Timing Moral Hazard under Deductibles in Health Insurance[*]

Véra Zabrodina

Abstract

This paper studies strategic timing behavior in healthcare consumption under insurance contracts with deductibles. I develop a dynamic model that allows individuals to shift the timing of planned care, and incorporates both moral hazard and deductible choice. I show that timing decisions involve a trade-off between out-of-pocket cost savings below the deductible, and utility gains from additional consumption after exceeding the deductible. The timing of risk realizations within a coverage period affects both when and how much healthcare is consumed. The timing and intensive margins—as well as their interaction—have distinct implications for healthcare costs, premiums, and selection. I derive a sufficient statistic for timing moral hazard that compares individuals who exceed their deductible at different times within the coverage period, facing varying price incentives. In the context of mandatory health insurance in Switzerland, I find evidence of substantial timing moral hazard after a temporary health shock, which varies with the month of the shock. Timing is a key component of price responses under deductibles, with implications for insurance design.

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

September 16, 2024

[*] Zabrodina: University of California, Berkeley and University of Basel; zabrodina@berkeley.edu. I am grateful to Janet Currie, Michael Gerfin, Tobias Klein, Karine Lamiraud, Helge Liebert, Owen O'Donnell, Nicola Pavoni, Martin Salm, Bettina Siflinger, Ulrike Unterhofer, Alexandros Theloudis, Richard van Kleef, Andrea Weber, Conny Wunsch and Nicolas Ziebarth, as well as seminar participants at the University of Basel, EALE SOLE, Erasmus University Rotterdam, ESPE, ESWC, EuHEA Seminar Series, European Workshop on Econometrics and Health Economics, Graduate Institute Geneva, Georgetown University, IRDES, University of Manchester, Ski and Labour Seminar, Swiss Health Economics Association, Tilburg University, Verein für Socialpolitik Annual Conference, and Young Swiss Economists Meeting for helpful discussions and comments. I thank Christian Schmid and Nicolas Schreiner from the CSS Institute for Empirical Health Economics for access to and help with the data. All errors are my own.

1 Introduction

Asymmetric information about the timing of risk realizations has significant implications for insurance markets. In healthcare, while some medical procedures are urgent, many are not. Under nonlinear cost-sharing, this can create a time lapse between the realization of a health risk and its actual treatment. For instance, under a deductible, individuals might strategically delay or advance their healthcare consumption to years in which they expect to meet their deductible and no longer have to pay out of pocket. Insurers cannot observe or contract on the date the health risk materializes; they can only reimburse based on the date of treatment. Because of this, they are exposed to a specific type of moral hazard, affecting when individuals address materialized health risks.

Timing moral hazard has distinct implications compared to the *classical* moral hazard that has been extensively studied in the literature. It refers to the shifting of planned healthcare consumption across coverage periods without changing the total amount consumed. In contrast, classical moral hazard leads to *additional* consumption that wouldn't occur if the individual faced a higher marginal price during the coverage period.[1] Timing moral hazard influences costs and premiums by covering care that would have been consumed anyway, whereas classical moral hazard generates additional consumption. An example of timing moral hazard is getting a check-up this year instead of next, while classical moral hazard involves getting an additional check-up because the deductible has been met. This distinction is crucial for accurately measuring and interpreting the price elasticities of healthcare consumption.

Both behaviors, along with their interaction, can lead to selection. Synchronizing healthcare consumption with coverage purchase undermines the functioning of insurance markets by reducing the uncertainty about future consumption and generating selection based on planned care consumption (Cabral, 2017). Understanding the dynamics of timing moral hazard and its interplay with other forms of asymmetric information is crucial for addressing the insurance-incentive trade-offs in health insurance.

[1] A note on terminology: Cabral (2017) introduced the concept of 'timing moral hazard,' which I use alongside 'classical moral hazard' to differentiate between moral hazard in timing and quantity, reflecting the focus on the latter in the health insurance literature. I follow the conventional (ab)use of terminology, treating both behaviors as forms of *ex post* moral hazard, excluding any feedback effects on health. Since the insured's actions (consumption) are observable, the information asymmetry arises from the insured's private knowledge of their own propensity for moral hazard, health risk, and planned procedures. See Einav et al. (2013) for a discussion about this terminology. In the context of healthcare, consumption (or demand) is measured one for one by spending, so I use these terms interchangeably (Kowalski, 2015).

This paper presents a framework to identify timing moral hazard in health insurance with deductibles. I develop a model which rationalizes the incentives for shifting healthcare consumption across coverage periods. I also emphasize how this behavior differs from and interacts with classical moral hazard. I then derive and implement a sufficient statistic which identifies pure timing moral hazard net of other types of asymmetric information and selection. I estimate this moment in reduced form by comparing the healthcare consumption of similar individuals who experience shocks at random times within the coverage year and face different dynamic price incentives.

The literature has suggested that classical moral hazard can be limited by offering contracts with nonlinear cost-sharing (Einav and Finkelstein, 2018). Motivated by this insight and growing healthcare spending, health insurance contracts with deductibles have become widespread. The insured cover their healthcare consumption out of pocket up to the deductible amount, above which any additional consumption is free. Costs reset at the end of the coverage period (typically a year). Deductibles are used in, e.g. mandatory health insurance in Switzerland—the setting for this study—and the Netherlands, as well as in both private and public health insurance markets in the United States. There, among covered employees, 58% have a deductible higher than USD 1,000 (Kaiser Family Foundation, 2021). However, little attention has been given to the salient timing incentives created by deductibles through their simple kinked price structure. Decisions about the amount and timing of consumption depend on expectations about future consumption within and across years, which evolve dynamically.

