HEDG

HEALTH, ECONOMETRICS AND DATA GROUP

THE UNIVERSITY of York

WP 22/23

# Timing Moral Hazard under Deductibles in Health Insurance

Véra Zabrodina

August 2022

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

# Timing Moral Hazard under Deductibles in Health Insurance*

Véra Zabrodina
*University of Basel*

**Abstract**

This paper develops a new approach to identifying timing moral hazard in health insurance contracts when deductible choice is endogenous. I set up a dynamic model of healthcare consumption where individuals exceed a high deductible after a large health shock. I show that individuals either strategically prepone care from the year after the shock and keep a high deductible, or do not retime and switch to a low deductible the year after. The identification of timing moral hazard exploits the randomness of shock timing within a calendar year. Empirical results show quantitatively large timing moral hazard responses, which decrease with the time left to the deductible reset. The insured do re-optimize on-the-go to minimize out-of-pocket costs, but face substantial frictions in retiming, which differ across types of care. These patterns bear implications for cost sharing and insurance policy.

*Keywords:* Health insurance, strategic timing, moral hazard, insurance plan choice.
*JEL codes:* D82, I11, I13.

*This version: August 8, 2022.*

* Correspondence to: vera.zabrodina@unibas.ch. University of Basel, Faculty of Business and Economics, Peter-Merian Weg 6, 4002 Basel, Switzerland. I am grateful to Janet Currie, Michael Gerfin, Helge Liebert, Nicola Pavoni, Ulrike Unterhofer, Andrea Weber, Conny Wunsch and Nicolas Ziebarth, as well as seminar participants at the Swiss Health Economics Association, Young Swiss Economists Meeting, Ski and Labour Seminar, EALE SOLE 2020, ESWC 2020, ESPE 2021, IRDES, Graduate Institute Geneva, and the University of Basel for helpful discussions and comments. I thank Christian Schmid and CSS Insurance for access to and help with the data. All errors are my own.

# 1 Introduction

The strategic timing of healthcare consumption can pose an important threat to health insurance markets (Cabral, 2017). The insured can use private information about their realized health needs and time their consumption[1] across coverage periods so as to reduce out-of-pocket costs or purchase additional coverage. Specifically, this so-called *timing moral hazard* arises because the insurer cannot perfectly observe (or contract on) the date the health needs realize (particularly non-emergent medical procedures, e.g. a planned hip replacement), but only reimburse medical procedures based on the actual date of treatment. Understanding the incentives driving this behavior, its extent, and its interactions with other sources of asymmetric information is crucial to design health insurance contracts that sustainably balance risk protection and incentive costs.

This paper provides new evidence on timing moral hazard, and how it relates to classical forms of moral hazard and adverse selection in health insurance.[2] Relative to these responses, timing moral hazard bears conceptually important specifics, with distinct implications for insurance markets. First, it leaves total spending unchanged, but shifts consumption in time and out-of-pocket costs onto the insurer. The strategic timing of consumption can drive up costs in the insurance pool and increase premiums. It requires forward-looking individuals to be able to anticipate and plan medical procedures, and time them across coverage periods with different relative prices. Meanwile, classical moral hazard generates consumption that would not occur if the individual faced a higher marginal prices within the current coverage period (e.g. further diagnostic imaging tests). Here, it is additional consumption that drives up costs in the pool. With few exceptions, the literature has so far focused on this type of moral hazard, which I term *within-year moral hazard* to emphasize that it does not directly alter spending in other periods. Second, timing moral hazard differs from standard, *ex ante* adverse selection as it is driven by *ex post* timing decisions. It can occur even among individuals comparable in risk and preferences, whose different risk realizations create diverging dynamic incentives. Furthermore, the decision to retime care is tied to the coverage purchase decision, and can generate time-varying adverse selection even holding the available coverage choices fixed.

These specifics and interactions imply that policies targeting other sources of information asymmetry might be inadequate to address timing moral hazard. A large literature suggests that increasing the marginal out-of-pocket price of healthcare for the insured by offering contracts with consumer cost sharing can limit moral hazard on the demand

[1] In the context of healthcare, consumption (or demand) is measured one for one by total spending (Kowalski, 2015), so I use these terms interchangeably.

[2] I follow the conventional (ab)use of terminology in the context of health insurance and consider both behaviors to be types of *ex post* moral hazard, as I exclude all feedback effect on health risk. As the insured's actions (consumption) are observable, the information asymmetry stems from the insured having private information about their own price sensitivity and timing of procedures. See Einav et al. (2013) for a discussion of this terminology.

side. In particular, health insurance contracts with deductibles have become popular in the debate on how to contain the growth of healthcare spending. Under a deductible, the insured cover the first part of their medical costs out of pocket and contributions reset at the end of the coverage period—typically a year. The marginal price is then a nonlinear function of cumulated healthcare spending and time, and decisions today depend on expectations about future spending.[3] Such contracts are used in, e.g., mandatory health insurance in Switzerland (the setting for this study) and the Netherlands, as well as both private and public health insurance markets in the United States. There, among covered employees, 58% have a deductible higher than USD 1000 for single coverage (Kaiser Family Foundation, 2021). Contracts with deductibles are thus a widely-relevant case study.

However, they generate salient strategic timing incentives, which change dynamically within and across years. The insured have potential incentives to prepone care once their deductible is exceeded, or to postpone in anticipation of exceeding in the next year. Additionally, the insured typically have the possibility to choose between different deductibles every year, which can amplify timing moral hazard. A large scope for synchronizing healthcare consumption with insurance coverage can lead to strong adverse selection, and even failure of insurance markets (Cabral, 2017; Diamond et al., 2021). These links have however received little attention, as they complicate the identification of timing moral hazard.

In this paper, I develop a new approach to identifying timing moral hazard. I begin by formulating a model of healthcare consumption where a rational, forward-looking individual chooses their monthly healthcare consumption and yearly deductible, with the possibility to retime planned care. I focus on individuals with a high-deductible plan who are pushed into free care by a large, unanticipated health shock. This innovation allows me to circumvent selection issues and focus on the choice between preponing care from the year after the shock or switching to a low deductible. These choices are tied, and depend on the individual's baseline risk and propensity for within-year moral hazard. Specifically, individuals with low price sensitivity keep a high deductible and prepone planned care from next year to the current year to reduce out-of-pocket spending. More price-sensitive individuals rather switch to a low deductible without retiming, so that they hit the deductible again and benefit from within-year moral hazard.

Within this framework, I show that the differences in healthcare spending across individuals with shocks at different times within the calendar constitute sufficient statistics for timing moral hazard. This result motivates a novel identification strategy using the random variation in shock timing, which varies the distance to the year-end deductible

[3]Nonlinear price schedules have been studied in many markets. See e.g. Ito (2014) for electricity consumption, Grubb and Osborne (2015) for cell phones, and Nevo et al. (2016) for broadband, and Saez (2010) for labour supply responses to income tax incentives.

reset. Together with the nonlinear price schedule, shock timing exogenously varies dynamic price incentives after the shock. Individuals who suffer a shock later in the calendar year face a zero price for a shorter period after the shock than individuals with a shock earlier in the year. However, the later the shock, the more likely it spills over into the year after. Later shocks thus weaken the incentive to prepone care towards the shock year, and strengthen the incentive to switch to a low deductible and engage in within-year moral hazard the next year. In a frictionless case however, the total amount preponed would not vary across groups conditional on preponing.

For the empirical implementation, I use recent individual-level basic health insurance (BHI) claims data from the largest health insurance firm in Switzerland for the years 2005 to 2016. The Swiss BHI market offers an attractive setting for this analysis, since BHI is mandatory and highly regulated, and covers a broad scope of medical procedures. Contracts are standardized and bear a yearly deductible, which the insured can freely choose without risk classification. I run an event study of healthcare spending, where treatment effects are allowed to vary over time and across treatment groups defined by the calendar month of the first spending shock of more than CHF 2,500.

The results yield several insights. First, they provide suggestive evidence for variation in dynamic incentives: Shocks are persistent and spill over more into the next year, the later they occur in the calendar year. They also support the validity of the identifying assumption of random shock timing. Second, they allow me to compute the differences in healthcare spending needed to identify timing moral hazard. The estimates of timing moral hazard are quantitatively important. Early shock groups prepone over CHF 2,500 (CHF 1 $\approx$ \$1) more than later shock groups—a lower bound on the total amount preponed. Average rates of switching to a low deductible are of 14%, without significant differences across groups. These patterns suggest that there are substantial frictions to preponing, and that dynamic changes in incentives matter in shaping strategic timing responses. These frictions may stem from transaction costs (e.g. scheduling of appointments), institutional constraints (e.g. need for referrals), the nature of treatments (emergent or elective), as well as behavioral and cognitive biases (e.g. expectations about future health needs, present bias). These results point to coverage length and differentiating co-payment schedules across types of care as relevant policy tools to address timing moral hazard.

**Related literature.** This paper adds to several strands of the broad literature on moral hazard in health insurance. Following the seminal work by Arrow (1963) and Pauly (1968) and the RAND Health Insurance Experiment (Newhouse and the Insurance Experiment Group, 1993), an extensive literature has exploited price nonlinearities in insurance contracts to measure the price-elasticity of healthcare demand as a sufficient statistic for (within-year) moral hazard.[4] A recent body of studies has documented ev-

[4]See Finkelstein (2014), Einav and Finkelstein (2018), and Gerfin (2019) for reviews and discussions

idence of dynamic responses to nonlinearities in cost-sharing, with individuals bunching their spending around convex kinks of the budget set and anticipating deductible resets. In Switzerland, Gerfin et al. (2015) find persistently higher expenditures after hitting the deductible, and a sharp drop in spending at the year-end deductible reset among individuals with high-deductible plans. Einav et al. (2015) show that individuals close to entering the coverage gap ('donut hole') in Medicare Part D reduce their expenditures towards the end of the year, and shift their consumption to the next year (where expenditures are covered again). They find no such responses among those who spend largely past the gap and have weaker incentives to shift. The authors emphasize that failing to account for timing responses may overestimate the within-year moral hazard response. Further reduced-form studies find suggestive evidence of strategic delay of healthcare in anticipation of higher coverage (Card et al., 2009; Simonsen et al., 2021).[5] Taken together, these findings suggest that individuals respond to future prices and time spending strategically.

Several recent papers study timing moral hazard specifically. Lin and Sacks (2019) develop a test for short-term intertemporal substitution, where they compare individuals who hit the deductible with individuals under free care plans in the last month of the coverage year using data from the RAND experiment. They find that individuals with high deductibles have higher spending in that period on average, suggesting that those individuals who hit the deductible 'stock up' on health capital. Most closely related to this paper is Cabral (2017), who studies the incentive for delaying realized health needs in order to purchase additional coverage in the context of dental care. Using a structural modelling approach, the author finds that about 40% of individuals postpone deferrable dental care when incentivized to do so under a maximum benefit structure, which explains the missing market for dental insurance.

My paper makes several contributions to this line of research. First, it proposes a novel strategy to quantify timing moral hazard in reduced-form while retaining a structural interpretation, in a literature that has mainly adopted fully structural modelling approaches (Einav et al., 2015; Cabral, 2017). My approach provides a clear characterization of the quantities estimated in reduced form, while imposing fewer conceptual and distributional assumptions. It does not require specifying all the primitives underlying healthcare consumption and timing decisions. Importantly, it accounts for within-year moral hazard and endogenous deductible choice, and their relation to strategic timing decisions. In that,

of the literature on moral hazard in health insurance. The estimate of -0.2 from Keeler and Rolph (1988) is considered as the benchmark. Among recent reduced-form analyses, Kowalski (2016) uses injuries to family members in family-level insurance plans as an instrument for individual prices to estimate price elasticities across quantiles of the annual expenditure distribution. Ellis et al. (2017) instrument individual prices with employer-year-plan-month average cost shares to estimate price elasticities by type of medical service on a monthly basis. These studies adopt a static perspective by assuming that individuals only consider current or within-year prices, and implicitly rule out strategic timing across years.

[5]Simonsen et al. (2021) find evidence of stockpiling of drugs in anticipation of a switch from a linear to a non-linear co-payment plan, and year-end coverage resets, even among individuals with little incentives to stockpile. They conclude that consumers over-react to short-run price changes, and misunderstand incentives of non-linear contracts over the whole coverage year.

my paper bridges into the literature that aims to understand the links between behavioral responses to insurance stemming from asymmetric information. Einav et al. (2013) introduce the mechanism of adverse selection on moral hazard, taking a static, annual perspective.[6] My model builds on these insights to set up a dynamic framework with the option to retime care. It highlights that within-year and timing moral hazard are linked as the timing of consumption affects the year-end price.

Second, my empirical analysis considers all healthcare spending under BHI of a representative sample of individuals with high deductibles in Switzerland. In that, it is more general than the mentioned studies, that consider a particular population (e.g. the elderly, or employees of a specific firm) or healthcare market (e.g. drugs or dental care, which are particularly amenable to stocking up or retiming). It thereby informs on the extent of timing distortions in mandatory health insurance, which typically covers a large set of services and where the magnitude of retiming has not been quantified. Third, it focuses on a setting with a deductible where incentives are to prepone rather than postpone care. The preponing responses found in my analysis confirm that individuals are forward-looking and respond to future price incentives. Preponing indeed requires sophisticated agents who can anticipate and advance future non-emergent or planned procedures. Postponing is less demanding as individuals can delay care as needs realize over time. Still, the substantial heterogeneity in the preponed amount point to frictions to preponing that relate to the remaining time horizon. This paper broadens our understanding of the role of dynamic, cross-year incentives in shaping spending decisions. The quantitative magnitude of the results support the view that timing moral hazard bears important implications for the interpretation of estimates of the price-elasticity of healthcare consumption.

This insight relates my study to the literature on dynamic price responses in healthcare consumption. Several studies explore the relative responsiveness of individual healthcare consumption to current, future and average prices. Aron-Dine et al. (2015) consider employees who are hired and enrolled in employer-provided health insurance at different times within the year, and hence face different expected future prices for a given initial current price. They estimate the effect of future prices on initial spending within the first three months of enrolling, and reject the null of no response. In Brot-Goldberg et al. (2017), individuals newly faced with a high-deductible plan reduce their spending under the deductible, including sicker individuals who may expect to exceed it. Using a difference-in-regression-discontinuities design, Klein et al. (2022) find that the drop in consumption at the reset increases in response to increases in the universal deductible in the Netherlands. Dalton et al. (2020) and Abaluck et al. (2018) adopt dynamic structural approaches to estimating price-elasticities. They find evidence against the benchmark

[6]This is a recurrent feature of health insurance choice models (e.g. Kowalski 2015), where rational and forward-looking individuals maximize their utility over annual healthcare spending and other goods on an annual (static) basis. Then, only the expected year-end price matters, and within and cross-year spending dynamics are irrelevant.

rational agent model, and in favour of models incorporating price salience and significant myopia. Overall, this evidence suggests that individuals respond more strongly to the current price, and fail to correctly account for the year-end price. My findings suggest that individuals are not fully myopic, and respond to relative price differences across years by advancing care from the year after the shock.

The paper proceeds as follows. The next section presents the theoretical framework for the analysis. Section 3 outlines relevant institutional features of the Swiss health insurance system, and describes the data. Section 4 elaborates on the empirical implementation. Section 5 presents the main results, while Section 6 discusses robustness checks and extensions. The final section concludes.

# 2 Theoretical framework

## 2.1 A model of healthcare consumption with retiming

I now present a model of individual healthcare consumption and deductible choice. The framework extends the healthcare consumption models used in, e.g. Einav et al. (2013), Abaluck et al. (2018) and Klein et al. (2022) to include planned care amenable to retiming. It allows understanding the strategic timing incentives generated by a large health shock under deductibles, and how such behavior relates to within-year moral hazard and deductible choice. It then provides a clear characterization of sufficient statistics that can be estimated in reduced form to identify timing moral hazard, while the need to specify a full structural model. The model is tailored to the Swiss mandatory health insurance setting (Section 3.1), but can be generalized to contracts with nonlinear prices.

**Setup and utility function.**— Take a rational, forward-looking individual who lives in two years, split into months $t = 1, \ldots, 24$.[7] Every month, the individual chooses how much care to consume given future prices and health needs. My innovation is in considering individuals who exceed a high deductible due to a health shock during year 1, so that their price drops unexpectedly from 1 to 0 for the rest of the year. After the shock, they choose their consumption, the deductible for the year after the shock, as well as the amount of care to prepone care from the next year (timing moral hazard). Importantly, individuals with shocks at different times in year 1 face different incentives to prepone care, due to variation in the time left until the deductible reset, and the shock spillovers into the next year.

The per-period utility trades off health and consumption. Specifically, the individ-

[7]Since this is a model of individual-level behavior, I omit the $i$ subscript for simplicity. Separating the decisions made by the patient from those made by the physician, and assessing the optimality of the consumption from a cost-efficiency perspective is beyond the scope of this paper.

ual maximizes spot healthcare consumption $c_t$, which is measured in monetary units of purchased healthcare. Utility is assumed quasilinear in money and additively-separable across periods:

$$\max_{c_t \geq 0} u_j(c_t; \lambda_t, \omega, m_t, R_{j,t}, s) = v_c(c_t; \lambda_t, \omega) - v_m(m_t; \mu_t, s) - C_j(c_t; m_t, R_{j,t}) \tag{1}$$

where $v_c(c_t; \lambda_t, \omega) = (c_t - \lambda_t) - \frac{1}{2\omega}(c_t - \lambda_t)^2$ is concave, following Einav et al. (2013). $\lambda_t$ measures exogenous, nondiscretionary health needs.[8] $\omega$ is the time-constant price sensitivity.[9] This primitive measures the preference for within-year moral hazard in that it determines how strongly individuals react to marginal price changes, but also drives selection on moral hazard in deductible choices. $m_t$ measures planned care consumption, which is subject to timing moral hazard. $v_m(m_t; \mu_t, s)$ captures the potential costs of retiming. I discuss the rationale behind these key elements further below.

