rate and thus have a shorter triggering period. This suggests a weak temporal contagious effect in this plan. We interpret this result as evidence that patients respond to the nominal price in addition to the shadow price.

Table 4: Estimation Results for Dynamic Parameters

	FREE	Prior-Deductible	Post-Deductible
	(1)	(2)	(3)
λ_1	-2.539		
	(0.834)		
λ_2		-2.501	
		(0.831)	
λ_3			-1.931
			(0.727)
μ_1	1.153		
	(0.401)		
μ_2		2.853	
		(0.919)	
μ_3			0.336
			(0.149)
a_1	1.235		
	(0.691)		
a_2		2.699	
		(0.581)	
a_3			1.186
			(0.356)
b		1.353	
		(0.273)	
Individual Hetergeneities	YES	YES	YES

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

Table 5: Estimation Results for Individual Heterogeneity Parameters

	FREE	Prior-Deductible	Post-Deductible
	(1)	(2)	(3)
age	-0.055	-0.059	-0.069
	(0.055)	(0.035)	(0.031)
sex	-0.956	-0.823	-0.992
	(1.796)	(0.689)	(0.633)
Socioeconomic status	0.961	1.681	0.706
	(0.537)	(0.591)	(0.488)

Age is measured as real age-18. Socioeconomic status measure is $LINC + 0.2EDU - 10.51$ where $LINC$ = log income in year, EDU = years of education. A similar measure is also used in Keeler and Rolph (1988). Numbers in the parentheses are estimated standard errors.

Free V.S. Post-Deductible. The results for the individual deductible plan after exceeding the DT are shown in Column (3) of Table 4. This plan differs significantly from the free plan in its decay rate. The difference in decay rates, $\hat{\mu}_1 - \hat{\mu}_3 = 0.817$, has a standard error of $(0.401^2 + 0.149^2)^{1/2} = 0.428$.

A slower decay rate in the exciting component of the post-deductible plan indicates higher and more prolonged impacts of past claims: The function driving the triggering effect of past claims on future claims decays more gradually under the post-deductible ID plan. Given that both plans impose no restrictions on patients, one possible explanation is the "retaliatory spending". Since individuals respond to both nominal and shadow prices, they may curtail some discretionary healthcare needs before reaching the DT. However, these needs do not necessarily disappear, and upon reaching the DT, individuals might restore these needs, which could explain the slower decay rate.

5.2 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 almost statistically identical, we focus on the free plan for interpretation. Time-invariant explanatory variables include age, sex, education (in schooling years), and log-income as individual factors. We create a socioeconomic measure as $LINC + 0.2EDU - 10.51$ where $LINC =$ log income, $EDU =$ years of education. A similar measure is also used in Keeler and Rolph (1988). For free plan, we observe that only socioeconomic status positively correlate with outpatient frequency. While age and gender do not significantly impact outpatient activity.

5.3 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 $\sum_{j:t_{ij}<t} \exp(a_k - \mu_k^2(t - t_{ij})), k = 1, 2, 3$ are purely results of and approximated to an unobserved heterogeneity η_i.

We have shown that patients may spend more after exceeding their DT. This should affect parameters controlling the background rate or exciting strength if state-dependence is spurious. However, the results in Column (3) of Table 4 indicate that only the decay rate parameter differs from the free plan, i.e., $\mu_3 - \mu_1 \neq 0$. Since individuals' unobserved heterogeneities are unlikely to change just from exceeding the DT, the only explanation for the changed parameter is the 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.

6. Quantification of Incentives

In the preceding section, we presented empirical evidence that individuals respond to both static and dynamic incentives and display retaliatory behaviors. In this section, we aim to quantify the effects of these incentives. We employ two quantification approach as described in the next two subsections.