I formulate a dynamic model of healthcare consumption to uncover the factors driving these decisions. A rational, forward-looking individual chooses their monthly healthcare consumption based on realized health shocks and the probability of exceeding the deductible. They can also time a fixed amount of planned healthcare consumption, and choose a yearly deductible. The model sheds light on how timing moral hazard interacts with other sources of asymmetric information, and resulting efficiency trade-offs.

In the literature, the key incentive for timing is reducing out-of-pocket costs. However, I show that timing moral hazard affects classical moral hazard by influencing the probability of exceeding the deductible. For instance, an individual who has exceeded their deductible can advance care to the current year. While they shift out-of-pocket costs of planned care onto the risk pool, they also consume less additional care in the following year due to the higher marginal price. In this case, timing creates an opportunity cost if individuals value the additional consumption from classical moral hazard. Conversely, individuals can also use timing to amplify classical moral hazard by bringing themselves closer to the deductible. This trade-off between out-of-pocket costs and the utility derived from classical moral hazard creates ambiguous effects on overall costs in the risk pool. In sum, the timing matters for the extent of insurance use.

Timing moral hazard influences coverage choices by affecting healthcare spending within a given year. Advancing care can lead to higher deductible choice in the following year, as it lowers expected spending. Delaying can be joint with a lowering the deductible. As a result, timing moral hazard acts as a key driver of *ex post* selection. This is based on private information about planned consumption, i.e. *realized* rather than expected risk, but also about the propensity for classical moral hazard. Even *ex ante* homogeneous individuals may become heterogeneous due to differences in when risks realize, when they are addressed, but also via their effect on expected classical moral hazard. This result generalizes the one in Cabral (2017), where delaying care generates *ex post* adverse selection on top of classical adverse selection on *ex ante* risk.

The model also provides a new sufficient statistic for timing moral hazard that differences out classical moral hazard. My test focuses on individuals who suffer a large, unanticipated health shock which pushes them above the deductible. After the shock, individuals may have an incentive to advance care planned for next year to the current year to reduce out-of-pocket costs and keep a high deductible the year after. On the other hand, they may also decide to delay care, choose a lower deductible the next year, and benefit from the additional consumption from classical moral hazard. If shock timing is random, it exogenously varies these timing incentives. Individuals who suffer a shock early in the calendar year face a zero price for a longer period until the year-end deductible reset than those with a shock late in the year. The later the shock, the more likely it persists into the year after. Later shocks thus weaken the incentive and the time available to advance care. The differences in healthcare spending across individuals with shocks at different times within the calendar year measure timing moral hazard. This is because individuals are otherwise comparable, and shock-related consumption and classical moral hazard are differenced out after the shock.

This theoretical result motivates a novel reduced-form identification strategy which relies on random shock timing. Specifically, I run an event study of healthcare spending, where treatment groups are defined by the calendar month of the first hospitalization. I implement this strategy in the context of health insurance in Switzerland. I use individual-level claims data from the largest private health insurance provider for the years 2012 to 2019. These data are representative of the Swiss population. Switzerland offers an attractive setting for studying strategic timing. Health insurance is mandatory, and contracts are highly regulated and cover a broad range of medical procedures. They bear one deductible, which the insured can choose yearly without risk classification. The options lie between CHF 300 and 2,500 (CHF 1 $\approx$ USD 1). To circumvent selection at baseline and increase the plausibility of shock timing being random, my test uses individuals with the highest deductible, a plausibly-unanticipated shock, and potentially strong timing incentives.

I document substantial timing moral hazard. Individuals with shocks in February consume nearly CHF 1,000 more of planned care than individuals with shocks later than June. Consistent with a stronger incentive to delay, individuals with shocks later than June consume essentially no planned care in the shock year. This difference represents approximately 10% of February's consumption in the year of the shock, and a price-elasticity of -0.14 which is close to the benchmark for classical moral hazard of -0.2 (Keeler and Rolph, 1988). These results suggest that existing price-elasticity estimates under nonlinear contracts may overstate classical moral hazard because they include timing responses.

I do not find heterogeneity in deductible choice in the year after the shock depending on shock timing. This result suggests that timing generates little *ex post* selection in my setup. It also points to a role for dynamic frictions related to the time available in driving heterogeneity in the amount of timing. Frictions may stem from supply-side constraints (e.g., appointment scheduling, referral requirements), hassle costs, or behavioral and cognitive biases (e.g., incorrect expectations about future health needs).

Categories of care differ in their amenability to retiming, with outpatient and drugs representing an overproportional share of the timing response. I find smaller timing responses when shocks are persistent, very urgent, and occur under plans with restricted provider choice—all situations where the ability to retime is plausibly limited. Taken together, the results point to coverage length, deductibles split by type of procedures or illness episode, and provider choice restrictions as relevant policy tools to address timing moral hazard.