**Cost function and deductible.**— The out-of-pocket cost function under a low or a high annual deductible, $j \in \{L, H\}$, is

$$C_j(c_t, m_t, R_{j,t}) \equiv \min\{c_t + m_t, R_{j,t}\} + n_j \tag{2}$$

which depends on total spending in that period, and the remaining deductible at the beginning of period $t$, denoted by $R_{j,t} \equiv \max\{0, R_{j,t-1} - c_{t-1} - m_{t-1}\}$, with $R_{j,t} = D_j$, the deductible, for $t \in \{1, 13\}$. Monthly premiums are $n_j$. This function bears a nonlinearity in the marginal year-end price of consumption at the deductible level.[10] For forward-looking individuals, only the marginal year-end price $p_t^e$ matters for spot consumption decisions (Keeler and Rolph, 1988; Abaluck et al., 2018; Klein et al., 2022) and is defined as follows

$$p_t^e = \frac{\partial C_j(c_t, m_t, R_{j,t})}{\partial c_t} = \begin{cases} 1 \text{ if } R_{j,T} > 0 \\ 0 \text{ if } R_{j,T} = 0 \end{cases} \quad \text{for } T \in \{12, 24\} \tag{3}$$

Individuals pay the full price of care below the deductible, and nothing above.

Furthermore, assume that the deductible in year 1 is given and high, i.e. $R_{H,1} = D_H$. Focusing on individuals with high deductibles in the first year circumvents selection effects

[8] This quadratic functional form is an approximation of any utility function in the difference between healthcare consumption and nondiscretionary health needs, and quasilinear in money. Income effects are assumed away, as is customary in the literature.

[9] Following the existing literature, health is modelled as a normal good, so that the price elasticity of demand is weakly negative with respect to prices. If several medical treatments are possible for the underlying illness at the same price, the individual chooses the one with the highest marginal return.

[10] This formulation follows existing literature in assuming an exogenous income, and no saving and borrowing, see Klein et al. (2022) for a discussion. It gives rise to a non-convex annual budget set in health and residual income (consumption of other goods), which introduces the possibility of multiple solutions, but excludes bunching at the kink as in nonlinear price schedules with maximum benefits (as in e.g. Abaluck et al. 2018; Cabral 2017; Einav et al. 2015).

at baseline.[11] The deductible resets in $t = 13$, i.e. the start of the second coverage year, and the individual can freely choose the deductible for year 2.

**Types of healthcare consumption.**— Notice that the model distinguishes between three types of healthcare consumption, which rationalize specific individual decisions, and have different implications for cost sharing. All components are known to the individual, while the insurer (and the researcher) only observes the individual's total healthcare consumption.

First, nondiscretionary needs $\lambda_t$ capture a minimal set of medical services that cannot be chosen nor retimed. Think of a patient who suddenly suffers a heart attack and requires an angioplasty, which bears a given price for emergency care and a series of follow-up treatments at a regulated price. The individual consumes at least $\lambda_t$ in any given period, even under a marginal price of 1. $\lambda_t$ is determined by the individual's risk type, with a higher level capturing sicker individuals and driving up healthcare consumption, and is thus the source of *ex ante* adverse selection in deductible choices.

Second, the parameter $\omega$ determines classical within-year moral hazard, i.e. discretionary care that the individual would not consume if they had to cover the cost themselves (e.g. an additional diagnostic imaging test, or more expensive treatments). This primitive is essentially the elasticity of healthcare consumption with respect to the out-of-pocket price (see Einav et al. 2013 for a discussion). Individuals with a higher $\omega$ increase their healthcare consumption more if they exceed their deductible. The amount of additional spending depends on the year-end price in the current year, but is not directly affected by price dynamics in other coverage years.

Third, the individual has a given amount of planned care every month, denoted by $\mu_t$, which is amenable to retiming. The amount actually consumed in period $t$ is endogenous and denoted by $m_t$. This amount and its variation across shock groups is the key object of interest. At every time $t$, the individual can consume no or at most all planned care in that period, such that $m_t \in [0, \sum_{t=1}^{24} \mu_t]$. The total planned spending over the two years is fixed, such that $\sum_{t=1}^{24} m_t = \sum_{t=1}^{24} \mu_t$. This modelisation postulates that there are medical procedures known in advance that can be shifted in time, on top of any nondiscretionary needs. Individuals can only shift care that they know they (will) need, depending on the relative year-end prices across years.[12] In other words, if the price is

[11] In particular, it excludes any incentive to prepone before (or in the absence of) the shock, as individuals initially expect to end year 1 below the deductible. In that sense, the first-year deductible is assumed to have been chosen optimally. The shock itself does not contradict that, but rather captures that the individual suffered an unfavourable realization of health needs. However, it remains possible that these individuals choose a low deductible in year 2 if total planned spending is high enough, even without the shock. The share of such individuals should be constant across shock groups.

[12] Another way to rationalize intertemporal substitution in healthcare is through health capital (in the spirit of Grossman 1972), whereby individuals invest into a durable 'stock' of health when prices are lower. Here, as in Cabral (2017), healthcare consumption does not translate into future benefits through greater health capital, i.e. it does not impact its marginal utility in other periods.

constant across years, there is no incentive to retime. Shifting planned spending does not affect its monetary value: For e.g. an elective hip replacement, the set and value of the medical care received are the same regardless of when it is performed and of the marginal price.[13] Hence, retiming as such does not affect total healthcare spending, but cost sharing between the insured and the insurer.

Planned care enters utility via the out-of-pocket costs it generates below the deductible. The utility gain from retiming stems from savings in out-of-pocket costs. However, retiming is potentially costly as it requires active action from the individual. The utility cost of retiming is captured by $v_m(m_t; \mu_t, s)$, which depends on the amount retimed, as well as shock timing. I model the cost of planned care as a function of the difference between $m_t$, the amount actually consumed, and $\mu_t$, the amount initially planned

$$v_m(m_t; \mu_t, s) = \mathbf{1}\{m_t > \mu_t\} \cdot \left\{ (m_t - \mu_t) - \frac{1}{2\rho(s)} (m_t - \mu_t)^2 \right\} \tag{4}$$

The cost is only paid once in the period to which care is retimed ($m_t > \mu_t$), but cost savings occur in the period where consumption was initially planned. It is concave in the amount consumed: The cost of retiming declines initially as individuals group procedures, but increases eventually as hassle costs dominate. Importantly, the function $\rho(s)$ creates a possibility for heterogeneity in the timing response among shock groups, but constant in $t$. In a frictionless case $\rho(s) = \rho$, comparable individuals would prepone the same amount regardless of shock timing. While the model is agnostic about specific sources of retiming costs and heterogeneity therein, I discuss possible micro-foundations and their implications in light of the results below.

Note that the additively separable structure of the utility function implies that spot and planned consumption only influence each other through out-of-pocket costs. This assumption enables identification. It is reasonable if one sees planned care as "mandatory" spending that is planned regardless of and thus independent from any current consumption, but that can be consumed at any time within the two years.[14]

**Health shock with random timing.**— The individual suffers an unanticipated, exogenous shock in period $s \in S = \{1, \ldots, 12\}$, which pushes their cumulated spending above the deductible and yields a marginal price of zero for the rest of the first year. I now use the superscript $s$ to emphasize where decisions might depend on shock timing.

The shock triggers a sequence of health needs $\lambda_k(s)$ in time relative to the shock $k \geq 0$ (assume for simplicity that $\lambda_k(s) = 0$ for $k < 0$). It creates an incentive to prepone care

[13] As long as discounting is negligible and planned care yields constant marginal returns, any direct utility therefrom is constant and drops out of the optimization problem.

[14] This setup could be generalized to a longer time horizon. Given that the focus is on characterizing preponing decisions towards the year of a health shock, the two-year model captures the key short-term incentives to do so. However, it reasonably abstracts from retiming from further in the future.

from year 2 to year 1 to benefit from reductions in future out-of-pocket costs, as the year-1 coverage choices and prices are fixed.[15] The timing of the shock creates differential incentives through two channels. First, it varies the time left until the year-end deductible reset: The later the shock, the less time left to prepone. Second, shock timing shifts shock-related health needs in calendar time, determining how strongly the shock affects year 2: The later the shock, the more likely are shock-related health needs to spill over.[16] Figure 1 illustrates the variation induced by different shock timing with exemplary shock months. Individuals with a shock in March (Panel a) have more time before the reset than those with a shock in June (Panel b), and have less spillovers into year 2 (black dashed line). The June group suffers larger spillovers into year 2 than the March group. As a result, the marginal price next year for the June group is larger than for March, as they are more likely to exceed the deductible, but it is endogenous to consumption and deductible choices.

The key identifying assumption is that shock timing is random. It is thus independent of any potential heterogeneity across individuals, e.g. in preferences (in particular $\omega$), underlying health status, or other features of insurance plan choice. This assumption also implies that expected nondiscretionary spending in relative time is on average the same for all $s, s' \in S$. I thus compare the behavior across individuals identical in their *ex ante* risk, who differ only in the timing of their risk realization. In the empirical implementation, this comparison is more credible than that of individuals with and without shock. I discuss the random timing assumption further below.

**Optimal decisions.**— In the wake of the shock, the individual makes the following choices by backwards induction, knowing all relevant elements:[17]

1. Choice of monthly spot consumption for the rest of year 1.
2. Choice of planned care consumption and deductible for year 2.
3. Choice of monthly spot consumption for year 2.

In year 1, since the deductible is exceeded due to the shock, all additional care is free with $p_t^e = 0$ for $t = s, \ldots, 12$. Any decisions regarding year 2 do not affect spot consumption decisions in year 1. Taking first-order conditions, optimal consumption

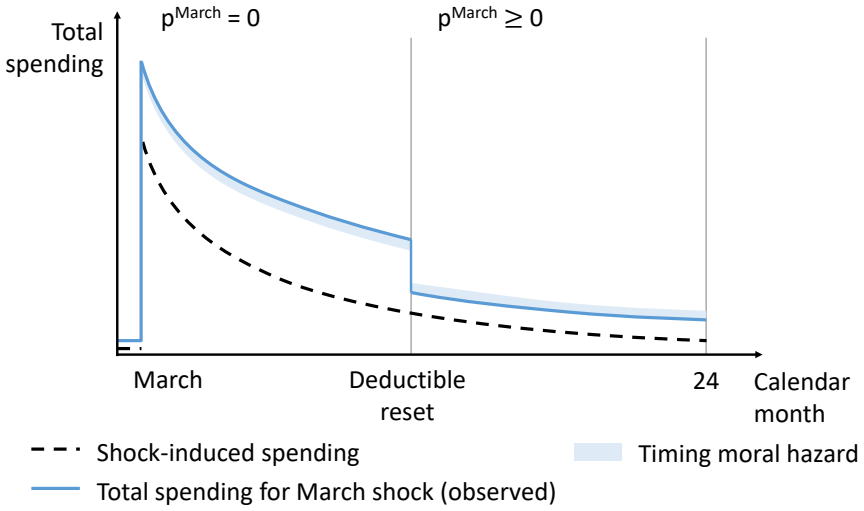
[15] Postponing would be incentivized by the possibility to purchase more coverage.

[16] The variation stems from the health needs of individuals in calendar time. This differs from previous studies that have focused on changes in the price schedule, e.g. an increase in the deductible (Brot-Goldberg et al., 2017; Klein et al., 2022). There, the current and future prices vary jointly, even if this variation is exogenous. Here, the current price (i.e. the marginal price of healthcare today) between the shock and the year-end reset is held constant, while future prices vary (i.e. the marginal price of healthcare tomorrow), similarly to Aron-Dine et al. (2015).

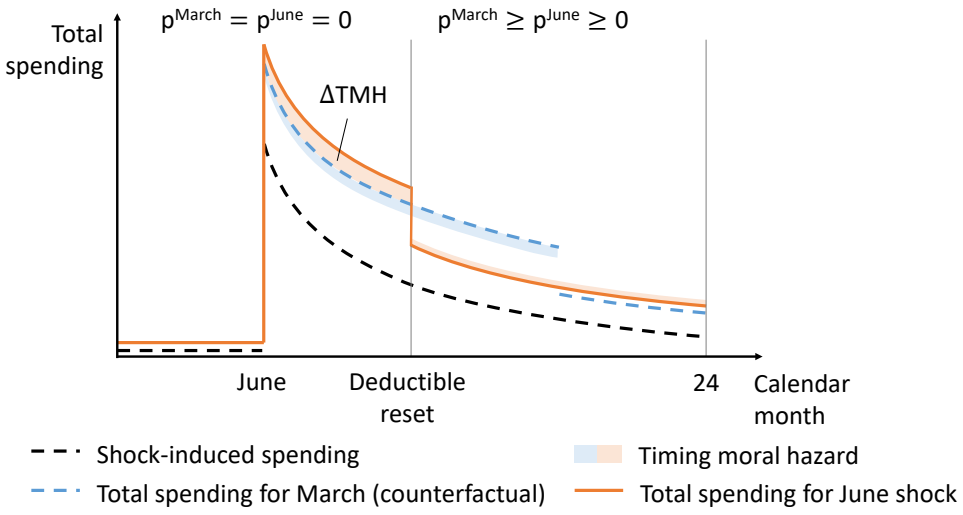
[17] I assume that all uncertainty with regards to health needs and prices is resolved after the shock, and ignore discounting for simplicity. Uncertainty would work against me finding evidence for timing moral hazard, since on average risk averse individuals are less willing to prepone (thus driving down differences in spending) to stay closer to the deductible and reduce the risk the next year.

Figure 1: Dynamics of healthcare spending following a health shock

(a) March shock




(b) March vs. June shock




*Notes:* The figure illustrates the intuition behind the identification of timing moral hazard based on shock timing and differences in spending dynamics in relative time. It depicts exemplary healthcare spending patterns of individuals with early (March) vs. mid-year (June) health shocks. Grey vertical lines mark year-end deductible resets (i.e. the end of the two calendar years). The black dashed line illustrates shock-induced nondiscretionary care, which cannot be chosen nor shifted in time. Health shocks push the individuals above the deductible, so that the marginal price of healthcare $p = 0$ for the rest of year 1. The solid blue and orange lines mark observed total healthcare spending. The dashed blue line illustrates the use of the mid-year's group spending as a counterfactual. The shaded areas illustrate timing moral hazard (TMH), i.e. care shifted from year 2 to year 1 following the shock realization.

equals nondiscretionary needs plus any within-year moral hazard spending

$$c_t^*(s) = \lambda_t(s) + \omega \quad \forall t = s, \ldots, 12 \tag{5}$$

Importantly, individuals have the same optimal spot consumption in time relative to the shock across $s$, and so until their year-end reset. The timing decision $m_t$ comes on top. Total observed healthcare spending is given by the sum of nondiscretionary, within-year moral hazard, and retimed care

$$h_t^*(s) = \lambda_t(s) + \omega + m_t^*(s) \quad \forall t = s, \ldots, 12 \tag{6}$$

where the single components are unobserved. After exceeding the deductible, any excess spending combines two responses: within-year and timing moral hazard. The goal is to identify $m_t^*(s)$, which depends on the decisions for year 2, while year-1 conditions are fixed. It already becomes apparent that any differences in observed total healthcare spending across shock timing groups are driven by differences in the amount of preponed care.

For year 2, I show in Appendix A that only two options are optimal. An individual is either a *preponer* who advances their optimal amount of planned care to year 1 and keeps a high deductible. Their path of planned spending consumption is

$$m_t^*(s) = \begin{cases} \mu_t & \text{for } t = 1, \ldots, s-1 \\ \mu_t + \rho(s) & \text{for } t = s, \ldots, 12 \\ 0 & \text{for } t = 13, \ldots, 24 \end{cases} \tag{7}$$

where any preponed care $\rho(s)$ is allowed to vary across shock groups. Otherwise, the individual is a *switcher* who keeps consumption as planned and switches to a low deductible, such that $m_t^*(s) = \mu_t$ for $t = 1, \ldots, 24$. All other options are dominated.

Let $V(D_j, m(s))$ be the utility value of deductible choice and planned consumption decisions after the shock. Preponers satisfy

$$V(D_H, m(s)) \geq V(D_L, \bar{\mu}(s)) \tag{8}$$

$$(\bar{n}_L - \bar{n}_H) + (D_L - \bar{\lambda}(s) - \rho(s)) - \bar{v}_m(\rho(s), s) - \frac{\bar{\omega}}{2} \geq 0 \quad \forall s \in S \tag{9}$$

The intuition behind this result is that individuals only prepone and keep a high deductible if the reduction in out-of-pocket costs is larger than the opportunity cost in terms of within-year moral hazard consumption in year 2, and the cost of retiming. A greater taste for within-year moral hazard increases the opportunity cost of preponing. So if expected needs in year 2 are high enough, individuals with higher $\omega$ switch to a low deductible (even if it comes with higher insurance premiums) and consume as planned to hit the deductible. This key result highlights the link between the two forms of moral

hazard. Interestingly, (selection on) within-year moral hazard attenuates the incentive to prepone care. Preponing is costly and might put the individual below the deductible and prevent them from drawing utility from within-year moral hazard under free care. From the insurer's viewpoint, preponing may be preferable if within-year moral hazard is reduced in the next year.