6.1 Quantification based on Variations of Nominal and Shadow Prices

The first approach (henceforth "Approach A") is grounded in the observation that nominal and shadow prices differ across plans and states, which enables us to measure the effects of these differences. The states are distinguished by whether cumulative individual expenditures exceed the DT. Table 6 provides a summary of the variations in these prices across different insurance plans and states, which we utilize as the foundation for our analysis. However, given the intricate intrinsic mechanisms of these incentives, we have opted to adopt a simulation strategy.

Table 6: Nominal and Shadow Prices at Different Plans in Different States

	State1: $x(t) < DT$	State2: $x(t) = DT$	State3: $x(t) > DT$
FREE	NP $= 0$ SP $= 0$	NP $= 0$ SP $= 0$	NP $= 0$ SP $= 0$
Individual Deductible	NP $= c_0$ SP $=$ $1 - \exp(b^2(x(t) - DT))$	NP $= c_0$ SP $= 0$	NP $= 0$ SP $= 0$

In this table, NP, SP are short for nominal price and shadow price, respectively. $x(t)$ is the cumulative expenditure at time t, and c_0 is the nominal coinsurance rate.

Specifically, we simulate counting processes for individuals within a time period $\mathcal{T}$ for each plan-state and obtain a set of outpatient counts $\mathcal{S} = \{N_i(\mathcal{T})\}_{i\in[n]}$. From this set, we can construct various statistics, such as the mean, maximum, and quantiles, denoted by Y, for each specific plan-state. We then use these statistics to quantify incentives. For example, the difference between the free and individual deductible plans in state 1 arises from the overall effects of both the static and dynamic incentives, measured by $Y_{Free} - Y_{ID-State1}$.

To simulate the counting process, we employ Ogata's thinning method (Ogata, 1981). The algorithm involves generating a Poisson process and then selecting a subset of the points based on the thinning probability. The thinning probability is calculated using the intensity function of the self-exciting process. By using Ogata's thinning method, we can simulate self-exciting point processes that exhibit clustering and temporal dependence. This method has proven to be very useful in modeling earthquake occurrences, neuronal spikes, and financial transactions. Appendix C describes detaild procedures of this algorithm. It should be noted that, to maintain economic interpretation, we impose constraints on the duration between consecutive events and the total number of outpatients. Specifically, we require the duration between events to be greater than one day and the total number of outpatients to be less than 365. The imposition of additional restrictions in the simulation algorithm results in a reduction of the number of events generated. Consequently, the simulation outcomes should not be utilized to assess the adequacy of the model.

Table 7 summarizes the measurements for various incentives. Interpretations of these quantifications are:

- Static+Dynamic Incentives: The amount of outpatients decreased when individuals responded to both dynamic and static incentives and did not exceed the DT within the time interval $\mathcal{T}$;
- Static Incentives: The amount of outpatients decreased when individuals only respond to static incentives and did not exceed the DT within the time interval $\mathcal{T}$;
- Dynamic Incentives: The amount of outpatients decreased when individuals only respond to dynamic incentives and did not exceed the DT within the time interval $\mathcal{T}$;
- Retaliatory Behaviors: The amount of outpatients increased due to the postponement of healthcare needs within the time interval $\mathcal{T}$.

Importantly, these strategies are based on artificial counting processes. That is, the processes corresponding to ID-States (State 1 to State 3) are not likely to occur in practice. For instance, ID-State1 represents a scenario where the deductible exists, but individuals would never exceed such a threshold within the time period $\mathcal{T}$. ID-State2 corresponds to a situation where patients only respond to spot price, while

Table 7: Quantification Strategies for Various Incentives, Approach A

Incentives	Strategies	Comments
Static+Dynamic	$Y_{Free} - Y_{ID-State1}$	Both NP and SP are different.
Static	$Y_{Free} - Y_{ID-State2}$	Only NP is different
Dynamic	$Y_{ID-State2} - Y_{ID-State1}$	Only SP is different
Retaliatory Behaviors	$Y_{ID-State3} - Y_{Free}$	NP and SP are the same, but the decay parameter differs

NP, SP are short for nominal price and shadow price, respectively. Y is a statistic derived from the count set $\mathcal{S}$. We are particularly interested in the mean, maximum and quantiles at 0.75, 0.5 and 0.25.