Related literature. This paper adds to 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 (classical) moral hazard.[2] These studies have adopted a static perspective by assuming that individuals only consider current or year-end prices, and have implicitly ruled out strategic timing across years. Meanwhile, empirical papers have found evidence suggestive of timing with, e.g., individuals anticipating deductible resets, or increasing consumption after becoming eligible for coverage (Manning et al., 1987; Gerfin et al., 2015; Simonsen et al., 2021; Card et al., 2009).[3] In contrast to these studies, I tease out the part of the dynamic response which is due to pure timing moral hazard.

Several papers explicitly test for strategic timing in healthcare consumption. 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). The authors find no such responses among those who spend largely past the gap and have weaker incentives to shift. Similar to mine, their results highlight that failing to account for timing responses may overestimate the classical moral hazard response. Lin and Sacks (2019) develop a test for short-term intertemporal substitution using data from the RAND experiment. The authors find that individuals with deductibles have higher spending than those in free care plans in the last month of the coverage year. This suggests that those under the deductible 'stock up' on health. Most closely related to my paper is Cabral (2017), who studies the strategic delay of dental care under contracts with maximum benefits.[4] Using a structural modeling approach, the author finds that about 40% of individuals postpone dental care when incentivized to do so. The resulting *ex post* adverse selection, whereby individuals delay care to purchase coverage, explains the largely-missing insurance market for dental care, which is easily deferrable.

This paper contributes to the literature by introducing a novel framework and empirical strategy for measuring timing moral hazard. First, my theoretical model explicitly incorporates multidimensional asymmetric information. It provides insights into how

[2] See Finkelstein (2014), Einav and Finkelstein (2018), and Gerfin (2019) for reviews and discussions of the literature on moral hazard in health insurance. Examples include Kowalski (2016), which 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) instruments individual prices with employer-year-plan-month average cost shares to estimate price elasticities by type of medical service on a monthly basis.

[3] Simonsen et al. (2021) finds 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.

[4] Maximum benefits are the opposite of deductibles, in that they cover spending up to a given yearly amount. The incentive is to shift care towards years where the spending is below the maximum benefit.

timing moral hazard endogenously shapes classical moral hazard and deductible choice. So far, the literature on the interactions between multiple sources of asymmetric information has remained sparse (Einav et al., 2015; Cabral, 2017; Hendren et al., 2021), despite the importance these interactions for the functioning of insurance markets. My framework illustrates how these interactions also complicate the separate identification of timing moral hazard.

My second contribution is in a novel identification strategy that overcomes this challenge and quantifies pure timing moral hazard. In a literature that has mainly adopted fully-structural methods (Einav et al., 2015; Cabral, 2017), my sufficient statistics approach provides a characterization of the quantity estimated in reduced form, with a clear source of variation. It imposes fewer conceptual and distributional assumptions, as it does not require specifying all the primitives underlying healthcare consumption and timing decisions. Third, my empirical analysis covers a broad range of medical services in a setting with mandatory health insurance, where timing responses have not been quantified so far. Previous studies have had a narrower focus (e.g. Medicare for the elderly Card et al. 2009; Einav et al. 2015; or dental care for employees of a firm Cabral 2017).

This paper adds to our understanding of the high-frequency dynamics of healthcare consumption both within and across years. My findings suggest that individuals are not fully myopic as they anticipate both non-emergent health needs and prices. This insight adds to the debate on sophistication and price perceptions in healthcare consumption decisions (Aron-Dine et al., 2015; Brot-Goldberg et al., 2017; Abaluck et al., 2018; Dalton et al., 2020; Klein et al., 2022). My approach also uncovers frictions to timing in healthcare consumption that relate to the time horizon available. Meanwhile, the literature on frictions in health insurance has focused on plan choices and their implications for adverse selection (see e.g. Hendren et al. 2021).

My findings contribute to the growing body of evidence that individuals use insurance not only to transfer resources across different risk states but also over time (Ericson and Sydnor, 2018; Gross et al., 2022; Hong and Mommaerts, 2024). The results demonstrate that price nonlinearities facilitate such behavior. My study adds to the expanding literature in public finance that examines how dynamic asymmetric information drives strategic timing behavior and its efficiency consequences for insurance and taxation systems. Examples include the strategic timing of labor supply to qualify for more generous unemployment insurance (Albanese et al., 2020; Citino et al., 2022; Brébion et al., 2022), or to reduce income taxes (Martínez et al., 2021).

The paper proceeds as follows. The next section outlines relevant institutional features of the Swiss health insurance system. Section 3 presents the theoretical model and derives the sufficient statistic for timing moral hazard. Section 4 elaborates on the data and reduced-form estimation. Section 5 presents the main results, as well as robustness checks and heterogeneity analyses. Section 6 further explores the microfoundations of timing moral hazard and dynamic frictions. Section 7 discusses implications for insurance markets. The final section concludes.

2 Institutional Setting

Privately-provided mandatory health insurance. The health insurance market in Switzerland offers a compelling setting to analyze strategic timing behavior.[5] Each resident is required by law to individually enroll with a private health insurance company, which is forbidden from denying coverage or selecting on risk. Premiums are community-rated by region of residence and broad groups of age, where the last bracket starts at age 26. These regulations prevent insurers from underwriting strategically timed claims, and offering contracts with coinsurance (i.e. no timing incentives). All mandatory health insurance contracts bear a deductible chosen every year between CHF 300 (the default), 500, 1,000, 1,500, 2,000 and 2,500—a rather small choice set. The financial stakes of deductible choice and timing are high and salient, since insurance premiums and healthcare costs represent a growing share of household budgets. Health insurance premiums amounted to nearly 9 percent of household disposable income in 2020. The market structure shares many similarities to the US, besides the universal coverage. The setting closest to Switzerland is the Netherlands, where health insurance is also mandatory and the choice of deductible lies between EUR 385 and 885.