The condition in (8) holds regardless of shock timing $s$, but the underlying primitives might vary across individuals within a shock group. Heterogeneity in $\omega$ implies that preponing is optimal only for a given share of individuals in every shock group $s$ who satisfy condition (8). Denote the share of preponers as $q(s)$. Since the retiming and deductible decisions are tied, $q(s)$ is identified by the share of individuals choosing a high deductible in year 2, which is observed.[18]

Shock timing affects $q(s)$ through two main channels. First, $q(s)$ decreases with $\bar{\lambda}(s)$ (recall that $\bar{\lambda}(s) \geq \bar{\lambda}(s')$ for $s > s'$). This comparative static highlights the within and cross-year spending dynamics induced by shock persistence and their role in shaping dynamic incentives. Second, the amount retimed and induced costs potentially depend on shock timing. As a consequence of the variation in incentives, even individuals with comparable risk at the start of the coverage year may in principle diverge in their moral hazard responses, depending solely on when their shock realizes. In other words, they can become differentiated by the amount of care they have retimed as a result of shock timing— the *ex post* adverse selection pinned down by Cabral (2017).

## 2.2 Identifying timing moral hazard using shock timing

I now combine the results above to quantify timing moral hazard. Figure 1 illustrates that the difference between total spending and the counterfactual for earlier shock groups identifies differential planned care consumption (shaded areas) in relative time after the shock. Formally, let $\Delta\gamma_k(s, \Delta s) \equiv h_k(s + \Delta s) - h_k(s)$, for $s, s + \Delta s \in S$ and $\Delta s > 0$ in relative period $k = 1, \ldots, 13 - s - \Delta s$. Plugging in total healthcare spending (6) gives

$$\begin{aligned}\Delta\gamma_k(s, \Delta s) &= \lambda_k(s + \Delta s) - \lambda_k(s) + \omega - \omega + m(s + \Delta s)q(s + \Delta s) - m(s)q(s) \\ &= \rho(s + \Delta s)q(s + \Delta s) - \rho(s)q(s) \end{aligned} \tag{10}$$

where nondiscretionary needs and within-year moral hazard in year 1 are differenced out, given random shock timing and holding constant any seasonality in $\mu$.[19] The remaining

[18] Unexpectedly exceeding the deductible leads to substitutability between timing moral hazard and coverage purchase. This differs from the case where the incentive is to delay procedures in order to purchase additional coverage (e.g. if the marginal price in year 1 is higher than in year 2). In that case, timing moral hazard and purchasing coverage are complements.

[19] In Figure 1, within-year moral hazard is comprised in the unshaded area between the nondiscretionary spending path and the timing moral hazard, but cannot be identified separately from $\lambda$ without further assumptions. The total difference in spending in year 2 is the sum of timing moral hazard, within-year

difference solely captures differences in retimed care driven by shock timing, and is independent of all other parameters. This difference is a weighted average of preponers and switchers' decisions, where switcher have zero retimed care. In sum, timing moral hazard can be teased out without having to identify the other components of total spending, in constrast to existing structural approaches.

Using the finding that the rate of switching to a low deductible is not significantly different across shock groups, i.e. $q(s) \approx q(s + \Delta s)$ (see Section 5), gives

$$\Delta\tilde{\gamma}_k(s, \Delta s) \equiv \frac{\Delta\gamma_k(s, \Delta s)}{q(s)\Delta s} \tag{11}$$

$$= \frac{\rho(s + \Delta s) - \rho(s)}{\Delta s} \tag{12}$$

$$= \lim_{\Delta s \to 0} \rho'(s) \tag{13}$$

where we notice that the first term is a function of quantities that can be estimated for different comparison pairs $\{s, s + \Delta s\}$. These moments then provide point estimates for the derivative of the optimal timing moral hazard choice function for preponers at different values of $s$.[20] Finally, by making a functional form assumption for the optimal preponed amount $m_t^*(s)$, we can fit the estimated moments for the derivative to integrate it up and infer the total retimed amount across shock groups (see Section 4).

# 3 Setting and data

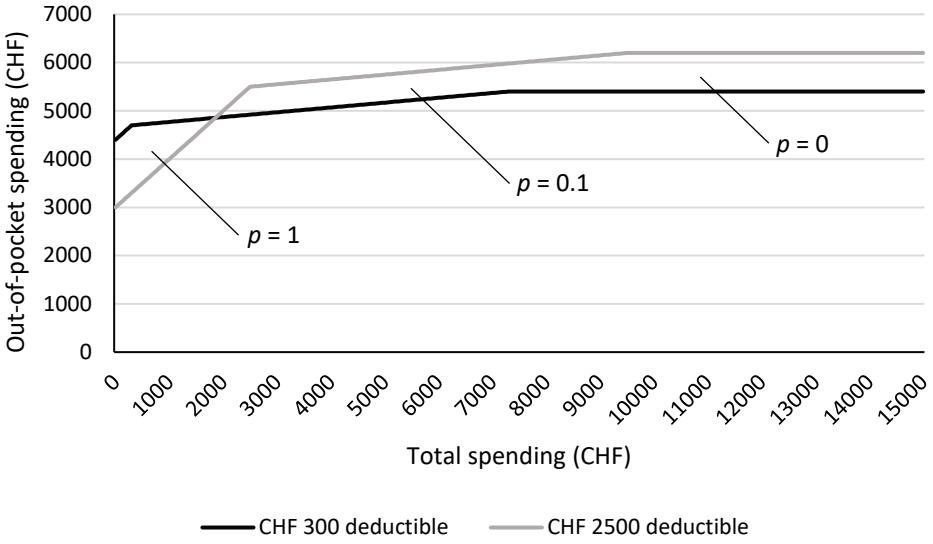
## 3.1 Health insurance in Switzerland

The Swiss market for basic health insurance (BHI) offers compelling features to analyse strategic timing behavior. Each resident is required by law to conclude an individual BHI contract with a private health insurance company (i.e. there are no family-level plans). The main features of BHI plans are highly standardized and regulated on a federal level. They comprise price nonlinearities in the spending and time dimensions that provide identifying variation in dynamic incentives. Specifically, the individual chooses an annual deductible of CHF 300 (the default), 500, 1,000, 1,500, 2,000 or 2,500. This choice set has remained unchanged since 2005 (i.e. the starting point for this analysis). Above the deductible, a co-payment rate of 10% applies up to a stop-loss of CHF 700. Any additional care is fully reimbursed. Total annual out-of-pocket spending under deductible $D_j$ can be

moral hazard, and deductible selection.

[20] In this model, non-null differences are a sufficient condition to establish the existence of timing moral hazard. However, null differences are not a sufficient condition to reject it.

Figure 2: Out-of-pocket healthcare spending as a function of total yearly healthcare spending




*Notes:* The figure presents out-of-pocket healthcare spending as a function of total healthcare spending within a calendar year for the CHF 300 and CHF 2,500 deductibles. Yearly insurance premiums for standard plans determine the intercepts at CHF 4,400 and 3,000, respectively, i.e. sample averages from 2015. Below the deductible, the insured face a marginal price $p = 1$, then $p = 0.1$ until the stop-loss of CHF 700 for the co-payment, and $p = 0$.

thus written as a function of total annual healthcare spending $H$ as follows:

$$OOP_j(H) = 12n_j + \min\{D_j, H\} + \max\{0, \min\{0.1(H - D_j), 700\}\} \tag{14}$$

where $n_j$ are monthly premiums which decrease with the deductible, and depend on the characteristics of the insurance plan. Figure 2 sketches this function for the lowest and highest deductibles. The marginal price drops from 1 to 0.9 when exceeding the deductible, and then to 0 after the stop-loss for the co-payment is reached. For simplicity, I ignore the co-payment in what follows. This price schedule generates non-convexities in the individual's budget constraint at the deductible and stop-loss in the total healthcare expenditure dimension. On average in my data, the maximum annual OOP spending net of premiums is CHF 1,000 for the lowest deductible and CHF 3,200 for the highest deductible.

As coverage is annual, out-of-pocket contributions to the deductible and co-payment reset at the end of each calendar year, generating a discontinuous increase in price for individuals ending the year above the deductible. The insured have until November 30th to switch plans or insurers, with the new contract taking effect on January 1st.[21] Individ-

[21]Some insurers allow notifying until December 31st if the insured wishes only to increase their de-

uals can freely purchase any insurance plan from insurers operating within their canton (region) of residence.[22] Apart from the deductible, individuals choose between standard or alternative plans (i.e. health maintenance organizations, gatekeeping family physician, or telemedicine). The standard plan offers the free choice of authorized healthcare providers. Alternative plans restrict the choice of healthcare providers, but come with lower premiums. The cost-sharing structure and set of covered services are homogeneous across all plan types.[23]

Insurers mainly compete on premiums, and are forbidden from denying any BHI contract, selecting or pricing individual premiums based on identified health risk. As highlighted by Cabral (2017), such regulations limit the possibility to underwrite potentially retimed claims. The mandatory nature of BHI implies nearly-universal coverage and eliminates selection into insurance at baseline, such that selection occurs at plan choice level only. This feature constrasts to settings with options to opt in and out, e.g., voluntary employer-sponsored health insurance.

BHI covers a comprehensive range of medical services generated by illnesses defined by federal law. This includes elective and emergent ambulatory and inpatient services, as well as all drugs prescribed by a physician, such that BHI claims data allows eliciting retiming responses for a broader range of spending types than previous studies. Finally, the prices of medical services covered by BHI are fixed: Ambulatory services are reimbursed through a fee-for-service system, while hospitalizations are reimbursed through prospective payment by diagnosis-related groups (DRGs). This cancels out the price-shopping channel, for which little evidence has been found in the literature (see e.g. Brot-Goldberg et al. 2017).

This contract structure creates incentives for individuals to retime care towards years where they expect to exceed the deductible, that touches upon a broad scope of healthcare services. In particular, an individual can prepone care from the following coverage year if they hit the deductible in a given year, or delay care if they expect to exceed the deductible the year after. In the latter case, individuals can (temporarily) purchase additional coverage by choosing a lower deductible and face a lower marginal price. In my sample, the average annual premiums for standard plans were about CHF 4,400 and 3,000, for the CHF 300 and CHF 2,500 deductibles, respectively. As seen in Figure 2, the CHF 300 deductible plan dominates the CHF 2,500 deductible plan in terms of out-

ductible.

[22] Cantons provide means-tested subsidies for low-income households to purchase BHI.

[23] Accidents are covered by accident insurance, which is concluded separately and therefore does not count towards BHI deductible and is not included in the dataset. Individuals have to purchase accident insurance themselves if they are employed less than 8 hours a week. Otherwise, the employer is mandated to purchase the insurance. Accident insurance can be subscribed independently of BHI. Supplementary or private health insurance can be purchased to benefit from a greater coverage of ambulatory services (e.g. alternative medicine) or choice of private hospitals. For these contracts, insurers can flexibly select individuals and underwrite pre-existing conditions. The data in this study do not include claims made under supplementary or private health insurance. Dental care is not covered by BHI.

of-pocket spending for healthcare spending above CHF 1,900, after accounting for the difference in premiums. Retimed care affects where the individual lies relative to this threshold. These incentives to retime care are formalized in the theoretical framework above.

### 3.2 Insurance claims data

The analysis uses an individual-level panel of BHI claims from a sample of insured from the largest health insurer in Switzerland for the years 2005 to 2016.[24] The raw sample is a random draw of 375,000 insured from the high-deductible population. An insured belongs to this population if they had a deductible of CHF 1,000 or more in one year at least over the observation period, and was enrolled with CSS for at least five years. The data contain information on the associated spending for each claim, with the start and end dates of the treatment spell, broad category of care (i.e. outpatient at an ambulatory clinic or physician, outpatient at a hospital, inpatient, prescription drugs).[25] They also include basic individual demographic information (gender, age, nationality, canton/region of residence), as well as full insurance plan information (premiums, deductible, type of plan, start and end dates of enrolment), and an indicator for whether they subscribe to accident insurance at CSS.

The data allow me to observe all BHI claims, also below the deductible, as well as individuals who do not have any claims. The deductible structure incentivizes the filing of all claims, as opposed to e.g. a maximum benefit. Claims are sent by the healthcare provider directly to the insurer, who then invoices the insured for any expenses below the deductible. This happens by default unless the insured specifically requests to pay the invoice to the healthcare provider and then claim reimbursement themselves.

With this, I construct full individual monthly spending paths by allocating the claimed amount proportionally across the months spanned by the claim.[26] I censor total monthly spending at CHF 20,000 (i.e. at approximately the $99.9^{\text{th}}$ percentile of the distribution in the baseline sample) to avoid extreme outliers. All spending is adjusted to prices in December 2016 (i.e. the last observation month) using the Swiss consumer price index from the Swiss Federal Statistical Office.

Following the model, the main analysis focuses on high-deductible individuals with a large spending shock, for whom a sharp change in price is most likely unanticipated

[24] CSS Insurance operates across all of Switzerland, with approximately 800,000 insured yearly and a stable market share of around 10% of BHI.

[25] Information on specific medical services is not available.

[26] See Appendix B.2 for further remarks on billing. The month as a time unit balances the trade-off between statistical power and the precision of elicited dynamics, while smoothing out within-month variation and billing effects.

and therefore timed randomly. Rational individuals who expect large expenditures would choose a low deductible, especially if they are risk averse. This also decreases the likelihood that these individuals select into a specific treatment group by shifting expected spending towards the year of the shock. High-deductible individuals represent around 37% of the yearly general population of insured.

# 4 Empirical implementation

This section describes the estimation of the relevant quantities to identify the derivative of timing moral hazard in (11). Using the assumption of random shock timing, I adopt a multiple treatment framework, where the calendar month of the first shock $S_i \in \{3, ..., 10\}$ defines mutually-exclusive treatment groups.[27,28]

The main definition of a shock is having monthly spending of more than CHF 2,500 for the first time, at least one year into the observation window. This definition ensures that all individuals exceed their deductible and uncertainty with respect to year-end prices is alleviated, as in the model setup. In Section 6, I discuss robustness checks with alternative definitions of the shock.

I setup an event study to evaluate how spending dynamics vary with shock timing and to estimate $\Delta\gamma_k(s, \Delta s)$ based on counterfactual spending paths. Under the identifying assumption of random shock timing and after taking out baseline spending, any differences in spending between the shock and the reset are driven by timing moral hazard responses. Let the event time $E_i$ denote the calendar period of the first shock, which together with $S_i$ characterizes the full treatment path. The main outcome is healthcare spending for individual $i$ in calendar month $t$, denoted by $h_{it}$. I estimate the following event study on the insured-month level:

$$h_{it} = \sum_{s=3}^{10} \mathbf{1}\{S_i = s\} \left( \sum_{k=-11}^{24} \gamma_k^s \mathbf{1}\{t - E_i = k\} + \gamma_{24+}^s \mathbf{1}\{t - E_i > 24\} \right) + \sigma_t + \nu_i + \zeta X_{it} + \varepsilon_{it} \tag{15}$$

This specification stems directly from the optimal consumption model laid out above, and allows constructing the empirical equivalent to timing the moral hazard responses. The coefficients of interest $\gamma_k^s$ flexibly capture the effect of treatment $s$ on spending in relative month $k \in \{-11, ..., 24\}$, with any longer term level effects captured by $\gamma_{24+}^s$.

[27] I exclude individuals who have a shock in January and February, as well as November and December in the empirical analysis to avoid turn-of-the-year spending and billing effects.

[28] The 'treatment' terminology does not refer to receiving specific medical treatments. In this setup, the first shock is an absorbing state (i.e. a sick state) that permanently distinguishes individuals who have already had a shock versus those that have not yet had one (not-yet-treated), and so in different calendar months. There is no control group that does not suffer any shock. Focusing on the first shock ensures that the individuals do not switch treatment groups.

This interval is chosen so as to estimate the counterfactual over the whole two years after the shock for the earliest possible shock group, and over the whole first year for the latest possible shock group. Results are not sensitive to varying this binning (Schmidheiny and Siegloch, 2020). Relative shock time is normalized at the individual level to the pre-shock period $k = 0$, so that $\gamma_1^s$ measures the spending increase at the time of the shock, with the reference group being March, $S = 3$.

Specification (15) allows for heterogeneous treatment effects on spending in two dimensions. First, treatment effects are dynamic across leads and lags of the shock, given the focus on spending dynamics, in particular anticipation effects and shock persistence. Second, dynamic effects are possibly heterogeneous across treatment groups, via the interaction of the relative period indicators with the treatment group indicators. However, dynamic treatment effects are assumed to be homogeneous across individuals within a treatment group.[29]

Seasonality in baseline planned spending is controlled for by $\sigma_t$, which includes calendar month dummies, and a time trend. Calendar month dummies (1-12) take out differences in seasonal healthcare spending that would occur even in the absence of the shock. The differential moral hazard responses of interest induced by incentives changing over the calendar year remain identified. That is, e.g. for all relative months corresponding to December, they take out baseline spending on seasonal flu (homogeneous across groups), but not year-end bunching following the shock (heterogeneous across treatment groups). Furthermore, the term includes a $3^{\text{rd}}$ degree polynomial trend to account for secular trends (e.g. changes in prices and insurance premiums, medical technology, aging of the sample) that may lead to differences in spending across cohorts.[30] Individual fixed effects $\nu_i$ subsume any time-invariant individual characteristics, observed (e.g. gender, nationality), or unobserved (e.g. preferences, education, genetic predispositions, chronic diseases) that determine baseline individual spending and potentially correlate with shock timing. However, identification relies on between-individual variation. Some estimations include a vector of time-varying individual characteristics $X_{it}$ (age, type of insurance plan, accident insurance, and region fixed effects). Finally, $\varepsilon_{it}$ is random noise. Estimations are performed using linear least squares with standard errors clustered to allow for arbitrary correlation at the individual level.