ID-State3 corresponds to a case where patients exhibit retaliatory behaviors that arise from nothing.

The selection of an appropriate time interval $\mathcal{T}$ is crucial to ensure reliable simulation results. A small time interval, such as one week, can severely limit the exhibition of properties of various counting processes, thereby rendering corresponding incentives ineffective. Conversely, a large time interval exceeding one year is practically unrealistic, as insurance plans usually reset at each contract year. As such, we set $\mathcal{T} = 52$ weeks.

Regarding simulation details, we use the intensity model (5) to simulate the Free, ID-State1, and ID-State3 counting processes. Meanwhile, we use model (6) to generate the ID-State2 counting process. Since estimates for individual heterogeneity parameters are statistically indifferent across plan-states (see Table 5), we will use estimates from the free plan (column 1 of Table 5). We also use individual heterogeneities (age, sex, and socioeconomic status) from the free plan to construct the simulation sample. Therefore, the sample size is $n = 723$. Dynamic parameters for different plan-states are summarized in Table 8. Lastly, we resample each outpatient cost with replacement from the free plan to form expenditures in ID-State1. Table 9 summarizes descriptive statistics for expenditure distribution in the free plan.

For each iteration k of the simulation, we obtain the counting set $\mathcal{S}_{plan-state}$ and use it to construct the statistics $Y^{(k)}_{plan-state}$. We perform $\kappa = 1000$ iterations in total. The resulting sample means of these statistics, $\bar{Y}_{plan-state} = \kappa^{-1} \sum_{k=1}^{\kappa} Y^{(k)}_{plan-state}$, are reported in Table 10. Results in this table are then used to quantify various incentives using strategies described in Table 7, which are eventually summarized in Table 11.

Table 8: Dynamic Parameters used for Simulations, Approach A

	Free (1)	ID-State1 (2)	ID-State2 (3)	ID-State3 (4)
λ	-2.539	-2.539	-2.539	-2.539
a	1.235	1.235	1.235	1.235
μ	1.153	2.853	2.853	0.336
b		1.353		

Table 9: Descriptive Statistics for Expenditure Distribution in the Free Plan

	Original Sample	Bootstrap Sample
Mean	39.171	39.311 (3.976)
Maximum	2975.789	1483.69 (655.048)
Quantile 75	36.842	36.755 (1.812)
Quantile 50	19.579	19.635 (1.149)
Quantile 25	12.632	12.717 (0.308)

We conduct $b = 1000$ iterations of bootstrap, within each iteration, we resample with replacement from the free plan expenditures. The resample size is $n = 600$. Numbers in the parentheses are empirical standard deviations.

Table 10: Simulation Results for Different Plan-States, Approach A

	Free (1)	ID-State1 (2)	ID-State2 (3)	ID-State3 (4)
Simulation Mean	6.663 (0.13)	1.376 (0.085)	5.597 (0.087)	9.933 (0.189)
Simulation Maximum	87.6 (8.335)	32.761 (5.568)	42.56 (4.245)	122.675 (5.49)
Simulation Quantile 75	7.85 (0.346)	1.004 (0.063)	7.611 (0.475)	9.226 (0.448)
Simulation Quantile 50	3.935 (0.247)	0.0 (0.0)	3.948 (0.222)	3.978 (0.147)
Simulation Quantile 25	1.144 (0.34)	0.0 (0.0)	1.166 (0.363)	1.042 (0.191)

This table reports sample means of various statistics, denoted by $\bar{Y}_{plan-state} = \kappa^{-1}\sum_{k=1}^{\kappa} Y^{(k)}_{plan-state}$. Numbers in the parentheses are empirical standard deviations calculated as $std\left(\{Y^{(k)}_{plan-state}\}_{k\in[\kappa]}\right)$.