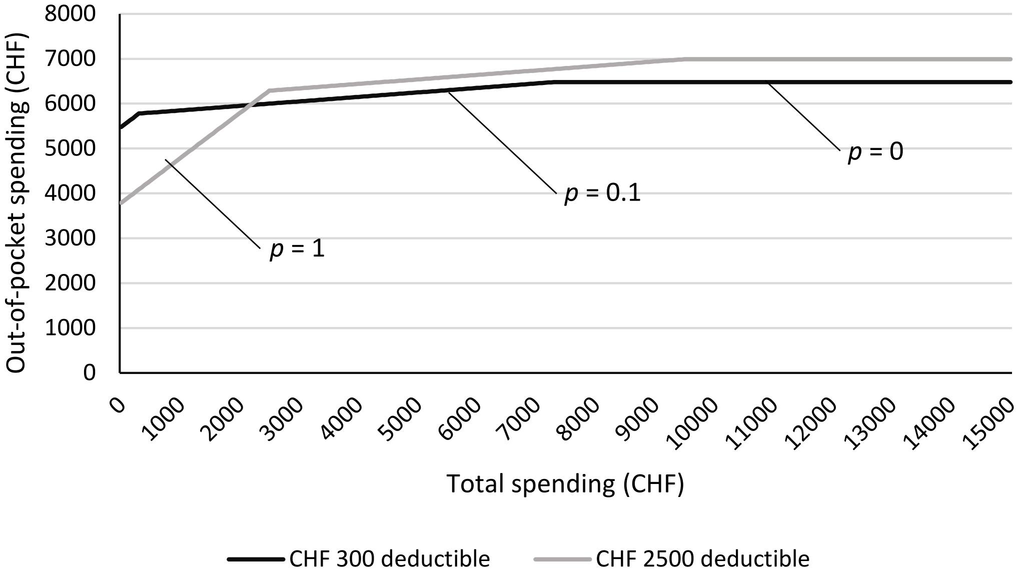
Contracts with deductibles. The insured pay for any covered healthcare consumption out of pocket up to the deductible. Above that, a co-payment rate of 10% applies up to a stop-loss of CHF 700. This cost-sharing schedule generates non-convexities in the individual's budget constraint at the deductible and stop-loss. The marginal price drops from 1 to 0.9 when exceeding the deductible, and to 0 after the stop-loss is reached. The total annual out-of-pocket spending under deductible D_j can be written as a function of total annual healthcare spending H as follows:

$$OOP_j(H) = 12n_j + \min\{D_j, H\} + \max\{0, \min\{0.1(H - D_j), 700\}\} \tag{1}$$

where n_j are monthly premiums which decrease with D_j, and depend on the charac-

[5] The laws regulating mandatory health insurance have remained broadly unchanged since 2005.

Figure 1: Out-of-pocket healthcare costs as a function of total yearly healthcare consumption



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. Average insurance premiums for standard plans from 2019 determine the intercepts at CHF 5,480 and 3,790, respectively. The insured face a marginal price of $p = 1$ up until the deductible, then $p = 0.1$ up until the stop-loss of CHF 700, and $p = 0$ above.

teristics of the insurance plan. Apart from the deductible, individuals choose between a standard plan, which offers free choice of authorized healthcare provider, and alternative plans, which restrict this choice within health maintenance organizations, or require referrals for specialist care (i.e., gatekeeping family physician, or telemedicine), but come with lower premiums. Figure 1 sketches the out-of-pocket cost function for the lowest CHF 300 and highest CHF 2,500 deductibles. The maximum annual out-of-pocket costs net of premiums is CHF 1,000 for the lowest deductible and CHF 3,200 for the highest deductible. The lowest deductible dominates the highest deductible in terms of out-of-pocket costs for healthcare consumption above CHF 2,200, after accounting for the difference in premiums. For simplicity, I ignore the co-payment in what follows. Out-of-pocket contributions reset at the end of each calendar year. This generates a discontinuous increase in the current price at the year-end reset for individuals who had exceeded their deductible. This setup incentivizes individuals to time their consumption strategically in years where they anticipate exceeding the deductible. I study these incentives formally in the next section.

The insured can switch plans or insurers until November 30th, with the new contract taking effect on January 1st.[6] Individuals can freely choose any insurance plan from insurers operating within their canton (region) of residence without risk classification. Cantons provide means-tested insurance premium subsidies for low-income households.

Covered medical services. Mandatory health insurance covers a wide range of ambulatory and inpatient services for illnesses, as well as prescribed drugs, all at fixed prices and under a single deductible. This allows studying timing moral hazard across different types of care. Ambulatory services are reimbursed through a fee-for-service system, while hospitalizations are reimbursed through a prospective payment system with diagnosis-related groups (DRGs). It does not cover accidents, dental care, accidents or care reimbursed under supplementary insurance. Insurance for the latter two categories is voluntary. Accident insurance is purchased by the employer if individuals work more than 8 hours a week, and on an individual basis otherwise. Supplementary or private health insurance covers additional ambulatory services (e.g. alternative medicine) and expands the choice of private hospitals. For these contracts, insurers are allowed to select on risk and underwrite pre-existing conditions.