[29] In Abraham and Sun (2021), this assumption corresponds to 'stationary' treatment effects, whereby each group of individuals with $S_i = s$ experiences the same average effect $\gamma_k^s$ in any given relative month. In other words, the cohort of individuals with a shock in March 2012 are assumed to have the same dynamic spending patterns as the March 2013 cohort, conditional on other included factors. Abraham and Sun (2021) show that if the effects are stationary, the estimates are consistent and have a causal interpretation.

[30] Calendar month dummies are identified using the spending patterns of all the individuals in the year before the shock, as well as the spending of the not-yet-treated. They would not be identified in a fully interacted specification, since the interaction term between the relative and treatment months is collinear with calendar month. A full set of period dummies cannot be identified because of the restrictions on the sample. The polynomial time trend is identified similarly.

To get estimates of $\Delta\tilde{\gamma}_k(s, \Delta s)$, I compute differences in spending by using estimated treatment effects across shock groups. To get $\hat{q}(s)$, I regress an indicator for keeping a high deductible in the year after the shock on the shock month, and adjust for age, gender, insurance plan type and year and region effects. This share turns out to be roughly constant in $s$, with or without adjusting. I use the average share of individuals with high deductibles. I plug the estimates into (11) as follows

$$\Delta\hat{\tilde{\gamma}}_k(s, \Delta s) \equiv \frac{\hat{\gamma}_k^{s+\Delta s} - \hat{\gamma}_k^s}{\hat{q}(s)\Delta s} \tag{16}$$

For instance, $\Delta\hat{\tilde{\gamma}}_k(3, 1)$ provides an estimate of the rescaled difference in spending between the March and April groups, holding all else equal. Comparing all shock groups pairs from March to October in relative months after the shock yields 140 data points for the derivative of the optimal timing moral hazard decision.

Recall that the last step in identifying timing moral hazard requires a functional form assumption for $m(s)$, as the system in differences is otherwise under-determined. I make this decision based on the data, rather than some theoretical functional form assumption. The observed negative linear relationship between the estimates $\Delta\hat{\tilde{\gamma}}_k(s, \Delta s)$ and shock timing (see Table 3) suggests that the integral $\rho(s)$ is a quadratic function of $s$

$$\rho(s) = \alpha + \beta s + \delta s^2 \tag{17}$$

where the planned care consumption for preponers comprises a constant, and a component that is allowed to vary across groups due to heterogeneous retiming costs. The difference between shock groups writes as

$$\rho(s + \Delta s) - \rho(s) = \alpha + \beta(s + \Delta s) + \delta(s + \Delta s)^2 - [\alpha + \beta s + \delta s^2] \tag{18}$$

$$\frac{\rho(s + \Delta s) - \rho(s)}{\Delta s} = \beta + 2\delta s + \delta\Delta s \tag{19}$$

where the last term captures approximation error from comparing shock months further away. I run the following regression to estimate the values of the parameters $\beta$ and $\delta$ that most closely predict the observed values of the derivative across comparison pairs $\{s, s + \Delta s\}$

$$\Delta\hat{\tilde{\gamma}}(s, \Delta s) = \beta + 2\delta s + \delta\Delta s + \epsilon \tag{20}$$

where and regressors are demeaned. Since $\rho(s)$ is not identified in levels, I bound the total timing moral hazard using $\rho(s) \geq 0$ for preponers, namely $\bar{\alpha} = -\beta s - \delta s^2$. Finally, I predict the total preponed amount (timing moral hazard) across shock groups using this bound and the estimated coefficients from the regression as

$$\hat{\rho}(s) = \bar{\alpha} + \hat{\beta} s + \hat{\delta} s^2 \tag{21}$$

### 4.1 Main estimation sample and shock definitions

The main analysis sample includes adults residing in Switzerland, aged between 19 (as minors have different contracts), and 90 (as the elderly have more likely particular end-of-life spending patterns). I keep individuals who move during the year without changing the other features of their plan, although they may face a change in premiums. However, incomplete insured-years with contract changes or interruptions are excluded (e.g., due to turning 18, emigration, military service). Individuals with temporary attrition are permanently excluded, as their full spending path is unobserved and may be influenced by other factors. I do not observe individuals before and after their contract with CSS, nor the reason for exiting the sample (e.g. switching insurers or death).

I impose further restrictions to increase the plausibility of the shock being unanticipated. I focus on individuals who had a high deductible of CHF 1,000 or above in the shock year, and did not exceed the deductible prior to the shock. Furthermore, I restrict the main sample to individuals who did not have a shock in the first or the last year of their observation window, since their spending dynamics are censored. This restriction excludes individuals who exit the sample in the year after the shock.

Table 1 provides descriptive statistics for individuals who enter the main analysis sample in Column (2), and compares it to high- and low-deductible samples in columns (1) and (3), respectively. As expected, the high-deductible sample without a shock is younger, and has a lower share of women than the low-deductible one, as these characteristics are strongly correlated with health status. The main analysis sample generally lies in-between the two in terms of average characteristics, insurance plan choices, and prices. It is on average 51 years old, 47% female, and 88% Swiss. Its premiums and other spending outcomes are higher on average than both other groups. These figures suggest that the sample includes individuals with relatively low baseline risk who suffer an adverse health event. High deductibles with a health shock constitute 9.2% of all insured-years in the data. The analysis thus relies on a selective but highly-relevant sample, as discussed further below.

### 4.2 Identifying assumptions

In this setup, the identification of causal spending responses relies on the timing of the health shock being exogenous. Selection into the shock is circumvented, as only individuals with a shock enter the analysis.[31] It implies that randomly-drawn individuals are comparable on average across treatment groups and provide valid counterfactual spending

[31] Appendix Figure B.1 illustrates the main steps of selection. Comparing individuals with different shock timing is somewhat more credible than comparing individuals with to those without a shock (see e.g. Kowalski 2016).

paths for each other. In other words, conditional on the shock, there are no time-varying unobservable factors that jointly influence spending and the probability of having a shock in a given calendar month (possibly conditional on observable characteristics and seasonality). Furthermore, the approach assumes that individuals do not (differentially) adjust their spending in anticipation of the shock or to meet its definition.[32]

Defining what constitutes a health shock requires particular care, as this affects the plausibility of the identifying assumptions, but also the estimation sample. Under alternative shock definitions, an individual might not enter the sample, or belong to another treatment group. Under a given shock definition, individuals should be comparable. Table 2 shows descriptive evidence that there are no systematic differences in observable characteristics across treatment groups for the main definition of a shock. This provides supportive evidence that there is no selection into treatment groups, nor dynamic compositional effects as the share of not-yet-treated individuals decreases over the calendar year. Age and gender in particular are strong predictors of morbidity, and are balanced across groups. The number of observations points to shocks occurring throughout the year, despite some seasonality which closely follows the same yearly pattern as average healthcare consumption. Appendix Table B.1 presents t-test of sample averages for each group relative to the pooled sample. Any significant differences are small in magnitude and do not display a seasonal pattern.

Defining the shock in terms of spending has the advantage of generality, as comparability is not guaranteed by knowing the exact nature of the shock (e.g. heart attacks may unobservably differ in their severity). To provide supportive evidence that shock timing is random, I compare the magnitude and composition of spending at the shock across timing groups. Table 2 reassures that the shock is comparable across groups in terms of severity. The magnitude and paths of spending are very similar to the main results, and the shock mainly consists of inpatient spending, which is less amenable to strategic timing and reduces the likelihood that the shock itself is a moral hazard response.

Treatment groups may still differ in unobservable characteristics. To bias the effect estimates, unobservable confounders have to be correlated with both shock timing and spending, e.g. unobservable health deteriorations (more severe patients are more likely be treated earlier in the year and spend more), or adverse non-health events such as job loss. If such time-varying confounders drive anticipatory spending before the shock, this would transpire in the leads on the treatment. I control for compositional differences in observable time-varying characteristics of the individuals and their shock. The results are not sensitive to these adjustments.

[32]Identification also requires that the stable unit treatment value assumption (SUTVA) holds. This stipulates that an individual's observed outcome under a given treatment equals their potential outcome under that treatment. Hence, one individual's outcome with a given shock timing does not affect the outcomes of those that suffer a shock at another time, which appears plausible in this setting. Health system capacity constraints can be controlled for by region fixed effects.

Table 1: Summary statistics

| | (1) High deductibles | (2) High deductibles with health shock | (3) Low deductibles |
|---|---|---|---|
| *Demographics* | | | |
| Age | 45.34 (13.87) | 51.20 (14.79) | 45.68 (15.67) |
| Female | 0.45 | 0.47 | 0.53 |
| Swiss | 0.86 | 0.88 | 0.86 |
| *Insurance plan* | | | |
| Premiums | 2,720 (797) | 2,946 (905) | 3,620 (847) |
| CHF 2500 deductible | 0.29 | 0.23 | 0.00 |
| Standard plan | 0.51 | 0.56 | 0.60 |
| Other plan type | 0.49 | 0.44 | 0.40 |
| Accident insurance | 0.36 | 0.47 | 0.47 |
| *Spending* | | | |
| Total out-of-pocket spending | 3,286 (1,179) | 3,824 (1,299) | 4,035 (976) |
| Total annual spending | 1,289 (4,356) | 3,130 (7,730) | 2,678 (6,684) |
| 25th percentile | 0 | 142 | 174 |
| 50th percentile | 206 | 784 | 850 |
| 75th percentile | 949 | 2,970 | 2,505 |
| *Prices* | | | |
| Exceeded deductible | 0.17 | 0.40 | 0.66 |
| Cost-sharing | 0.88 | 0.71 | 0.48 |
| Insured-years | 2491316 | 291767 | 673749 |

*Notes:* The table presents means and standard deviations (in parentheses) for samples of insured-years. High deductibles (column 1) are insured-year observations with annual deductibles of CHF 1,000 to 2,500, and low deductibles are CHF 300 and 500 (column 3). A health shock is defined as monthly spending above CHF 2,500 for the first time in the observation window. High deductibles with health shock (column 2) refers to the main analysis sample of a shock defined as monthly spending of CHF 2,500 or more for the first time in the observation period. All spending is in Swiss Francs (CHF). Cost-sharing is calculated as out-of-pocket spending (net of premiums) over total yearly healthcare spending. Total out-of-pocket spending includes insurance premiums.

Table 2: Summary statistics by shock timing

| | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Pooled |
|---|---|---|---|---|---|---|---|---|---|
| *Demographics* | | | | | | | | | |
| Age | 51.44 | 50.98 | 51.12 | 50.80 | 49.79 | 50.31 | 49.91 | 49.73 | 50.58 |
| Female | 0.48 | 0.47 | 0.49 | 0.48 | 0.45 | 0.46 | 0.47 | 0.47 | 0.47 |
| Swiss | 0.86 | 0.87 | 0.87 | 0.86 | 0.86 | 0.88 | 0.86 | 0.87 | 0.87 |
| *Insurance plan* | | | | | | | | | |
| Premiums | 2,817 | 2,769 | 2,766 | 2,777 | 2,737 | 2,723 | 2,740 | 2,704 | 2,759 |
| CHF 2500 deductible | 0.27 | 0.27 | 0.28 | 0.29 | 0.31 | 0.32 | 0.31 | 0.31 | 0.29 |
| Standard plan | 0.54 | 0.53 | 0.53 | 0.52 | 0.51 | 0.52 | 0.50 | 0.51 | 0.52 |
| Accident insurance | 0.49 | 0.46 | 0.47 | 0.45 | 0.45 | 0.45 | 0.45 | 0.43 | 0.46 |
| *Spending at shock* | | | | | | | | | |
| Total spending | 4,860 | 4,797 | 4,845 | 4,836 | 4,994 | 4,868 | 4,863 | 4,805 | 4,856 |
| Share of physician outpatient spending | 0.13 | 0.11 | 0.13 | 0.13 | 0.11 | 0.12 | 0.14 | 0.12 | 0.12 |
| Share of hospital outpatient spending | 0.16 | 0.16 | 0.16 | 0.17 | 0.16 | 0.14 | 0.16 | 0.15 | 0.16 |
| Share of hospital inpatient spending | 0.63 | 0.65 | 0.63 | 0.62 | 0.65 | 0.66 | 0.62 | 0.66 | 0.64 |
| Share of drugs spending | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.03 | 0.04 | 0.04 | 0.04 |
| Share of other spending | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 |
| *Pre-shock spending* | | | | | | | | | |
| Cumulated spending 12 months before shock | 1,388 | 1,215 | 1,104 | 1,020 | 916 | 865 | 878 | 841 | 1,054 |
| Cumulated spending before shock | 441 | 513 | 547 | 587 | 593 | 610 | 669 | 700 | 573 |
| Total years observed | 10.60 | 10.61 | 10.64 | 10.60 | 10.45 | 10.60 | 10.62 | 10.47 | 10.58 |
| Insured | 4599 | 3724 | 3605 | 3420 | 3018 | 2852 | 3211 | 3153 | 27582 |

*Notes:* The table presents insured-level sample means by calendar month of the shock (treatment group), as measured in the year of the first health shock (treatment). The sample is insured with high deductibles with a shock. All spending in Swiss Francs (CHF). Cost-sharing is calculated as out-of-pocket spending (net of premiums) over total yearly healthcare spending. Appendix Table B.1 reports t-tests of sample differences.

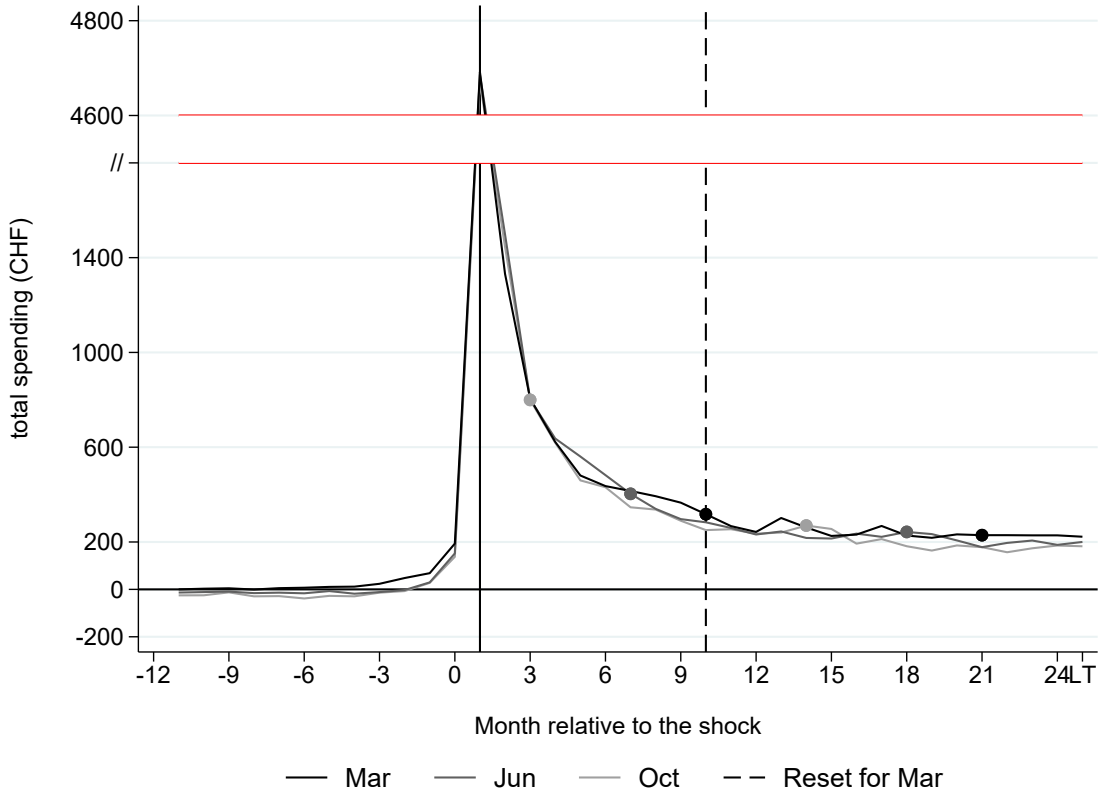
# 5 Results

## 5.1 Dynamic spending patterns and incentives

This section describes the reduced-form effect of the shock on spending dynamics, and provides first evidence of persistence and responses varying with shock timing. Figure 3 depicts the coefficients on dynamic treatment effects for selected shock groups from the event study. Several striking patterns emerge. First, average pre-shock differences in spending are small and would not lead to exceeding the deductible. This result supports the assumption that treatment groups do not differ systematically in their pre-shock spending. There is however a slight increase in spending of around CHF 100 two months before the shock. The approach is still valid if the timing of this health deterioration is random and its magnitude comparable across groups. Second, the close magnitude of the spikes at the shock suggest that these are comparable across groups. The estimated differences in spending in the month of the shock are of the same order of magnitude as in other periods (mean CHF 15, SD=99) for all shock months, which supports that shock groups are comparable in severity. Third, spending gradually decreases as expected and stabilises roughly one year after the shock, with a long-term effect of around CHF 200 per month, i.e. CHF 2,400 per year. This figure is close to the highest deductible of CHF 2,500. This pattern supports that shocks persist, and that the strength of these spillover effects on prices in the post-shock year varies across groups. The persistence is sufficiently large to induce differential incentives across shock groups in the year after the shock. Finally, these spending dynamics support the binning choice for lags and leads in the event study, as past and future effects of the shock converge across groups.