Table 11: Quantification Results for Incentives, Approach A

	Mean (1)	Maximum (2)	Quantile 75 (3)	Quantile 50 (4)	Quantile 25 (5)
Static+Dynamic Incentives	5.268 (0.155)	55.174 (9.655)	6.819 (0.38)	3.931 (0.253)	1.128 (0.327)
Static Incentives	1.047 (0.156)	45.04 (9.091)	0.212 (0.605)	-0.017 (0.337)	-0.038 (0.489)
Dynamic Incentives	4.221 (0.121)	10.134 (7.002)	6.607 (0.479)	3.948 (0.222)	1.166 (0.363)
Retaliatory Behaviors	3.289 (0.229)	35.075 (9.735)	1.403 (0.584)	0.047 (0.293)	-0.086 (0.379)

This table reports quantifications of various incentives. The quantities are obtained by following strategies documented in Table 7. Numbers in the parentheses are empirical standard deviations calculated as $\left(std^2_{plan-state} + std^2_{plan-state'}\right)^{1/2}$, where $std_{plan-state}$ is the empirical standard deviation described in Table 10.

On average, our analysis shows that both static and dynamic incentives significantly impact individuals' healthcare behavior, as indicated by their deviation from zero. However, they differ in ways that affect the temporal dependence structure and scales. Specifically, on average, dynamic incentives have roughly four times greater impact compared to static incentives, suggesting that individuals are more responsive to changes in shadow prices. Moreover, our findings indicate that the effects of incentives are not uniform across individuals.

Static incentives impact the cluster structure, with the cluster size smaller in the ID plan compared to the free plan, as the decay rate in this plan is much higher than that of the free plan. However, if individuals have fewer healthcare needs intrinsically (e.g., healthier individuals), impacts from the static incentives would be greatly reduced. This heterogeneity in responses is evidenced by the results presented in row 2 of Table 11.

Regarding dynamic incentives (row 3 of Table 11), the opposite observation holds, with most individuals responding to shadow prices while heavy users often ignore them. This finding may be explained by the fact that individuals who frequently utilize healthcare services tend to have poorer health conditions, leading them to exhaust the deductible in the contract year with a high degree of certainty. As a result, the perceived likelihood of exceeding the threshold reduces the shadow price to near zero effectively. In terms of dependence structure, dynamic incentives would not affect the decay rate, but they would reduce the overall excitement strength.

On average, individuals exhibit retaliatory behaviors. However, these effects are not homogeneous across all individuals. In our simulation sample, approximately half of the patients would not time or postpone their healthcare needs. The other half, particularly heavy users, would have retaliatory spending patterns.

6.2 Quantification based on Parametric Specifications

The second approach (henceforth "Approach B") that we used to quantify various incentives is based on distinct parametric specifications of counting processes. It differs from Approach A mainly in how it accounts for dynamic incentives. Considering three artificial coutning processes: (1) individuals only responding to dynamic incentives; (2) individuals only responding to static incentives; and (3) individuals exhibiting retaliatory behaviors when exceeding the threshold. By referencing the free plan, we find that the second and third counting processes correspond to the

specifications described in columns (3) and (4) of Table 8. Similarly, we could parameterize the first counting process by incorporating the shadow price effect into the free plan. Table 12 provides a summary of these specifications, while Table 13 presents the quantification strategies.

Table 12: Dynamic Parameters used for Simulations, Approach B

	Free (1)	Dynamic Incentive (2)	Static Incentive (3)	Retaliatory Behavior (4)
λ	-2.539	-2.539	-2.539	-2.539
a	1.235	1.235	1.235	1.235
μ	1.153	1.153	2.853	0.336
b		1.353		

Table 13: Quantification Strategies for Various Incentives, Approach B

Incentives	Strategies	Comments
Static	$Y_{Free} - Y_{Static}$	Only μ different
Dynamic	$Y_{Free} - Y_{Dynamic}$	Add shadow price effects
Retaliatory Behaviors	$Y_{Retaliation} - Y_{Free}$	Only μ different

The subscripts "Dynamic", "Static" and "Retaliation" refers to the specifications presented in columns (2), (3) and (4) of Table 12. Y is a statistic derived from the count set $\mathcal{S}$. We are particularly interested in the mean, maximum and quantiles at 0.75, 0.5 and 0.25.