3 Theoretical Model

In this section, I develop a dynamic model of individual healthcare consumption with deductible choice, making three key contributions. First, the model distinguishes timing and classical moral hazard, and their implications for healthcare costs within the risk pool. Second, the model uncovers the interaction between these behaviors, as well as it generates selection. Third, I derive a sufficient statistic for timing moral hazard net of classical moral hazard and coverage choice, which can be estimated in reduced form by leveraging the randomness in the timing of health shocks.

3.1 Healthcare Consumption with Strategic Timing

Setup and timing. Consider a rational, forward-looking individual living through two calendar (coverage) years, divided into 24 months, $t = 1, \ldots, 24$.[7,8] Each year, individuals choose a deductible $D_t \in \mathbf{D}$, which resets in period $t = 13$. The choice

[6] Some insurers allow notifying until December 31st if the insured only wishes to increase their deductible.

[7] I omit the individual subscript for simplicity since I focus on individual-level behavior. Separating patient decisions from physician decisions and evaluating the cost-efficiency of consumption is beyond the scope of this paper.

[8] Modeling two years provides a tractable framework to describe timing incentives over a realistic time horizon, where health risk stays constant such that there is no *ex ante* incentive to retime care. The model can be extended to a longer time horizon with, e.g., repeated shocks and risk increasing with age. Certain procedures, such as hip replacements, can be shifted over longer periods.

set includes multiple deductible levels under mandatory health insurance (as in the Swiss or the Dutch setting).[9] Each month, the individual chooses two components of total healthcare consumption, $h_t = c_t + m_t$, which vary in terms of discretion over amount and timing. First, they choose a path for *planned care consumption*, m_t, which is known in advance and can be rescheduled over time. Second, they determine *spot consumption*, c_t, based on health shocks realizing every period. Components of healthcare consumption are private information to the individual, as the insurer (and the researcher) only observes the total. Each component also has distinct implications for costs within the risk pool.

I present the decision problem by backwards induction. The problem is dynamic because each unit of consumption increases the probability of exceeding the deductible. The deductible creates a discontinuity in the marginal price of healthcare as a function of total healthcare consumption in a given year. The turn-of-the-year reset creates a discontinuity in the price as a function of time.

Spot consumption. Spot consumption is chosen for all possible combinations of planned care consumption and deductible choice. Individuals choose c_t to maximize the trade-off between monetized health and money. They have the following value function:

$$\begin{aligned} V_t^c(a_t, m_t) = \max_{c_t}(c_t - \lambda_t) - \frac{1}{2\omega}(c_t - \lambda_t)^2 - C(c_t, m_t; R_t) \\ + \mathbb{E}\big[V_{t+1}^c(a_{t+1}, m_{t+1})\big] \end{aligned} \tag{2}$$

where individuals take as given planned care consumption, as well as a set of state variables $a_t = \{\lambda_t, \omega, R_t, D_t\}$. λ_t are health shocks drawn every period from a distribution $F(\lambda_t)$ which, importantly, allows for serial correlation over time. Per-period utility is concave in c_t and quasilinear in money.[10] The out-of-pocket cost function is $C(c_t, m_t, R_t) \equiv \min\{c_t + m_t, R_t\} + n$, where n are monthly premiums.[11] The cost

[9] The model can be extended to feature the possibility to opt in and out of insurance (as in, e.g., employer-sponsored health insurance in the United States), or any nonlinear price scheme that generates timing incentives across years. Coinsurance does not create incentives for timing moral hazard, as the marginal price of healthcare is constant.

[10] The structure of the spot consumption decision problem follows existing models of healthcare consumption under insurance with classical moral hazard (Einav et al., 2013; Abaluck et al., 2018; Klein et al., 2022). The quadratic functional form is an approximation of any utility function in the difference between healthcare consumption and nondiscretionary health needs. The quasi-linearity rules out income effects.

[11] This formulation incorporates the per-period budget constraint in the utility function. It follows the existing literature in assuming an exogenous income, and no saving and borrowing, see Klein et al. (2022) for a discussion. The deductible 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 away from the kink, as opposed to, e.g. maximum benefits (as in e.g. Cabral, 2017), which incentivize

depends on total consumption in that period, and the remaining deductible at the beginning of period t, $R_t \equiv \max\{0, R_{t-1} - c_{t-1} - m_{t-1}\}$, with $R_t = D_t$ for $t \in \{1, 13\}$. Individuals pay the full price of care out of pocket up to the deductible, and enter free care above.

This price schedule implies the following optimal spot consumption:

$$c_t(m_t) = \begin{cases} \lambda_t + \omega(1 - P_t(m_t)) & \text{if } R_t > 0 \\ \lambda_t + \omega & \text{if } R_t = 0 \end{cases} \tag{3}$$

where P_t is the year-end marginal price, which equals one minus the probability of exceeding the deductible by the end of the year. Once the deductible is exceeded, $P_t = 0$.[12] This solution rationalizes two components of spot consumption (Einav et al., 2013). First, λ_t can be interpreted as *nondiscretionary consumption* induced by health shocks that is urgent, and cannot be foregone nor timed. Consider a patient who suffers a heart attack and requires emergency care as well as specific follow-up medication to survive. Second, the preference parameter $\omega > 0$ determines *classical moral hazard.*[13] It measures the additional consumption that occurs if individuals exceed the deductible, e.g. a diagnostic test that individuals get when they do not have to pay out of pocket.