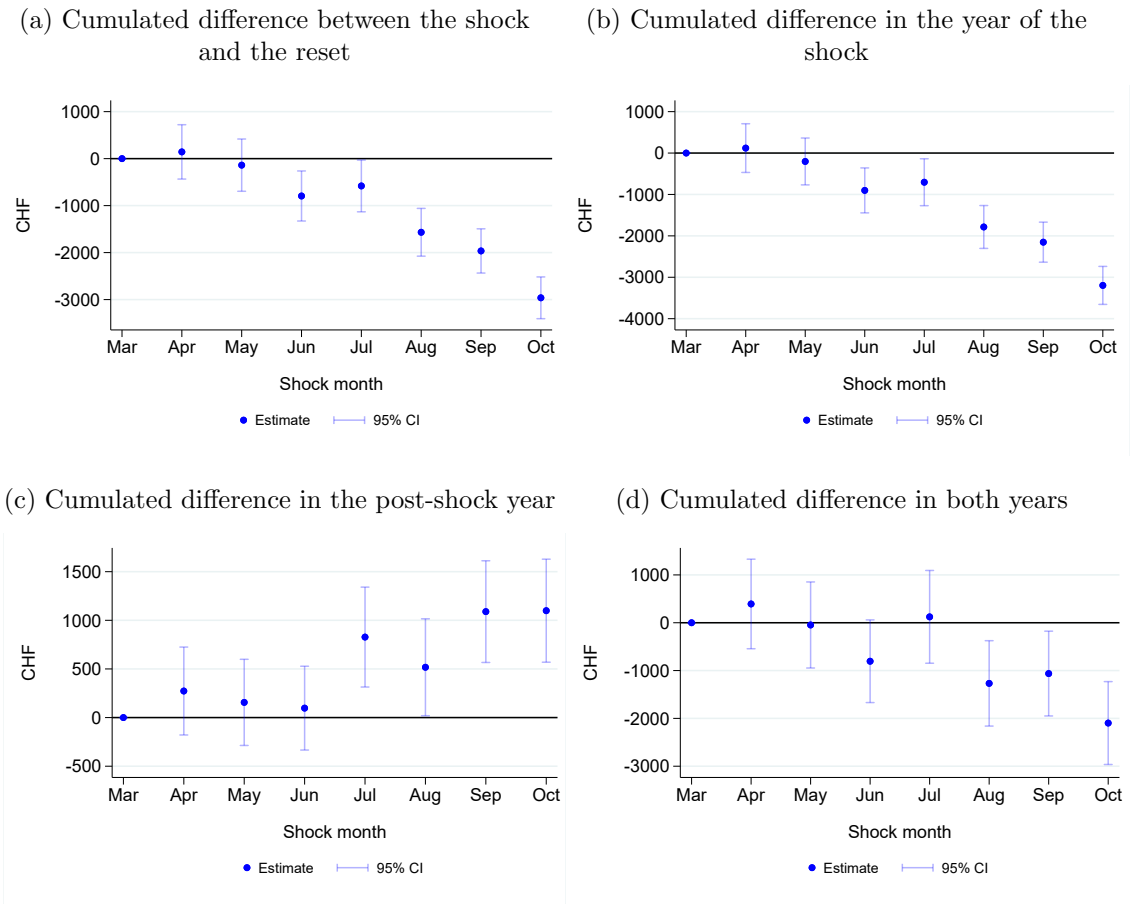
I provide additional supportive evidence of differential shock spillovers by showing that cumulated spending in the year of the shock, and the year after systematically vary across groups. I compute differences in dynamic treatment effects in calendar time between the shock and the year-end reset, over the whole shock year, the post-shock year, as well as in both years taken together. Details are given in Appendix B.3. The measures are presented in Figure 4, and confirm that shock timing induces spillovers. As shown in panel (a), there is a mechanical significant negative relationship between shock timing and total spending between the shock and the year-end reset, as individuals do not have the same amount of time to offset the shock in its calendar year. The difference exceeds CHF 3,000 for the October relative to the March group. Panel (b) shows similar patterns with respect to the cumulated differences over the whole calendar year of the shock. This confirms that there are no significant differences in cumulated spending prior to the shock. Panel (c) indicates the presence of spillovers—the later the shock occurs, the higher the total spending in the year after. However, this difference only goes up to CHF 1100. Taking both years together in panel (d), larger spending in the post-shock year offsets

Figure 3: Event study of spending around the health shock




*Notes:* The figure depicts the coefficient estimates on monthly treatment effects of the shock from the event study for selected treatment groups with shocks in March, June and October, for the main analysis sample of insured with a high deductible. These effects are normalized to the average spending of the March group up to 12 months before the shock. The dots indicate the last month before the year-end reset in years after the shock. The last point estimate denotes the long-term effect (LT) of the shock, i.e. the average after 24 months.

Figure 4: Differences in cumulated spending in calendar time




*Notes:* The figures depict cumulated differences in unadjusted dynamic treatment effects in calendar time between (a) the shock and the year-end reset, (b) over the whole shock year 1, (c) the post-shock year 2, as well as (d) in both years taken together. Details are presented in Appendix B.3. All differences are in Swiss Francs (CHF), and taken relative to the March group. Confidence intervals at the 5% level based on bootstrapped standard errors with 49 replications, clustered at the individual level.

lower spending in the shock year for early shock groups, but not for later ones.

To support that the price sequences observed empirically align with the model, I regress prices at different times over the shock and post-shock years on shock timing, as well as observable individual characteristics. Appendix Table B.7 shows coefficient estimates. Year-end prices in the shock year (column 1) are similar across groups, which suggests that incentives between the shock and the reset are aligned. The differences are precisely estimated, but the magnitude of up to CHF 0.02 is not economically meaningful. They are due to individuals mechanically accumulating spending over different time horizons throughout the rest of the shock year, and exiting the co-payment region. This result supports that the dichotomized price structure in the model is a reasonable approximation. The framework also postulates that shock persistence is either low or high enough to

exogenously vary prices in the post-shock year. Year-end prices in the second year (column 2) are not significantly different, but are harly interpretable as endogenous to the moral hazard and selection decisions, and further realized health events. Later shock groups have a higher probability of exceeding the (chosen) deductible in January of the post-shock year (column 3), and hit the deductible earlier (column 4). The latter two patterns support differential shock spillovers.

### 5.2 Estimates of timing moral hazard

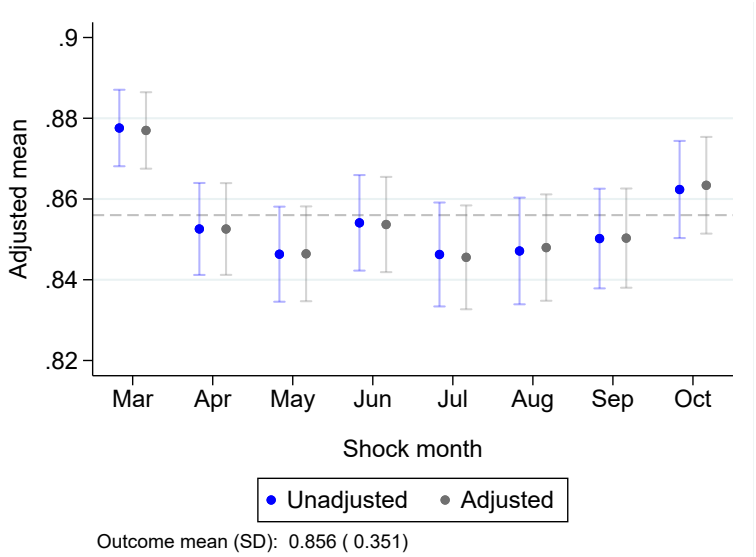
This section presents estimates for the inputs for timing moral hazard. First, Figure 5 shows the raw and the adjusted shares of preponers as identified by individuals who keep a high deductible in the year after the shock. The share is roughly constant in both specifications at nearly 86% across shock groups. Only individuals with a shock in March have a significantly larger share than the average, but the difference of 2 percentage points is not economically meaningful. In other words, only about 14% of individuals in the main estimation sample switch to a low deductible.[33] This result provides support for the assumption that $q(s) \approx q(s+\Delta s)$ Within the model, comparative statics for condition (8) imply that for later shock groups, the larger nondiscretionary needs may be offset by lower preponed amounts and preponing costs, such that the share of preponers stays roughly constant. This can explain the homogeneity in the share of individuals keeping a high deductible.[34]

Second, Table 3 displays summary statistics for estimates of monthly differences in total healthcare spending. The raw average difference in coefficients $\Delta\hat{\gamma}_k(s, \Delta s)$ from the event study is nearly CHF 18, and ranges between CHF -216 and 248. The differences adjusted for covariates yield very similar results. The table also displays the rescaled difference $\Delta\hat{\tilde{\gamma}}_k(s, \Delta s)$, as in equation (11). The estimates are smaller on average at CHF 7.50. The share of estimates that are statistically significantly different from 0 at the 10% level is 26%. Hence, individuals have significantly different spending patterns after the shock depending on its timing. This important intermediate result confirms the existence of timing moral hazard, as differences in responses are sufficient to establish the existence of timing moral hazard within my framework.

[33] In Appendix Table B.7, I check that the share of individuals choosing a standard plan is constant across shock groups, such that there is no differential selection into alternative managed care, family doctor or telemedicine plans that limit the flexibility in choosing healthcare providers.

[34] Although beyond the scope of this paper, a large literature has shown that individuals do not choose their utility-maximizing health insurance plan, due to e.g., switching costs, inattention or inertia (see e.g., Abaluck and Gruber 2011; Handel 2013; Abaluck and Gruber 2016; Handel and Kolstad 2015; Heiss et al. 2016, and Winter and Wuppermann 2019 for a review of the recent literature). Re-optimizing health insurance plan choice might be more challenging in the face of a large health shock. The panel only contains individuals who are observed for at least five years, so that they are prone to keeping the same insurer.[35]

Figure 5: Adjusted share of individuals keeping a high deductible (preponers)




*Notes:* The figure displays the adjusted share of preponers $q(s)$, as identified by individuals who keep a high deductible in the year after the shock, and so across shock months. The blue dots are raw averages, the grey are adjusted at the mean for demographic characteristics (gender, age, nationality, canton of residence) and time fixed effects. The dotted horizontal line denotes the sample average. Confidence intervals are at the 95% level, based on robust standard errors.

Third, I investigate the relationship between the estimated monthly differences and shock timing. This step serves to inform a data-driven choice of the functional form for the optimal timing moral hazard function (17). Table 4 presents regression results from regressions as in (20). The estimate for $\beta$ (i.e. the constant in the regression) is not significantly different from 0. The coefficient on shock timing, i.e. $2\delta$ in the regression, is negative at around CHF -13. It is robust in magnitude and becomes statistically significant when controlling for the size of the underlying difference in timing $\Delta s$. These results support the choice of a quadratic functional form for the timing moral hazard decision.[36] Moving forward, I use coefficient estimate from column (3) given they are most precise.

Importantly, this regression indicates that there is a significant variation in the amount of care preponed as a function of shock timing. It allows me to reject the frictionless case, where the total preponed amount would be the same. It also provides estimates for the parameters necessary to predict timing moral hazard as in (21). Figure 6 shows estimates of the predicted yearly timing moral hazard (i.e. the monthly prediction multiplied by the number of months between the shock and the reset). It suggests that the retiming response among preponers is substantial, and reaches nearly CHF 2800 for the earliest

[36] I cannot reject the null that the coefficient on $s$ equals twice the coefficient on $\Delta s$ (p-value = 0.725). This specification also minimizes the Akaike information criterion. Higher order terms are not significant.

Table 3: Summary statistics for monthly spending differences

| Panel (a). Unadjusted | Mean | SD | Min | Max |
|---|---|---|---|---|
| Monthly difference | 18.46 | 85.43 | -213.32 | 248.41 |
| Rescaled monthly difference | 7.64 | 61.69 | -177.15 | 232.41 |
| Significant at 10% | 0.26 | 0.44 | 0.00 | 1.00 |
| Observations | 140 | | | |

| Panel (b). Adjusted | Mean | SD | Min | Max |
|---|---|---|---|---|
| Monthly difference | 19.48 | 85.48 | -211.65 | 250.65 |
| Rescaled monthly difference | 8.09 | 61.75 | -176.99 | 233.78 |
| Significant at 10% | 0.26 | 0.44 | 0.00 | 1.00 |
| Observations | 140 | | | |

*Notes:* The table displays summary statistics for the estimated monthly differences in spending. The raw average difference corresponds to estimates of $\Delta\gamma_k(s, \Delta s)$. The rescaled difference corresponds to $\Delta\tilde{\gamma}_k(s, \Delta s)$, as in equation (11). Panel (a) presents figures using unadjusted event study estimates, and panel (b) covariate-adjusted ones.

shock groups. In other words, part of the sample is forward-looking and responds to incentives enough to prepone care. The response however decreases substantially with shock timing. The latest shock groups have the lowest retiming response, due to $\rho'(s) < 0$. I use their prediction as a lower bound to pin down the response level $\bar{\alpha}$. This result is not sensitive to using unadjusted difference estimates.

Table 4: Estimates of the timing moral hazard parameters

| | Monthly difference in total spending | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Coefficient on $s$ ($2\delta$) | -13.91 | -13.83 | -16.14* | -16.05* |
| | (7.29) | (7.30) | (6.60) | (6.61) |
| Coefficient on $\Delta s$ ($\delta$) | | | -5.89 | -5.85 |
| | | | (3.58) | (3.63) |
| Constant ($\beta$) | 7.64 | 8.09 | 7.64 | 8.09 |
| | (10.66) | (10.74) | (10.70) | (10.78) |
| Observations | 140 | 140 | 140 | 140 |
| Adjusted | No | Yes | No | Yes |

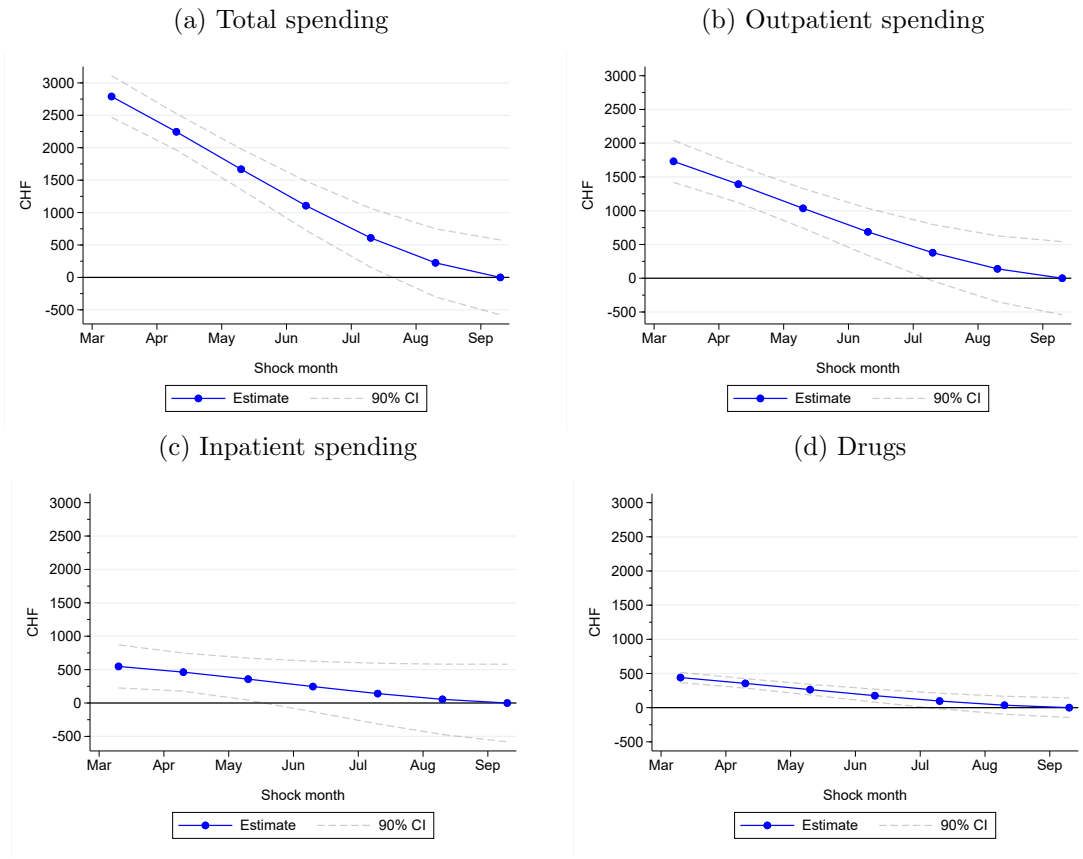
*Notes:* The table displays coefficient estimates from a regression of monthly differences in spending $\Delta\tilde{\gamma}_k(s, \Delta s)$ on shock timing, as in (20), where regressors are demeaned. Adjusted regressions in columns (2) and (4) use differences from a covariate-adjusted event study. Confidence intervals at the 5% level based on bootstrapped standard errors with 49 replications, clustered at the individual level.