The simulation details are identical to those described in Approach A. Tables 14 and 15 show the results. The simulated quantifications for various incentives are statistically equivalent to those of Approach A, and the conclusions presented earlier remain unchanged.

7. Counterfactual Analysis

In this section, we quantify dynamic incentives under two sets of counterfactual analyses: (1) High-Deductible and (2) Copayment with Deductible Plans. These counterfactual plans are based on modifications to the shadow price, allowing for a measurement of changes in dynamic incentives only. The quantification approach uti-

Table 14: Simulation Results for Different Plan-States, Approach B

	Free (1)	Dynamic Incentive (2)	Static Incentive (3)	Retaliatory Behavior (4)
Simulation Mean	6.644	1.669	5.597	9.933
	(0.13)	(0.134)	(0.087)	(0.189)
Simulation Maximum	87.6	65.187	42.56	122.675
	(8.039)	(12.532)	(4.245)	(5.49)
Simulation Quantile 75	7.823	1.003	7.611	9.226
	(0.375)	(0.055)	(0.475)	(0.448)
Simulation Quantile 50	3.931	0.0	3.948	3.978
	(0.253)	(0.0)	(0.222)	(0.147)
Simulation Quantile 25	1.128	0.0	1.166	1.042
	(0.327)	(0.0)	(0.363)	(0.191)

This table reports sample means of various statistics, denoted by $\bar{Y}_{specification} = \kappa^{-1}\sum_{k=1}^{\kappa} Y^{(k)}_{specification}$. Numbers in the parentheses are empirical standard deviations calculated as $std\left(\{Y^{(k)}_{specification}\}_{k\in[\kappa]}\right)$. The results in columns (1), (3) and (4) are identical to those of Table 10 as their specifications are also the same.

Table 15: Quantification Results for Incentives, Approach B

	Mean (1)	Maximum (2)	Quantile 75 (3)	Quantile 50 (4)	Quantile 25 (5)
Static Incentives	1.047	45.04	0.212	-0.017	-0.038
	(0.156)	(9.091)	(0.605)	(0.337)	(0.489)
Dynamic Incentives	4.975	22.413	6.82	3.931	1.128
	(0.187)	(14.889)	(0.379)	(0.253)	(0.327)
Retaliatory Behaviors	3.289	35.075	1.403	0.047	-0.086
	(0.229)	(9.735)	(0.584)	(0.293)	(0.379)

This table reports quantifications of various incentives. The quantities are obtained by following strategies documented in Table 7. Numbers in the parentheses are empirical standard deviations calculated as $\left(std^2_{specification} + std^2_{specification'}\right)^{1/2}$, where $std_{specification}$ is the empirical standard deviation described in Table 10.

lized in this section is grounded in Approach B, as it provides a more straightforward and natural method of measuring dynamic incentives.

7.1 High-Deductible

A High-Deductible Health Plan (HDHP) is a type of health insurance plan that features a higher deductible but lower premiums than a traditional insurance plan. The adoption of HDHPs has been growing, with 28% of firms in the U.S. offering HDHP options as of 2022 (Health Benefits Survey, 2022 Edition, Kaiser Family Foundation). We examine the impacts of adopting HDHPs by modifying the deductibles in our model. One crucial assumption is needed: individuals would form shadow prices in the same way regardless of deductible settings. However, this assumption is rather strong. To illustrate, considering an extreme case where the deductible is set at 900 USD for outpatient events. Using our estimates and comparing to the free plan, the intensity would be discounted at $\exp(-1.353 \cdot 9) = 5.148 \cdot 10^{-6}$ at the beginning of each contract year. This discount is not realistic, as we would expect higher intensity values even if there is no insurance coverage. To obtain reasonable counterfactual results, we restrict our attention to deductibles set at $150, 180$ and 210 USD, where the deductible at 150 USD is the one used in the Rand ID Plan.