Classical moral hazard allows for discretion in the amount of healthcare consumption. It raises individual utility, but also costs in the risk pool. Forward-looking individuals only consider P_t *within the current year* in their spot consumption decisions. Importantly, planned consumption affects c_t via P_t. Individuals can choose their planned care consumption to get consumption closer or further from the deductible, and shape their opportunity for classical moral hazard. Above the deductible, c_t includes maximum classical moral hazard, and is independent of any further timing decisions.

Planned consumption. Individuals have a given amount of *planned care* $\mu_t \geq 0$ in every period t. In contrast to λ_t, μ_t is known in advance, non-urgent and can be shifted in time.[14] The actual planned care consumption path m_t is endogenous and

bunching at the price kink.

[12] Appendix A provides further details on the definition of the price.

[13] This parametrization implies that health is a normal good.

[14] In classical models of healthcare consumption without timing (e.g., Einav et al. 2013), λ_t captures the individual's private information about their not-yet-realized risk type which generates *ex ante* adverse selection. Here, μ_t captures additional private information about planned care, which can be seen as already-realized risk that is the source of *ex post* adverse selection introduced in Cabral (2017). Individuals can only shift care that they know they (will) need (e.g. a check-up), which may or may not be related to the shock.

the key object of interest. For every possible deductible, the individual chooses the path that solves the following Bellman equation, taking into account the optimal spot consumption that ensues:

$$V_t^m(a_t) = \max_{m_t} V_t^c(a_t, m_t) - g(m_t; \mu_t) + \mathbb{E}\big[V_{t+1}^m(a_{t+1})\big] \tag{4}$$

$$\text{s.t.} \quad m_t \geq 0 \tag{5}$$

$$m_t \leq \sum_{\tau=0}^{T} \mu_\tau - \sum_{\tau=0}^{t-1} m_\tau \tag{6}$$

$$\sum_{\tau=0}^{T} m_\tau = \sum_{\tau=0}^{T} \mu_\tau \tag{7}$$

At any time t, individuals can consume any portion of their remaining planned care, but the total amount must eventually be consumed. Because the total is fixed, in contrast to spot consumption, individuals allocate this consumption over time taking *relative* prices across years into account.[15] One reason to shift planned consumption to years when the price is lower is to reduce out-of-pocket costs. Interestingly, it is not the only driver for deciding when to consume planned care. The key insight of the model lies in the interaction between multiple sources of private information.

First, timing moral hazard complements classical moral hazard. Higher planned consumption within a given year reduces the year-end price P_t, which increases c_t and utility gains from classical moral hazard. As a result, individuals may choose to consume more planned care within that year to benefit from the increased utility of additional consumption, even if it means incurring retiming costs and covering part of the consumption out-of-pocket. Importantly, this interaction occurs even without the possibility of choosing the deductible. The magnitude of μ matters because it determines the flexibility individuals have in positioning themselves close to or further from the deductible. If planned care consumption is low, it will matter little for the year-end price.

Second, the planned care consumption path is closely tied to deductible choice. Regrouping care within a single year increases expected spending directly and indirectly through classical moral hazard, which can incentivize purchasing more coverage despite higher premiums. This interaction introduces a new source of *ex post* adverse selection. Individuals select their deductible based not only on the amount of planned care

[15] Choosing how to allocate an exogenously-fixed amount of consumption over time is qualitatively different from choosing the amount of consumption every period to reach a binding budget constraint. I implicitly assume that individual income allows covering all planned care out-of-pocket. This assumption is non-trivial in contexts where liquidity constraints affect healthcare consumption or health insurance decisions (Ericson and Sydnor, 2018; Gross et al., 2022).

they have shifted (as in Cabral 2017) but also on their expected classical moral hazard (as in Einav et al. 2013), and on the interaction between the two. As shocks realize under the current deductible, individuals can reoptimize their allocation of planned care and adjust their deductible for the following year. Delaying care raises expected spending and may prompt individuals to purchase more coverage in anticipation, leading to increased classical moral hazard in the following year. Conversely, advancing care reduces expected spending, which might decrease both the need for coverage and classical moral hazard in the next year. The differences in premiums play a critical role in these decisions.

Third, timing moral hazard may induce a utility cost $g(m_t; \mu_t)$.[16] One can see planned care as a risk that has realized and takes a known path μ_t. Deviating from this path requires active action which may be costly, e.g. due to hassle costs, supply-side scheduling frictions, or health effects. The model remains agnostic on the exact nature of this cost. I discuss possible microfoundations in Section 6.

3.2 Identifying Timing Moral Hazard Using Random Shock Timing

I now explain how variation in the timing of a large health shock in the first year allows the reduced-form identification of pure timing moral hazard in this framework.