Table 5: Estimates of yearly timing moral hazard responses for alternative samples

| | Total care spending | Outpatient care spending | Inpatient care spending | Drugs spending |
|---|---|---|---|---|
| Lower bound $\bar{\alpha}$ | 1667.87 | 1034.83 | 344.71 | 263.89 |
| | (170.48) | (158.42) | (171.04) | (42.57) |
| March prediction | 2790.11 | 1731.07 | 547.86 | 440.66 |
| | (93.65) | (54.12) | (78.69) | (18.46) |
| September prediction | 0.00 | 0.00 | 0.00 | 0.00 |
| | (170.48) | (158.42) | (171.04) | (42.57) |
| Mean prediction | 1234.91 | 766.20 | 258.60 | 195.48 |
| | (98.18) | (82.94) | (94.88) | (23.63) |

*Notes:* The table presents a summary of the total yearly timing moral hazard response, predicted as in (21) using estimates from column (3) of Table 4 and Appendix Table B.3 for the respective types of spending. The last shock month serves as a lower bound. Standard errors in parentheses based on bootstrapped standard errors with 49 replications, clustered at the individual level.

Figure 6: Predicted timing moral hazard




*Notes:* The figure presents estimates of the total yearly timing moral hazard response across shock months, predicted as in (21) using estimates from column (3) of Table 4. The last shock month serves as a lower bound. Confidence intervals at the 5% level based on bootstrapped standard errors with 49 replications, clustered at the individual level.

# 6 Discussion and robustness checks

## 6.1 Magnitude

Is the magnitude of the estimated timing moral hazard responses realistic? I relate the estimates to total and cumulated differences in spending. Appendix Table B.6 presents summary statistics for spending across shock months. Notice that the total timing response does not exceed actual spending in the year of the shock for any of the months. The pattern and magnitude of the timing response also align with the cumulated spending differences in Figure 4. This comparison suggests that a large part of the differences in spending in the year of the shock are related to timing responses. Spending is lower for early shock groups in the year after, which lends support to the interpretation that spending is shifted in time, given similar deductible switching rates. The model does not restrict planned care $\mu$ to be unrelated to the shock. It is then possible that some of the needs planned for year 2 are due to the shock, but can be adressed more or less early. Hence, their planned spending after the shock is relatively large, and so is the scope for preponing. Kowalski (2016) finds larger spending responses to family injuries occurring in the first half of the year, which might be consistent with differential timing moral hazard.

## 6.2 Possible sources of frictions

The high degree of heterogeneity across shock groups points to the existence of dynamic frictions in preponing. I now discuss the possible nature and sources of these frictions. While my approach allows remaining agnostic on this matter, it is relevant for policies aiming to adress timing moral hazard.

**Constraints.**— Various constraints specific to healthcare markets can restrict individuals from retiming flexibly. Time constraints can stem from healthcare supply, through e.g., the imperfect control over the timing of appointments or the need for obtaining referrals. They can make preponing more difficult with little time left until the reset. In particular, as many patients increase their spot consumption towards the end of the year after exceeding the deductible (as in e.g., Lin and Sacks 2019; Gerfin et al. 2015), capacity constraints might prevent them from timing additional appointments to that period.

Another feature of healthcare consumption is its lumpiness (Einav et al., 2015; Cabral, 2017). Patients can typically not retime a continuous amount of consumption, as medical treatments come in bundles reimbursed at a given price, and are possibly administered over the course of a series of appointments that cannot be compressed. These bundles might be easier to retime as a whole into a longer time horizon. Certain treatments are also less amenable to being retimed. While emergent procedures have to be received

immediately, certain elective procedures could in principle be shifted over several years (e.g., a hip replacement). I further analyse and discuss heterogeneity in preponing across types of care in the next section.

**Explicit costs.**— Following Handel (2013), explicit costs are driven by factors that induce direct losses of utility or money. Retiming induces time and effort costs, e.g., to schedule doctor's appointments. The costs of searching for a healthcare provider who can accommodate the necessary treatments might also prevent preponing, especially on short notice. Such transaction costs may be exacerbated when individuals are still overcoming the acute phase of the shock. Then, individuals who experience the shock close to the deductible reset might be less inclined to schedule additional appointments within a short time horizon.

**Implicit costs.**— Implicit costs are driven by factors that indirectly lead to utility losses from suboptimal choices. The model above describes the behavior of a rational, forward-looking individual. However, the recent literature has highlighted several behavioral biases on the demand side that prevent the insured from achieving their optimal healthcare consumption. A longer horizon until the reset leaves more time for individuals to internalize the consequences of the shock, and to respond to the incentive to prepone. Meanwhile, myopic individuals might not foresee the incentives to prepone or future planned care (Abaluck et al., 2018; Dalton et al., 2020). Individuals with high deductibles have been found to underspend early in the year to avoid out-of-pocket costs under the deductible even if they can expect to exceed it, which may lead to unmet health needs (Brot-Goldberg et al., 2017). My findings provide support for individuals not being fully myopic after a shock. Procrastination (O'Donoghue and Rabin, 1999) would favour delaying rather than advancing treatments, leading to smaller preponing for those who are close to the deductible reset. By nature, these costs differ from the explicit transaction and search costs described above. Characterizing the sources and implications of bounded rationality for strategic timing of healthcare consumption is an interesting avenue for future research.

### 6.3 Heterogeneity across categories of spending

I now provide evidence that healthcare spending that is more amenable to retiming constitutes the largest share of the total preponing response. I estimate the timing moral hazard response for three categories of healthcare spending: outpatient, inpatient, and prescribed drugs.[37] Table B.5 presents summary statistics for the estimated responses

[37] Outpatient spending includes all ambulatory care covered by BHI received at practices and hospitals. Inpatient care is defined as stationary care received during a hospitalization with an overnight stay. Drugs are filled prescriptions issued by a physician.

across spending categories.[38] Notice that adding the timing responses across types of care matches the total response, up to some estimation error. This supports the robustness of the approach. The retiming of outpatient care constitutes over 60% of the total timing response, and reaches CHF 1678 for early shock groups. This category arguably includes medical procedures that are easier to retime, e.g. follow up care after the shock. In contrast, inpatient care amounts to only 20% of the timing response, up to CHF 540. The estimates for this category are generally less precise as inpatient spending is more variable. As for spending on drugs, the response goes up to CHF 427, i.e. 16% of the total response. Drugs are particularly easy to stock up on, and have been the focus of, e.g., Einav et al. (2015). Figure 6 shows that timing responses decrease in later shock groups across all categories.

These findings yield important insights on the extent and composition of timing responses. They support the interpretation of the estimated spending differences as being driven by medical procedures that can be shifted in time. The composition of the timing response constrasts with that of total healthcare spending, with inpatient care representing more than half of the total healthcare spending in the year of the shock (see Table 2). Conceptually, this contrast fits with the distinct modelling of a nondiscretionary shock with random, unalterable timing, and planned care amenable to retiming in the theoretical framework. From a health insurance policy perspective, an implication of this heterogeneity is that medical procedures differ in terms of their propensity for being retimed. In the Swiss context, BHI covers equally emergent and non-emergent procedures. My findings suggest that the two types of care interact, as nondiscretionary needs shape timing incentives. Whether this pooling is desirable from a welfare perspective is an important question for policy. Importantly, retiming frictions are relevant for all three categories of care, as the responses decreases with shock timing. A shorter time horizon decreases the amount preponed for all types of care.

### 6.4 Shock timing randomness

As discussed above, identification relies on individuals not systematically manipulating the timing of the shock. This assumption requires particular care when the shock is defined in terms of spending, as the shock itself should consist of nondiscretionary spending (as in the model), and not be initiated by a moral hazard response.

Consider an individual below the deductible who learns that they require a costly procedure around the end of the year. They have an incentive to delay it to benefit from a lower price over the whole next year. By doing so, they would enter a shock group in the next calendar year, which may then yield a selected group of individuals prone

[38] Appendix B.4 presents further results on the underlying inputs, computed with the same steps as for total healthcare spending.

to timing moral hazard. However, rational individuals would choose a low deductible for that year in anticipation of the large spending (excluding any large switching costs, which are not modelled here), as the low deductible plan dominates the high one for such amounts. Given the sample restriction of having a high deductible in the year of the shock, they would then not enter the analysis sample. Furthermore, my analysis conservatively excludes the earliest and latest months to avoid any turn-of-the-year effects.

The differential timing response should also not be driven by individuals with early shocks having more severe health deteriorations. As individuals mechanically accumulate spending throughout the year (in increments smaller than the shock definition), they may approach the deductible prior to the shock itself. Individuals close to the deductible can decide to move to the zero price segment endogenously, and engage in either type of moral hazard. A large spending induced by a moral hazard response would then qualify them for entering a treatment group, and is more likely for later shock groups.

Appendix B.5 presents results for alternative samples and shock definitions. The main result of decreasing preponing with shorter horizons remain qualitatively robust to altering the key sample and shock definition parameters. First, I restrict the estimation to a subsample of individuals who accumulated less than half of their deductible in spending before the shock (panel a). The estimates for total timing moral hazard are slightly larger than the main sample. Second, I consider individuals who exceed the deductible for the first time in the observation period (panel b), who have smaller responses than the main sample. Third, I restrict the sample to those who have the highest deductible of CHF 2,500 in the year of the shock. Those have larger responses. Fourth, I reduce the threshold for spending that defines a shock to CHF 1,500, which yields smaller responses. Taken together, these results suggest that individuals who are less severe at baseline have stronger incentives to prepone after a large shock, as they expect smaller nondiscretionary long-term needs than the main sample. However, smaller shocks may also yield smaller amounts of planned care to shift.

### 6.5 External validity

The analysis relies on a selected sample of individuals with high deductibles, who become high spenders due to a large health shock. Understanding the behavior of this population is particularly relevant for policy. Costly health events may have lasting consequences on individual and collective economic outcomes, beyond health and healthcare spending. Dobkin et al. (2018) find that unanticipated hospitalizations increase out-of-pocket healthcare spending and negatively affect earnings, income, access to credit, and consumer borrowing. A specificity of healthcare markets is that a small share of high-spending individuals generate a large share of costs. In my data, the main sample accounted for on average 7% of the insured per year, and 24% of all healthcare spending observed in the

year of the shock. However, several characteristics of the setup should be noted when considering external validity.

First, the large spending shock makes exceeding the deductible particularly salient for individuals who have initially chosen a high deductible. Some of the highest price-elasticity estimates for within-year moral hazard were found using exogenous variation in exceeding the deductible (e.g. Kowalski 2016). Smaller price changes or health shocks, as well as lower deductibles might render incentive changes less salient and prevent individuals from forming correct expectations about the year-end price (Brot-Goldberg et al., 2017). Further evidence using alternative sources of variation in relative prices across years would be useful to determine whether a similar mechanism applies to the timing elasticity. Second, despite the salient shift in incentives, it is possible that preference-wise the sample has a lower taste for within-year moral hazard, as individuals with higher price-sensitivity select into greater coverage *ex ante* (Einav et al., 2013). Third, the present setting and available data do not rely on restricting the nature of the shock or illness. The estimates capture the response to any health event that triggers a large spending amount, instead of focusing on narrow treatments or diseases. The approach could in its essence be applied with alternative definitions of a shock leading to exceeding the deductible, e.g. using data on medical procedures and identifying those that reflect arguably emergent and unanticipated health needs.

# 7 Conclusion

In this paper, I introduce a new approach to identifying timing moral hazard in health insurance, with a setup that allows for within-year moral hazard and deductible choice. I consider high-deductible individuals who suffer a large health shock, and exploit the timing of a health shock within the coverage year as a source of variation. This paper brings important theoretical and empirical insights into how individuals respond to the strategic timing incentives created by nonlinear cost-sharing schedules. With the model, I show that timing moral hazard is tied to the choice of coverage purchase, and influenced by the taste for within-year moral hazard. Individuals with a higher price-sensitivity retime so as to draw additional utility benefits from higher consumption. I also highlight that not only the *ex ante* risk, but also the timing of the shock matters for *ex post* spending responses. Empirically, I find that the total amount of preponed spending is substantial, and reaches over CHF 2,500. However, individuals who have a shock late in the calendar year have a significantly lower retiming response than those who have a shock earlier.

These results have implications for our understanding of health insurance markets and designing policies aiming at containing collective healthcare spending. First and in line with previous studies, timing moral hazard affects our interpretation of existing estimates

for price-elasticities of healthcare consumption. Part of the increase in spending after hitting a deductible is due to preponing planned care, rather than pure price responses mainly studied so far. Distinguishing between price and timing margins is key in analyzing price-elasticities. Second, individuals are not fully myopic. They re-optimize in the face of a bad risk realization so as to minimize future out-of-pocket costs. To this end, they are able to advance potentially large amounts of planned healthcare consumption.

Third, the avoidance of the co-payment in nonlinear cost-sharing can generate externalities via increases in premiums in the insurance pool. The insurer can indeed observe the realized timing of the shock and subsequent consumption, and price plans accordingly. Based on my partial equilibrium results, the extent of these externalities might depend dynamically on the length of the time period towards which care can be retimed. In light of this, the length of the coverage period seems a relevant (and salient) policy tool for shaping strategic timing incentives. Shorter coverage lengths can limit preponing, but might lead to adverse selection and increased within-year moral hazard due to shock spillovers into the next coverage period. However, my findings suggest that retiming is relatively easier than switching for the considered population, given the heterogeneity in retiming and the homogeneity in deductible switching rates.

Finally, since medical services differ in their amenability for retiming, an important policy question is whether a single co-payment schedule should apply to all healthcare consumption. Further understanding the sources of frictions to retiming, and how they affect future health outcomes, would allow designing targeted contract features.

## References


Abaluck, Jason and Jonathan Gruber (2011). "Choice inconsistencies among the elderly: evidence from plan choice in the Medicare Part D program", *American Economic Review*, 101(4): 1180–1210.

——— (2016). "Evolving choice inconsistencies in choice of prescription drug insurance", *American Economic Review*, 106(8): 2145–84.

Abaluck, Jason, Jonathan Gruber, and Ashley Swanson (2018). "Prescription drug use undere Medicare Part D: A linear model of nonlinear budget sets", *Journal of Public Economics*, 164: 106–138.

Abraham, Sarah and Liyang Sun (2021). "Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects", *Journal of Econometrics*, 225(2): 175–199.

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

Arrow, Kenneth J. (1963). "Uncertainty and the Welfare Economics of Medical Care", *American Economic Review*, 53(5): 941–973.

Brot-Goldberg, Zarek C, Amitabh Chandra, Benjamin R Handel, and Jonathan 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(3): 1261–1318.

Cabral, Marika (2017). "Claim Timing and Ex Post Adverse Selection", *Review of Economic Studies*, 84 1–44.

Card, David, Carlos Dobkin, and Nicole Maestas (2009). "Does Medicare Save Lives?", *Quarterly Journal of Economics*, 124(2): 597–636.

Dalton, Christina M., Gautam Gowrisankaran, and Robert Town (2020). "Salience, Myopia, and Complex Dynamic Incentives: Evidence from Medicare Part D", *Review of Economic Studies*, 87: 822–869.

Diamond, Rebecca, Timothy Dickstein, Michael J.and McQuade, and Petra Persson (2021). "Insurance without Commitment: Evidence from the ACA Marketplaces", *NBER Working Paper 24668*.

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo (2018). "The Economic Consequences of Hospital Admissions", *American Economic Review*, 108(2): 308–352.

Einav, Liran and Amy Finkelstein (2018). "Moral Hazard In Health Insurance: What We Know And How We Know It", *Journal of the European Economic Association*, 6(4): 957–982.

Einav, Liran, Amy Finkelstein, Stephen P Ryan, Paul Schrimpf, and Mark R Cullen (2013). "Selection on moral hazard in health insurance", *American Economic Review*, 103(1): 178–219.

Einav, Liran, Amy Finkelstein, and Paul Schrimpf (2015). "The response of drug expenditure to nonlinear contract design: Evidence from medicare part D", *The Quarterly Journal of Economics*, 130(2): 841–899.

Ellis, Randalll P, Bruno Martins, and Wenjia Zhu (2017). "Health care demand elasticities by type of service", *Journal of Health Economics*, 55: 232–243.

Finkelstein, Amy (2014). *Moral hazard in health insurance*, Columbia University Press.

Gerfin, Michael (2019). "Health Insurance and the Demand for Healthcare", in *Oxford Research Encyclopedia of Economics and Finance*, Oxford University Press.

Gerfin, Michael, Boris Kaiser, and Christian Schmid (2015). "Healthcare demand in the presence of discrete price changes", *Health economics*, 24(9): 1164–1177.

Grossman, Michael (1972). "On the Concept of Health Capital and the Demand for Health", *Journal of Political Economy*, 80(2): 223–255.

Grubb, Michael D. and Matthew Osborne (2015). "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock", *American Economic Review*, 105(1): 234–271.

Handel, Benjamin R. (2013). "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts", *American Economic Review*, 107(7): 2643–2682.

Handel, Benjamin R. and Jonathan T. Kolstad (2015). "Sinking, Swimming, or Learning to Swim in Medicare Part D", *American Economic Review*, 105(8) 2449–2500.

Heiss, Florian, Daniel McFadden, Joachim Winter, Amelie Wuppermann, and Bo Zhou (2016). "Inattention and Switching Costs as Sources of Inertia in Medicare Part D", *NBER Working Paper 22765.*

Ito, Koichiro (2014). "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing", *American Economic Review*, 104(2) 537–563.

Kaiser Family Foundation (2021). "Employer Health Benefits, 2021 Annual Survey",Technical report.

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

Klein, Tobias J., Martin Salm, and Suraj 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.

Kowalski, Amanda (2016). "Censored Quantile Instrumental Variable Estimates of the Price Elasticity of Expenditure on Medical Healthcare", *Journal of Business & Economic Statistics*, 34(1): 107–117.

Kowalski, Amanda E (2015). "Estimating the tradeoff between risk protection and moral hazard with a nonlinear budget set model of health insurance", *International Journal of Industrial Organization*, 43: 122–135.