Simulation details are similar to those in the previous section. We use parameters described in the first and second columns of Table 12 to simulate the free plan and HDHPs, respectively. Tables 16 and 17 present the results.

As the deductible increases, our results reveal reductions in dynamic incentives. However, we find that these reductions are not linear. More specifically, we observe significant reductions in outpatient counts from 150 USD to 180 USD deductibles. In contrast, reductions from 180 USD to 210 USD deductibles are less apparent. These findings suggest that the power of dynamic incentives is marginally decreasing.

7.2 Copayment with Deductible

In contrast to coinsurance rates, which obligate patients to pay a fixed percentage of healthcare costs, copayments entail a fixed payment amount from patients, irrespective of the actual expenses incurred for healthcare services. When comparing coinsurance plans to copayment plans, patients with coinsurance plans must navigate two sources of random variation: the occurrence and cost of each outpatient service. In contrast, patients with copayment plans only experience randomness related to

Table 16: Simulation Results for Different Deductible Settings

	150 USD (1)	180 USD (2)	210 USD (3)
Simulation Mean	1.669	0.967	0.551
	(0.134)	(0.106)	(0.08)
Simulation Maximum	65.187	53.959	41.653
	(12.532)	(14.267)	(16.287)
Simulation Quantile 75	1.003	1.0	0.278
	(0.055)	(0.0)	(0.438)
Simulation Quantile 50	0.0	0.0	0.0
	(0.0)	(0.0)	(0.0)
Simulation Quantile 25	0.0	0.0	0.0
	(0.0)	(0.0)	(0.0)

This table reports simulation results for artificial counting processes under different deductibles (150, 180 and 210 USD) restrictions. The results in the first column are identical to those in Table 14. Numbers in the parentheses are empirical standard deviations.

Table 17: Dynamic Incentives (Counts) for Different Deductible Settings

	150 USD (1)	180 USD (2)	210 USD (3)
Count Mean	4.975	5.677	6.093
	(0.187)	(0.168)	(0.153)
Count Maximum	22.413	33.641	45.947
	(14.889)	(16.376)	(18.163)
Count Quantile 75	6.82	6.823	7.545
	(0.379)	(0.375)	(0.577)
Count Quantile 50	3.931	3.931	3.931
	(0.253)	(0.253)	(0.253)
Count Quantile 25	1.128	1.128	1.128
	(0.327)	(0.327)	(0.327)

This table reports quantifications of dynamic incentives (counts) under different deductibles (150, 180 and 210 USD) restrictions. The quantification strategy is described in Table 13. The results in the first column are identical to those in the second row in Table 15. Numbers in the parentheses are empirical standard deviations.

the occurrence of outpatients. Additionally, a low (high) copayment amount has two implications for patients. On one hand, the effects of static incentives should be reduced (increased) since the nominal price is low (high). On the other hand, a low (high) copayment also requires more (fewer) outpatients to be seen before reaching the deductible, increasing (decreasing) the effects of dynamic incentives.

In this subsection, we investigate the dynamic incentive effects of adopting copayments while fixing the deductible level at 150 USD. The simulation and quantification details are consistent with those in the previous section. Tables 18 and 19 present the results.

Our findings suggest that a reduction in copayment has a similar effect as an increase in deductible in terms of simulated results and dynamic incentives. For instance, copayments of 30 and 10 USD produce similar results to deductibles at 150 and 180 USD levels, respectively. Consequently, the quantifications of dynamic incentives among these copayment and deductible plans are statistically identical. However, copayments result in lower out-of-pocket costs for patients compared to deductibles.

To illustrate, consider a comparison between a copayment of 30 USD and a deductible at the 150 USD level. The average cost per outpatient visit is approximately 39 USD for the deductible plan (see Table 9). In the copayment plan, individuals only pay 30 USD, with the difference being covered by the insurance plan. However, insurers may increase premiums slightly to cover the loss, resulting in a net gain for patients.