Health shock in year 1. Assume that all individuals suffer an unanticipated, exogenous shock (a hospitalization) in a random period $s \in S = \{1, \ldots, 12\}$ of year 1, when the year-1 deductible is already chosen. The shock pushes their nondiscretionary consumption above the deductible, so that all consumption is free for $t = s, \ldots, 12$.[17] Total healthcare consumption is given by:

$$h_t(s) = \lambda_t(s) + \omega + m_t(s) \quad \text{for } t = s, \ldots, 12 \tag{8}$$

where the notation emphasizes that consumption within a period t varies depending on the timing of the shock s. By (3), lassical moral hazard is maximal for the rest of

[16] As implicit in (4), planned care can be shifted from any to any period. The utility cost is symmetric in the time-direction of the shift. It is incurred in the consumption period if it is advanced, or in period from which it is delayed. Any out-of-pocket cost savings occur when consumption was initially planned. In this model, as in, e.g. Cabral (2017), healthcare consumption does not translate into future benefits through greater health capital in that it does not impact its marginal utility in other periods. Another way to rationalize intertemporal substitution in healthcare is with a durable health capital (in the spirit of Grossman, 1972), in which individuals invest more when prices are lower.

[17] I focus on the incentives triggered by the first shock, but the serial correlation in λ allows for repeated shocks in relative time.

year 1, regardless of shock timing. Importantly, all uncertainty regarding the price is resolved, and any further decisions do not affect c_t for $t = s, \ldots, 12$. In other words, the shock shuts down the interaction between classical and timing moral hazard in year 1. This result will allow identifying $m_t(s)$.

The shock creates an incentive to reoptimize the planned care path and deductible choice for year 2, which may go in two directions. Individuals *advance* care from the second to the first year if the out-of-pocket cost savings are higher than the utility cost of advancing, and the utility loss from foregoing classical moral hazard in the second year. Furthermore, advancing care reduces the need for coverage in the next year, and yields savings in premiums. Conversely, individuals might *delay* care if the benefits from classical moral hazard outweigh the combined out-of-pocket costs of delayed care, and the utility cost of delaying. Delaying may come with higher coverage purchase in the following year due to higher expected consumption.

In this sequence, a higher preference for classical moral hazard ω increases the benefits of delaying over advancing care. A lower λ in year 2 makes advancing more attractive, as the chances of exceeding the deductible in year 2 are lower. Note that if nondiscretionary consumption exceeds the deductible in both years (e.g. due to a persistent shock), there is no incentive for timing moral hazard.

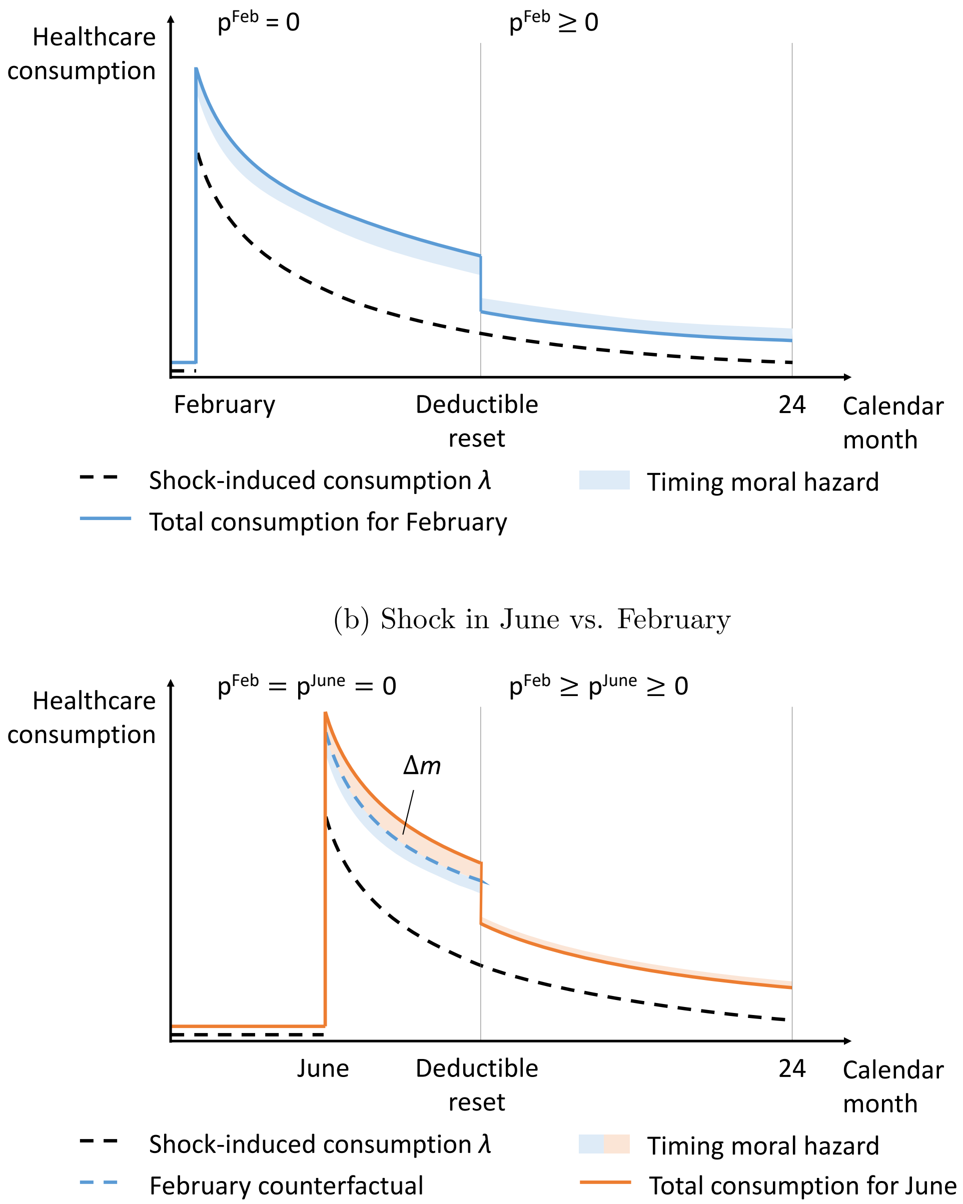
Role of shock timing. The timing of the shock affects individual behavior via two channels, as illustrated in Figure 2. First, it determines the time left until the year-end deductible reset: The later the shock, the less time to the reset. Individuals with a shock in February (Panel a) have more time to engage in timing moral hazard than those with a shock in June (Panel b). I allow $g(m_t; \mu_t, s)$ to depend flexibly on s to capture that the time horizon may affect the utility cost of retiming.