Lin, Haizhen and Daniel W. Sacks (2019). "Intertemporal substitution in health care demand: Evidence from the RAND Health Insurance Experiment", *Journal of Public Economics*, 175: 29–43.

Nevo, Aviv, John L. Turner, and Jonathan W. Williams (2016). "Usage-Based Pricing and Demand for Residential Broadband", *Econometrica*, 84(2): 411–443.

Newhouse, Joseph P and the Insurance Experiment Group (1993). *Free for All?*, Harvard University Press.

O'Donoghue, Ted and Matthew Rabin (1999). "Doing It Now or Later", *American Economic Review*, 89(1): 103–124, DOI: `http://dx.doi.org/10.1257/aer.89.1.103`.

Pauly, Mark V. (1968). "The Economics of Moral Hazard: Comment", *American Economic Review*, 58(3): 531–537.

Saez, Emmanuel (2010). "Do Taxpayers Bunch at Kink Points?", *American Economic Journal: Economic Policy*, 2(3): 180–212.

Schmidheiny, Kurt and Sebastian Siegloch (2020). "On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization", *CEPR Discussion Paper 13477*.

Simonsen, Marianne, Lars Skipper, Niels Skipper, and Anne Illemann Christensen (2021). "Spot price biases in non-linear health insurance contracts", *Journal of Public Economics*, 203 104508.

Winter, Joachim and Amelie Wuppermann (2019). "Health Insurance Plan Choice and Switching", *Oxford Research Encyclopedia of Economics and Finance*.

# Appendix

# A Microeconomic foundation

## A.1 Optimal choice of deductible and timing moral hazard

The decisions for year 2 are made by backwards induction. First, the individual chooses optimal spot consumption, taking the deductible and timing decisions, and all other parameters as given. First-order conditions imply that $c_t^*(s) = \lambda_t(s)$ if the individual ends the year below the deductible, and $c_t^*(s) = \lambda_t(s) + \omega$ if above. Only total yearly consumption matters for individuals with perfect foresight, and I ignore discounting. Let $\bar{x}(s) \equiv \sum_{t=13}^{24} x_t(s)$. Optimal spot consumption gives rise to the following value function

$$V(D_j, \bar{m}(s)) = \begin{cases} -\bar{\lambda}(s) - \bar{v}_m(m,s) - \bar{m}(s) - \bar{n}_j & \text{if } \bar{\lambda}(s) + \bar{m}(s) < D_j \\ \frac{\bar{\omega}}{2} - D_j - \bar{n}_j & \text{if } \bar{\lambda}(s) + \bar{m}(s) \geq D_j \end{cases} \tag{22}$$

where in the first case, the individual stays below the deductible in terms of total healthcare spending, and exceeds it in the second case.

Second, the individual compares all timing and deductible choices given shock timing. This yields several cases.

**Case 1.**— If nondiscretionary and planned care add up to less than $D_L$, the total utility value under both deductibles can be written as

$$V(D_j, \bar{m}(s)) = -\bar{\lambda}(s) - \bar{v}_m(m,s) - \bar{m}(s) - \bar{n}_j \tag{23}$$

In that case, $V(D_H, \bar{m}(s)) > V(D_L, \bar{m}(s)) \ \forall \bar{m}(s)$ since $\bar{n}_L > \bar{n}_H$. Hence, it is always optimal to choose the higher deductible and prepone.

**Case 2.**— If preponing allows the individual to position themselves above or below both deductibles, the following intuition applies. An individual who can prepone enough to spend below $D_L$ will choose $D_H$, as in the previous case. An individual who chooses a low deductible does not prepone, so as to avoid the cost of retiming. The individual prepones and keeps a high deductible if the following condition is satisfied

$$V(D_H, \bar{m}(s)) \geq V(D_L, \bar{\mu}(s)) \tag{24}$$

$$(\bar{n}_L - \bar{n}_H) + (D_L - \bar{\lambda}(s) - \rho(s)) - \bar{v}_m(m,s) - \frac{\bar{\omega}}{2} \geq 0 \quad \forall s \in S \tag{25}$$

They choose $D_H$ and prepone if the sum of the following terms is positive: the savings in premiums; the out-of-pocket cost difference from preponing; the costs of planned care

still consumed in year 2; the utility cost of retiming; and the opportunity cost in terms of foregone utility from within-year moral hazard.

**Case 3.**— If nondiscretionary care is higher than the high deductible, such that $\bar{\lambda}(s) \geq D_H$, both deductibles would be exceeded regardless of planned care consumption. It is then optimal to choose a low deductible since $D_L + \bar{n}_L < D_H + \bar{n}_H$, and not to pay the cost of retiming.

Hence, in all cases, the individual either chooses to prepone the optimal amount and keep a high deductible, or not to prepone anything and switch to a low deductible. The individual will choose either depending on the indifference condition (24).

By the functional form assumption in (17), shifted amounts are evenly allocated throughout the target year as $\rho(s)$ does not depend on $t$. Note that below the deductible $j$, the first order condition implies that, conditional on preponing, the individual prepones the amount such that the marginal cost of retiming equals the out-of-pocket cost savings. Above the deductible, planned consumption is not shifted as there are no savings to be achieved.

This yields two optimal planned care consumption paths under the two options. *Preponers* advance all planned care to year 1 and keep the high deductible, so that their path of planned spending consumption is

$$m_t^*(s) = \begin{cases} \mu_t & \text{for } t = 1, \ldots, s-1 \\ \mu_t + \rho(s) & \text{for } t = s, \ldots, 12 \\ 0 & \text{for } t = 13, \ldots, 24 \end{cases} \tag{26}$$

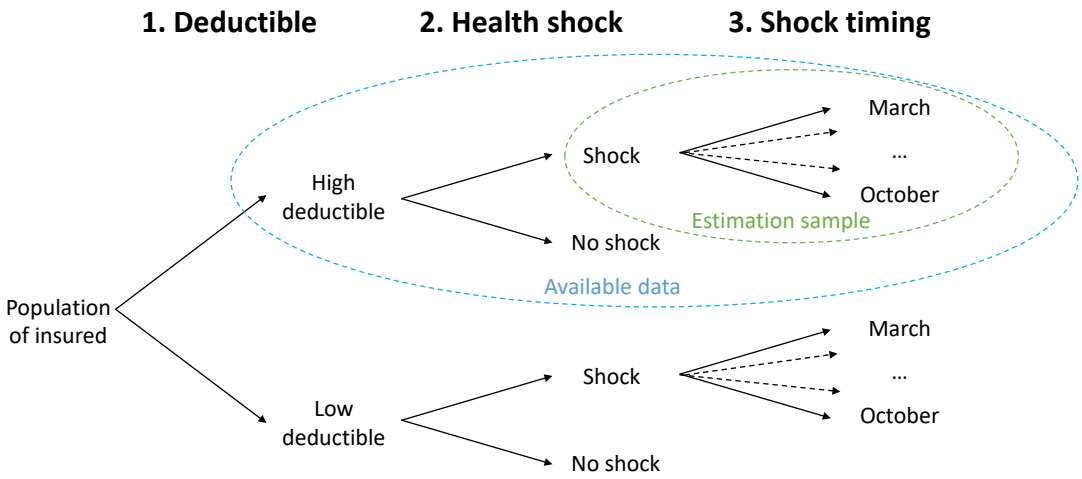
*Switchers* consume all $\mu_t$ as planned, so that their path of planned spending consumption is

$$m_t^*(s) = \mu_t \text{ for } t = 1, \ldots, 24 \tag{27}$$

# B Empirical analysis

## B.1 Sample definition

Figure B.1: Selection levels




*Notes:* The figure is a simplified depiction of the steps of selection into the available sample (blue oval), as well as the estimation sample (green oval). January, February, November and December months are excluded from the analysis to avoid turn-of-the-year confounding effects.

Table B.1: Differences in observable characteristics by shock timing

| | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | |
|---|---|---|---|---|---|---|---|---|---|
| *Demographics* | | | | | | | | | |
| Age | 0.85*** | 0.39 | 0.53** | 0.22 | −0.79*** | −0.28 | −0.67** | −0.86*** | |
| Female | 0.01 | −0.00 | 0.02* | 0.01 | −0.02* | −0.02 | −0.00 | −0.00 | |
| Swiss | −0.01 | 0.00 | 0.00 | −0.00 | −0.00 | 0.01* | −0.01 | 0.01 | |
| *Insurance plan* | | | | | | | | | |
| Premiums | 58.52*** | 10.03 | 7.57 | 18.47 | −21.23 | −35.27** | −18.69 | −54.63*** | |
| CHF 2500 deductible | −0.02*** | −0.02** | −0.01* | −0.00 | 0.02** | 0.03*** | 0.01 | 0.01 | |
| Standard plan | 0.02** | 0.01 | 0.01 | 0.00 | −0.01 | 0.00 | −0.02** | −0.02 | |
| Accident insurance | 0.03*** | 0.00 | 0.01 | −0.01 | −0.01 | −0.00 | −0.01 | −0.03*** | |
| *Spending at shock* | | | | | | | | | |
| Total spending | 4.05 | −59.17 | −11.51 | −19.68 | 137.42** | 11.51 | 7.34 | −50.93 | |
| Share of physician outpatient spending | 0.00 | −0.01*** | 0.01 | 0.01** | −0.01*** | −0.00 | 0.02*** | −0.01* | |
| Share of hospital outpatient spending | −0.00 | 0.00 | 0.01 | 0.01** | −0.00 | −0.01** | 0.00 | −0.01 | |
| Share of hospital inpatient spending | −0.01 | 0.01 | −0.01 | −0.02*** | 0.01* | 0.02*** | −0.02** | 0.02** | |
| Share of drugs spending | 0.00 | 0.00 | 0.00 | 0.00 | −0.00 | −0.00 | −0.00 | 0.00 | |
| Share of other spending | 0.00 | −0.00 | −0.00 | 0.00 | 0.00 | −0.00 | 0.00 | −0.00 | |
| Insured | 4599 | 3724 | 3605 | 3420 | 3018 | 2852 | 3211 | 3153 | 27582 |

*Notes:* The table presents differences between the sample average of each treatment group, and the average pooling all insured in the analysis sample (individuals with high deductibles and a health shock). The significance symbols are based on unadjusted t-tests, $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

## B.2 Allocation of claims over time

Each claim in the data contains a start and an end date, which may include several medical services provided at different times (e.g. initial and follow-up consultations). Over two thirds of the claims span one day, and nearly 90% span no more than a calendar month. For claims spanning longer periods, I do not observe the exact month where care was consumed. Note that all claims end on December 31st because of the annual coverage of the insurance contract. There is a spike in the number of claims closed on December 31st, with a smaller share closing in the very beginning of the year.

I censor duration at 365 days (around 0.2% of all claims). I then allocate spending proportionally to the months spanned by the claim spell using the start and end dates. Furthermore, I split any claims that span more than one calendar year (around 0.5% of all claims) in two, so that the first one ends on December 31st, and the second starts on January 1st, and allocate spending proportionally. For the claims that start in the last year of observation, I similarly split the claim and drop the part allocated to the year after.

### B.3 Cumulated differences in spending

Cumulated difference between shock and reset

$$\Delta\text{ShockReset}(s) = \sum_{m=1}^{13-s} \hat{\gamma}_m^s - \sum_{m=1}^{10} \hat{\gamma}_m^3 \tag{28}$$

Cumulated difference in year 1

$$\Delta\text{Year1}(s) = \sum_{m=2-s}^{13-s} \hat{\gamma}_m^s - \sum_{m=-1}^{10} \hat{\gamma}_m^3 \tag{29}$$

Cumulated difference in year 2

$$\Delta\text{Year2}(s) = \sum_{m=14-s}^{25-s} \hat{\gamma}_m^s - \sum_{m=11}^{22} \hat{\gamma}_m^3 \tag{30}$$

Cumulated difference in both years

$$\Delta\text{BothYears}(s) = \sum_{m=2-s}^{25-s} \hat{\gamma}_m^s - \sum_{m=-1}^{22} \hat{\gamma}_m^3 \tag{31}$$

All cumulated differences computed for $s = 4, \ldots, 10$, relative to the March group.

## B.4 Results by healthcare spending category

Table B.2: Summary statistics for monthly spending differences by spending category

| *Outpatient* | Mean | SD | Min | Max |
|---|---|---|---|---|
| Monthly difference | 9.79 | 57.76 | -178.04 | 160.84 |
| Rescaled monthly difference | 5.03 | 43.72 | -182.51 | 188.62 |
| Significant at 10% | 0.29 | 0.45 | 0.00 | 1.00 |
| Observations | 140 | | | |

| *Inpatient* | Mean | SD | Min | Max |
|---|---|---|---|---|
| Monthly difference | 18.37 | 71.03 | -188.24 | 253.76 |
| Rescaled monthly difference | 7.42 | 53.70 | -198.36 | 309.81 |
| Significant at 10% | 0.17 | 0.38 | 0.00 | 1.00 |
| Observations | 140 | | | |

| *Drugs* | Mean | SD | Min | Max |
|---|---|---|---|---|
| Monthly difference | 2.49 | 18.88 | -60.03 | 56.73 |
| Rescaled monthly difference | 1.44 | 15.03 | -73.98 | 67.49 |
| Significant at 10% | 0.14 | 0.34 | 0.00 | 1.00 |
| Observations | 140 | | | |

*Notes:* The table displays summary statistics for the estimated monthly differences in spending. The raw average difference corresponds to estimates of $\Delta\gamma_k(s)$. The rescaled difference corresponds to $\Delta\tilde{\gamma}_k(s)$, as in equation (11).

Table B.3: Estimates of the timing moral hazard parameters

| | Monthly difference in spending | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| *Panel (A). Outpatient* | | | | |
| Coefficient on $s$ ($2\delta$) | -8.18 | -8.18 | -9.95* | -9.96* |
| | (4.55) | (4.56) | (4.27) | (4.28) |
| Coefficient on $\Delta s$ ($\delta$) | | | -4.66** | -4.68** |
| | | | (1.57) | (1.59) |
| Constant ($\beta$) | 4.74 | 5.03 | 4.74 | 5.03 |
| | (4.33) | (4.35) | (4.34) | (4.36) |
| Observations | 140 | 140 | 140 | 140 |
| *Panel (B). Inpatient* | | | | |
| Coefficient on $s$ ($2\delta$) | -3.91 | -3.83 | -4.30 | -4.19 |
| | (4.18) | (4.21) | (3.73) | (3.77) |
| Coefficient on $\Delta s$ ($\delta$) | | | -1.02 | -0.96 |
| | | | (2.44) | (2.48) |
| Constant ($\beta$) | 7.24 | 7.42 | 7.24 | 7.42 |
| | (7.83) | (7.91) | (7.86) | (7.94) |
| Observations | 140 | 140 | 140 | 140 |
| *Panel (C). Drugs* | | | | |
| Coefficient on $s$ ($2\delta$) | -2.03 | -2.04 | -2.55 | -2.56 |
| | (1.87) | (1.86) | (1.64) | (1.64) |
| Coefficient on $\Delta s$ ($\delta$) | | | -1.37 | -1.37 |
| | | | (0.93) | (0.93) |
| Constant ($\beta$) | 1.37 | 1.44 | 1.37 | 1.44 |
| | (2.40) | (2.40) | (2.41) | (2.41) |
| Observations | 140 | 140 | 140 | 140 |
| Adjusted | No | Yes | No | Yes |

*Notes:* The table displays coefficient estimates from a regression of monthly differences in spending $\Delta\tilde{\gamma}_k(s)$ on shock timing, as in (20), where regressors are demeaned. Adjusted regressions in columns (2) and (4) use differences from a covariate-adjusted event study. Confidence intervals at the 5% level based on bootstrapped standard errors with 49 replications, clustered at the individual level.