Our results demonstrate that adopting copayments can be a viable alternative to high-deductible plans, especially for patients who require frequent outpatient services. Copayments maintain similarly dynamic incentive levels while reducing out-of-pocket fees. However, more work is needed to identify and quantify the static incentives induced by copayments, in order to better understand their impacts.

8. Conclusion

This study identifies and quantifies both static and dynamic incentives induced by the deductible. The static incentives are driven by individuals' responses to the nominal price, specifically the nominal coinsurance rate. In contrast, the dynamic incentives are a result of individuals' reactions to the shadow price, a stochastic

Table 18: Simulation Results for Different Copayment Settings

	10 USD (1)	20 USD (2)	30 USD (3)
Simulation Mean	0.881	1.301	1.744
	(0.058)	(0.102)	(0.124)
Simulation Maximum	32.722	57.889	66.204
	(15.196)	(13.449)	(11.773)
Simulation Quantile 75	1.0	1.0	1.002
	(0.0)	(0.0)	(0.045)
Simulation Quantile 50	0.0	0.0	0.0
	(0.0)	(0.0)	(0.0)
Simulation Quantile 25	0.0	0.0	0.0
	(0.0)	(0.0)	(0.0)

This table reports simulation results for artificial counting processes under different copayments (10, 20 and 30 USD). Numbers in the parentheses are empirical standard deviations.

Table 19: Dynamic Incentives (Counts) for Different Copayment Settings

	10 USD (1)	20 USD (2)	30 USD (3)
Count Mean	5.763	5.343	4.9
	(0.142)	(0.165)	(0.18)
Count Maximum	54.878	29.711	21.396
	(17.191)	(15.668)	(14.256)
Count Quantile 75	6.823	6.823	6.821
	(0.375)	(0.375)	(0.378)
Count Quantile 50	3.931	3.931	3.931
	(0.253)	(0.253)	(0.253)
Count Quantile 25	1.128	1.128	1.128
	(0.327)	(0.327)	(0.327)

This table reports cumulative costs under different copayments (10, 20 and 30 USD). The quantification strategy is described in Table 13. Numbers in the parentheses are empirical standard deviations.

process over time that is defined as the conditional probability of not exceeding the deductible at the end of the contract year given individual covariates and cumulative spending up to the current time.

Our analysis begins with a discussion on the selection of appropriate dependent variables. We believe that using individuals' healthcare spending, as commonly done in prior literature, can result in biased estimates due to physicians' moral hazards. Specifically, a relatively healthier patient may incur higher healthcare expenditures than a sicker patient, leading to a non-monotonic mapping from health states to healthcare spending. Consequently, comparisons between different health insurance plans based on such dependent variables are invalid.

Our preferred variable of interest for this analysis is outpatient counts, for two primary reasons. Firstly, healthcare counts are monotonically correlated with health states. In other words, on average, a healthier patient will have fewer healthcare events compared to a sicker patient. Secondly, and more importantly, we found that this monotonic correlation is not impacted by variations in supply-side incentives.

In order to represent count data, we employed the self-exciting process, which offers several advantages. Firstly, it is a special counting process that does not aggregate data, thus avoiding the loss of information, particularly dynamic information regarding the temporal dependence structure among outpatient events. Secondly, the self-exciting process is characterized by the occurrence of an event that influences the occurrence of the same event in the near future. In other words, the event itself excites more of the same event to happen subsequently. By modelling outpatient events as a self-exciting process, we are better able to determine the impact of cost-sharing policies on both static and dynamic incentives.

Our analysis relies on the Rand HIE, which is a randomized field experiment involving various insurance plans conducted from the early 1970s to the 1980s. Due to the randomness of the assignments and the nonlinear cost-sharing features, the Rand HIE data is particularly suitable for studying both static and dynamic incentives. Among the available insurance plans, we focus on the free plan and the individual deductible plan. The former places no restrictions on either patients or physicians, while the latter imposes a deductible and coinsurance rate of 95% on patients.