## B.5 Results for alternative samples

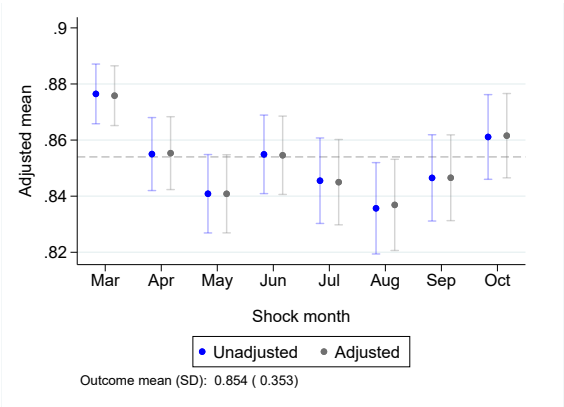
Table B.4: Summary statistics for alternative samples and shock definitions

| | (1) High deductibles with health shock | (2) Cumulated spending below 1/2 of deductible | (3) First time exceeded deductible | (4) CHF 2,500 deductible only | (5) Shock of CHF 1,500 definition |
|---|---|---|---|---|---|
| *Demographics* | | | | | |
| Age | 50.58 (14.82) | 49.98 (14.75) | 51.57 (15.12) | 51.03 (14.52) | 49.03 (14.55) |
| Female | 0.47 | 0.46 | 0.54 | 0.45 | 0.49 |
| Swiss | 0.87 | 0.86 | 0.86 | 0.88 | 0.86 |
| *Insurance plan* | | | | | |
| Premiums | 2,759 (753) | 2,724 (744) | 2,844 (771) | 2,546 (653) | 2,754 (753) |
| CHF 2500 deductible | 0.29 | 0.31 | 0.28 | 1.00 | 0.23 |
| Standard plan | 0.52 | 0.51 | 0.54 | 0.55 | 0.53 |
| Other plan type | 0.48 | 0.49 | 0.46 | 0.45 | 0.47 |
| Accident insurance | 0.46 | 0.45 | 0.49 | 0.47 | 0.43 |
| *Spending* | | | | | |
| Total out-of-pocket spending | 5,030 (818) | 5,002 (816) | 5,102 (821) | 5,495 (708) | 4,873 (824) |
| Total annual spending | 10,262 (10,920) | 10,172 (11,386) | 9,801 (9,616) | 10,839 (11,545) | 8,185 (9,470) |
| *Prices and deductible switching* | | | | | |
| Exceeded deductible | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Cost-sharing | 0.34 | 0.35 | 0.33 | 0.42 | 0.40 |
| Year-end price in shock year | 0.07 (0.05) | 0.07 (0.05) | 0.07 (0.05) | 0.07 (0.05) | 0.06 (0.05) |
| Year-end price in post-shock year | 0.55 (0.46) | 0.57 (0.46) | 0.51 (0.46) | 0.61 (0.46) | 0.54 (0.46) |
| Hit deductible in January of year 2 | 0.11 (0.31) | 0.10 (0.31) | 0.09 (0.29) | 0.11 (0.31) | 0.07 (0.26) |
| Month hit deductible in year 2 | 5.01 (3.51) | 4.96 (3.52) | 5.33 (3.49) | 4.77 (3.58) | 4.86 (3.49) |
| Insured | 27582 | 19834 | 12566 | 8055 | 39139 |

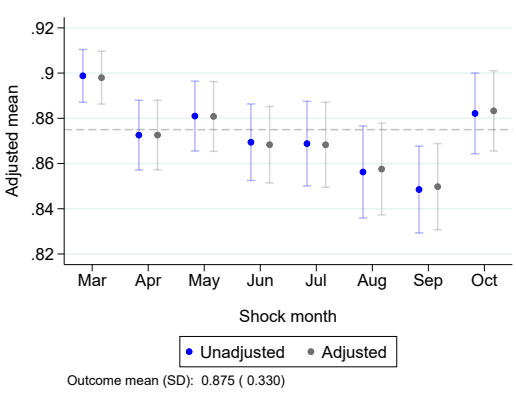
*Notes:* The table presents means and standard deviations (in parentheses) for samples of insured-years. High deductibles (column 1) are the main analysis sample—insured-year observations with annual deductibles of CHF 1,000 to 2,500 and the main shock definition (i.e. monthly spending above CHF 2,500 for the first time in the observation window). Column (2) additionally imposes that cumulated yearly spending was below half of the deductible before the shock occurred. Column (3) restricts to insured with the highest deductible only. Column (4) imposes defined the shock based on inpatient spending. Column (5) sets the shock definition at CHF 1,500. Cost-sharing is calculated as out-of-pocket spending (net of premiums) over total yearly healthcare spending. Total out-of-pocket spending includes insurance premiums.

Figure B.2: Adjusted share of individuals keeping a high deductible (preponers), alternative samples

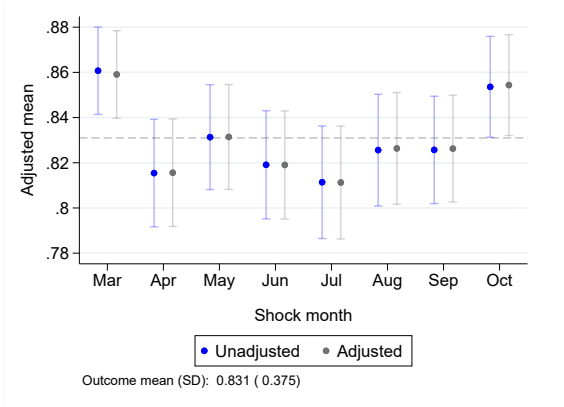
(a) Shock CHF 2,500, Cumumulated spending below half of deductible



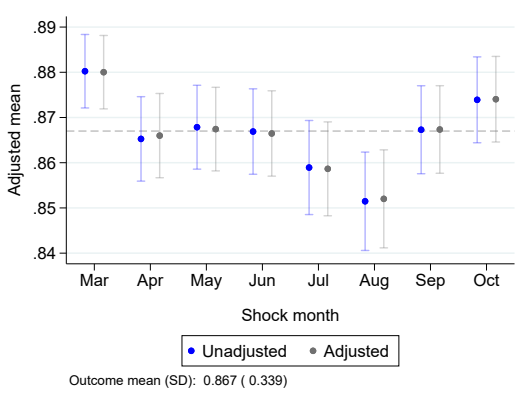

(b) Shock CHF 2,500, first time exceeded deductible




(c) Shock CHF 2,500, CHF 2,500 deductible



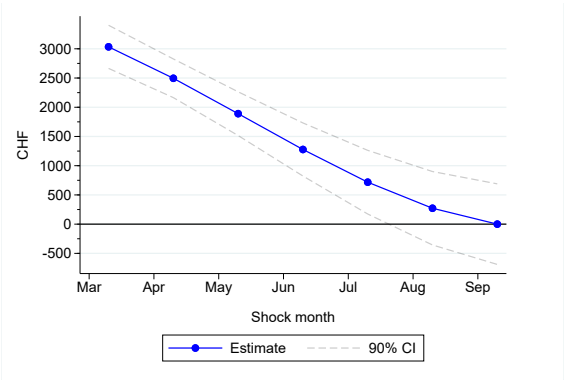

(d) Shock CHF 1,500



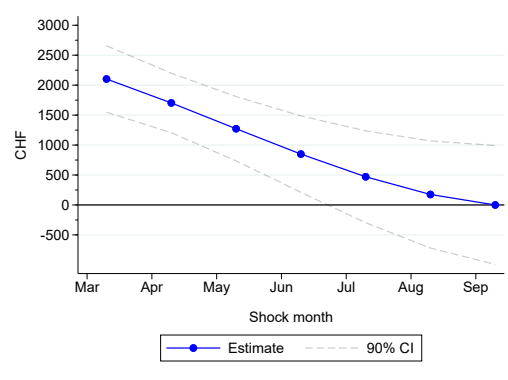

*Notes:* The figure displays the adjusted share of preponers $q(s)$, as identified by individuals who keep a high deductible in the year after the shock, and so across shock months. The blue dots are raw averages, the grey are adjusted at the mean for demographic characteristics (gender, age, nationality, canton of residence) and time fixed effects. The dotted horizontal line denotes the sample average. Confidence intervals are at the 95% level, based on robust standard errors.

Figure B.3: Predicted timing moral hazard by shock month – Alternative samples and shock definitions

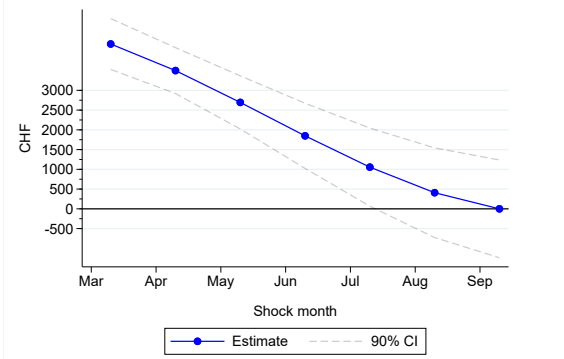
(a) Shock CHF 2,500, Cumumulated spending below half of deductible



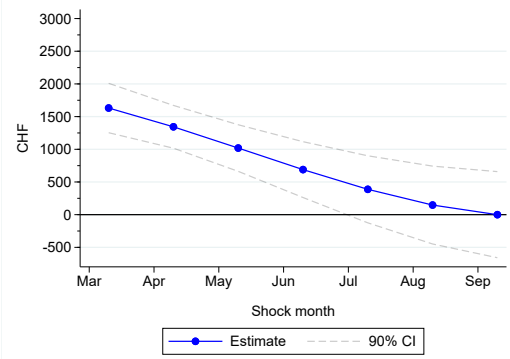

(b) Shock CHF 2,500, first time exceeded deductible




(c) Shock CHF 2,500, CHF 2,500 deductible




(d) Shock CHF 1,500




*Notes:* The figure presents estimates of the total yearly timing moral hazard response across shock months, predicted as in (21). The last shock month serves as a lower bound. Confidence intervals at the 5% level based on bootstrapped standard errors with 49 replications, clustered at the individual level.

Table B.5: Estimates of yearly timing moral hazard responses

| | Total care spending | Outpatient care spending | Inpatient care spending | Drugs spending |
|---|---|---|---|---|
| *Panel (A). Shock CHF 2,500, Cumumulated spending below half of deductible* | | | | |
| Lower bound $\bar{\alpha}$ | 1857.36 | 1245.67 | 320.83 | 241.99 |
| | (203.93) | (167.57) | (190.38) | (68.11) |
| March prediction | 3033.68 | 2067.14 | 448.99 | 398.89 |
| | (104.86) | (76.52) | (65.41) | (32.73) |
| September prediction | 0.00 | 0.00 | 0.00 | 0.00 |
| | (203.93) | (167.57) | (190.38) | (68.11) |
| Mean prediction | 1383.80 | 924.26 | 247.81 | 179.86 |
| | (116.39) | (93.42) | (101.15) | (39.24) |
| *Panel (B). Shock CHF 2,500, first time exceeded deductible* | | | | |
| Lower bound $\bar{\alpha}$ | 1265.90 | 1046.32 | 93.45 | 209.83 |
| | (290.45) | (236.78) | (168.74) | (41.21) |
| March prediction | 2102.66 | 1768.02 | 122.63 | 351.37 |
| | (80.33) | (66.68) | (32.58) | (23.97) |
| September prediction | 0.00 | 0.00 | 0.00 | 0.00 |
| | (290.45) | (236.78) | (168.74) | (41.21) |
| Mean prediction | 939.04 | 772.63 | 75.01 | 155.32 |
| | (148.88) | (121.60) | (97.49) | (23.58) |
| *Panel (C). Shock CHF 2,500, CHF 2,500 deductible* | | | | |
| Lower bound $\bar{\alpha}$ | 1500.77 | 511.94 | 749.46 | 79.95 |
| | (263.97) | (157.82) | (211.83) | (46.11) |
| March prediction | 2444.05 | 823.97 | 1186.80 | 188.62 |
| | (130.38) | (84.88) | (92.58) | (70.85) |
| September prediction | 0.00 | 0.00 | 0.00 | 0.00 |
| | (263.97) | (157.82) | (211.83) | (46.11) |
| Mean prediction | 1118.98 | 382.84 | 562.75 | 108.96 |
| | (147.82) | (89.35) | (117.74) | (41.73) |
| *Panel (D). Shock CHF 1,500* | | | | |
| Lower bound $\bar{\alpha}$ | 701.09 | 512.90 | 46.03 | 229.00 |
| | (127.52) | (118.99) | (56.16) | (28.77) |
| March prediction | 1166.76 | 845.65 | 90.85 | 387.68 |
| | (62.81) | (40.07) | (135.64) | (13.20) |
| September prediction | 0.00 | 0.00 | 0.00 | 0.00 |
| | (127.52) | (118.99) | (56.16) | (28.77) |
| Mean prediction | 519.80 | 381.20 | 59.53 | 169.01 |
| | (71.04) | (62.72) | (73.06) | (16.31) |

*Notes:* The table presents a summary of the total yearly timing moral hazard response, predicted as in (21) using estimates from column (3) of Table 4. The last shock month serves as a lower bound. Standard errors in parentheses based on bootstrapped standard errors with 49 replications, clustered at the individual level.

## B.6 Prices and deductible switching

Table B.6: Summary statistics for prices, plan switching and spending by shock timing

| | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Pooled |
|---|---|---|---|---|---|---|---|---|---|
| *Prices* | | | | | | | | | |
| Price at beginning of month | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Price at end of month | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 |
| Year-end price in shock year | 0.06 | 0.06 | 0.06 | 0.07 | 0.07 | 0.07 | 0.07 | 0.08 | 0.07 |
| Hit deductible in January of year 2 | 0.08 | 0.10 | 0.10 | 0.10 | 0.12 | 0.11 | 0.13 | 0.14 | 0.11 |
| Year-end price in post-shock year | 0.55 | 0.53 | 0.55 | 0.55 | 0.55 | 0.56 | 0.54 | 0.56 | 0.55 |
| Hit deductible by June of year 2 | 0.32 | 0.34 | 0.33 | 0.31 | 0.32 | 0.32 | 0.34 | 0.34 | 0.33 |
| Month hit deductible in year 2 | 5.26 | 5.14 | 5.02 | 5.23 | 4.90 | 4.97 | 4.83 | 4.56 | 5.01 |
| *Plan switching* | | | | | | | | | |
| Switched to low deductible in year 2 | 0.12 | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 | 0.14 | 0.14 |
| Switched to lower deductible in year 2 | 0.14 | 0.17 | 0.17 | 0.17 | 0.17 | 0.17 | 0.17 | 0.15 | 0.16 |
| Standard plan in year 2 | 0.50 | 0.50 | 0.49 | 0.48 | 0.49 | 0.48 | 0.46 | 0.47 | 0.48 |
| *Spending* | | | | | | | | | |
| Total spending in year of shock | 11,397.29 | 11,474.75 | 11,103.70 | 10,379.48 | 10,435.78 | 9,432.26 | 9,001.25 | 7,951.68 | 10,261.96 |
| Total spending in year after shock | 4,292.66 | 4,530.65 | 4,358.37 | 4,273.27 | 4,860.71 | 4,627.15 | 5,138.61 | 5,136.07 | 4,622.61 |
| Total spending in year before shock | 1,212.26 | 1,008.85 | 916.51 | 840.31 | 759.23 | 732.44 | 756.85 | 707.82 | 890.16 |
| Insured | 4599 | 3724 | 3605 | 3420 | 3018 | 2852 | 3211 | 3153 | 27582 |

*Notes:* The table presents insured-level sample means by calendar month of the shock (treatment group). The sample is insured with high deductibles with a shock. All spending in Swiss Francs (CHF).

Table B.7: Regressions of prices on shock timing

| Shock month | (1) Year-end price in shock year | (2) Year-end price in post-shock year | (3) Hit chosen deductible in January in post-shock year | (4) Month hit chosen deductible (if do) |
|---|---|---|---|---|
| March (reference) | – | – | – | – |
| April | 0.001 | −0.019 | 0.018** | −0.130 |
| | (0.001) | (0.010) | (0.006) | (0.108) |
| May | 0.001 | −0.001 | 0.022*** | −0.239* |
| | (0.001) | (0.010) | (0.006) | (0.110) |
| June | 0.005*** | 0.003 | 0.021** | −0.030 |
| | (0.001) | (0.010) | (0.006) | (0.114) |
| July | 0.004*** | −0.001 | 0.040*** | −0.352** |
| | (0.001) | (0.011) | (0.007) | (0.118) |
| August | 0.008*** | 0.004 | 0.029*** | −0.279* |
| | (0.001) | (0.011) | (0.007) | (0.120) |
| September | 0.008*** | −0.013 | 0.052*** | −0.410*** |
| | (0.001) | (0.011) | (0.007) | (0.116) |
| October | 0.012*** | 0.002 | 0.060*** | −0.689*** |
| | (0.001) | (0.011) | (0.007) | (0.117) |
| Mean dep. var. | 0.067 | 0.548 | 0.105 | 5.010 |
| Insured | 27582 | 27582 | 27582 | 13467 |

*Notes:* The table displays coefficient estimates from linear regressions at the insured level. Robust standard errors in parentheses. All models include age-gender group dummies (in 10-year bands), insurance plan type, and year and region dummies. All prices are in Swiss Francs (CHF). $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B.8: Regressions of insurance plan choice in post-shock year on shock timing - Main sample

| Shock month | (1) Kept a high deductible | (2) Standard plan |
|---|---|---|
| March (reference) | – | – |
| April | −0.024** | −0.002 |
| | (0.008) | (0.005) |
| May | −0.031*** | −0.003 |
| | (0.008) | (0.005) |
| June | −0.023** | −0.008 |
| | (0.008) | (0.005) |
| July | −0.031*** | 0.003 |
| | (0.008) | (0.005) |
| August | −0.029*** | −0.004 |
| | (0.008) | (0.005) |
| September | −0.027*** | −0.007 |
| | (0.008) | (0.005) |
| October | −0.014 | −0.006 |
| | (0.008) | (0.005) |
| Mean dep. var. | 0.856 | 0.485 |
| Insured | 27582 | 27582 |

*Notes:* The table displays coefficient estimates from linear regressions at the insured level. Robust standard errors in parentheses. All models include age-gender group dummies (in 10-year bands), insurance plan type, and year and region dummies. All prices are in Swiss Francs (CHF). $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B.9: Attrition – Main sample

| Shock month | (1) Dropped out of the sample before 2016 | (2) Total years observed |
|---|---|---|
| March (reference) | – | – |
| April | −0.012 | 0.011 |
| | (0.008) | (0.045) |
| May | −0.008 | 0.054 |
| | (0.008) | (0.045) |
| June | −0.001 | 0.023 |
| | (0.008) | (0.046) |
| July | 0.003 | −0.101* |
| | (0.008) | (0.049) |
| August | 0.001 | 0.010 |
| | (0.008) | (0.049) |
| September | −0.010 | 0.078 |
| | (0.008) | (0.046) |
| October | 0.012 | −0.102* |
| | (0.008) | (0.049) |
| Mean dep. var. | 0.166 | 10.578 |
| Insured | 27582 | 27582 |

*Notes:* The table displays coefficient estimates from linear regressions at the insured level. Robust standard errors in parentheses. All models include age-gender group dummies (in 10-year bands), insurance plan type, and year and region dummies. All prices are in Swiss Francs (CHF). $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.