We conceptualize different states based on the relative position of cumulative healthcare spending to the deductible: prior-deductible and post-deductible. For the free plan, both nominal and shadow prices are zero in both states. However, for the

individual deductible plan, nominal and shadow prices differ in these states. In the prior-deductible state, both nominal and shadow prices are non-zero. However, when cumulative healthcare spending equals the deductible, the nominal price is non-zero but the shadow price is zero. Finally, in the post-deductible state, both nominal and shadow prices are zero. The variations in nominal and shadow prices allow for our identification and quantification strategy. We therefore built self-exciting process models based on these plan-states.

Regarding the empirical results, we have identified the presence of both static and dynamic incentives in individuals' healthcare behaviors, i.e., individuals would respond to both nominal and shadow prices. Furthermore, our quantification studies have revealed that static and dynamic incentives differ in significant ways that impact the temporal dependence structure and scales. Our results indicate that, on average, dynamic incentives have roughly four times greater impact compared to static incentives. Additionally, we found that static incentives would shrink the cluster size, while dynamic incentives would reduce the overall excitement strength in terms of dependence structure. Furthermore, incentive effects are not uniform across different individuals. Static incentives have a greater impact on "heavy users," while we found no significant dynamic incentive effects in this group. In contrast, dynamic incentives affect "light users" more, and individuals in this group do not respond to static incentives.

Finally, we conduct counterfactual analyses on two sets of scenarios: high-deductibles and copayments with deductibles. While a high-deductible plan would reduce dynamic incentives, we found that a copayment cost-sharing policy could achieve the same goal while keeping individuals' out-of-pocket fees low.

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.

Arrow, K. J. (1963): "Uncertainty and the welfare economics of medical care," *American Economic Review*, 941–973.

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.

Currie, J., W. Lin, and W. Zhang (2011): "Patient knowledge and antibiotic abuse: Evidence from an audit study in China," *Journal of health economics*, 30, 933–949.

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.

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 N. Mahoney (2018): "Provider incentives and healthcare costs: Evidence from long-term care hospitals," *Econometrica*, 86, 2161–2219.

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.

Eliason, P. J., P. L. Grieco, R. C. McDevitt, and J. W. Roberts (2018): "Strategic patient discharge: The case of long-term care hospitals," *American Economic Review*, 108, 3232–65.

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.

Gottschalk, F., W. Mimra, and C. Waibel (2020): "Health services as credence goods: A field experiment," *The Economic Journal*, 130, 1346–1383.

Gruber, J., J. Kim, and D. Mayzlin (1999): "Physician fees and procedure intensity: the case of cesarean delivery," *Journal of health economics*, 18, 473–490.

Gruber, J. and M. Owings (1996): "Physician financial incentives and cesarean section delivery," *The RAND Journal of Economics*, 27, 99–123.

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.

Jacobson, M., C. C. Earle, M. Price, and J. P. Newhouse (2010): "How Medicare's payment cuts for cancer chemotherapy drugs changed patterns of treatment," *Health Affairs*, 29, 1391–1399.

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.

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*.

Nguyen, N. X. and F. W. Derrick (1997): "Physician behavioral response to a Medicare price reduction." *Health services research*, 32, 283.

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.

Rice, T. H. (1983): "The impact of changing Medicare reimbursement rates on physician-induced demand," *Medical care*, 21, 803–815.

Rossiter, L. F. and G. R. Wilensky (1984): "Identification of physician-induced demand," *Journal of Human resources*, 231–244.

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 price 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. Simulation Algorithm and Simulation Studies

C.1 Simulation Algorithm

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 $a(t)$ in the interval as

$$a_{max} = \max_{t\in(\tau,\tau+\delta)} a(t)$$

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

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

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

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

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

C.2 Simulation Results

To examine the performance of the minimum distance estimators, we conduct simulation studies. The data generating process for these 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.

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

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 20: 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 21: 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 22: 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%