Second, shock timing shifts nondiscretionary consumption (black dashed line) into year 2: The later the shock, the more likely is the shock to persist into the year after.[18] The June group has higher nondiscretionary spending on average in year 2. All else equal, these spillovers push down the year-2 marginal price. Individuals with shocks later are thus more likely to delay, switch to a lower deductible, and engage in classical moral hazard in year 2. While the first two predictions can be tested, the latter cannot be given that the price is endogenous to the other two choices.

[18] In other words, the variation in dynamic price incentives stems from differences in health needs 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).

Figure 2: Dynamics of healthcare consumption following a health shock

(a) Shock in February



(b) Shock in June vs. February

Notes: The figure illustrates the intuition behind the identification of timing moral hazard based on shock timing and differences in consumption dynamics in relative time. It depicts exemplary healthcare consumption patterns of individuals with early (February) 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 consumption. The dashed blue line illustrates the use of the February group's consumption as a counterfactual. The area Δm marks the difference in timing moral hazard when care shifted from year 2 to year 1 following the shock realization.

If shock timing is random, all individual characteristics and preferences are orthogonal to s (conditional on having the shock). In other words, individuals differ in the timing of their risk realization but are otherwise comparable *ex ante*, in particular in the shock-induced λ and ω. I provide support for shock timing being random in the empirical implementation. This assumption implies that spot consumption is the same for all s in relative time between the shock and the deductible reset. Let $\Delta x_k(s,s') \equiv x_k(s') - x_k(s)$ for $s, s' \in S$ and $s' > s$, and take the difference in monthly healthcare consumption in (8) between shock groups in relative months $k = 1, \ldots, 13 - s'$:

$$\begin{aligned} &\Delta h_k(s,s') = \lambda_k(s') - \lambda_k(s) + \omega - \omega + m_k(s') - m_k(s) \\ \Rightarrow\ &\Delta h_k(s,s') = \Delta m_k(s,s'), \end{aligned} \tag{9}$$

where I use that nondiscretionary consumption is the same on average in relative time, i.e. $\mathbb{E}[\lambda_k(s)] = \mathbb{E}[\lambda_k(s')]$ for all $s, s' \in S$; and classical moral hazard equals ω every month until the deductible resets, since individuals are all in free care. The comparison in relative time has to be made on a monthly basis and only until the deductible reset. Individuals are no longer comparable afterwards, because the price is again endogenous to healthcare consumption and coverage purchase decisions.

The moment measures the difference in monthly planned care consumption, i.e. timing moral hazard, due to shock timing. It nets out shock-induced needs and classical moral hazard. Panel (b) of Figure 2 illustrates this result. In this example, the February group advances the same amount of care as the June group, but spreads it over a longer time interval between the shock and the reset. In that case, June has a lower monthly planned care consumption $\Delta m_k(2,6) < 0$, i.e. the orange line lies above the blue. If the amount for June is much smaller, e.g. because there is too little time to advance care, $\Delta m_k(2,6) > 0$ and the orange line would lie below the blue. Notice that the difference in (9) is sufficient but not necessary to reject the null of no timing moral hazard. A null difference may mask identical monthly $m_k(s)$, e.g. if no planned consumption is shifted. In Section 4, I show how to estimate the monthly differences in reduced form by comparing individuals across s. I then integrate these monthly differences up to compute total timing moral hazard in year 1.

The model uncovers a new source of *ex post* dynamic selection. Even individuals with *ex ante* identical preferences and health risk can become differentiated over time depending on the exogenous *timing of identical health shocks* within the coverage period. The nature of selection is asymmetric depending on the time-direction of the response. The possibility to advance care reduces coverage purchase via lower expected consumption in the next period. Some may even opt out of insurance if that option is available. Delaying creates adverse selection as individuals select higher coverage based on planned consumption. As I discuss in Section 7, shock timing provides information on behavioral responses to the insurer.

Robustness. I now discuss the robustness of the difference in (9) as a sufficient statistic for timing moral hazard under the assumption of random shock timing. Exploiting the large shock avoids having to specify the primitives and risk preferences shaping m_t. Interestingly, the moment is robust to including other choice options that influence m_t, even deductible choice in year 2. In that sense, timing moral hazard is identified even if we remain agnostic about the primitives driving timing and coverage choice. The moment identifies timing moral hazard as long as the shock triggers maximal classical moral hazard in year 1, where c_t is insensitive to any further decisions and can be differenced out. Identification is enabled by the additive separability of the utility in c_t and m_t. That is, spot and planned consumption only influence each other through out-of-pocket costs. This assumption is plausible if one sees planned care as fixed spending that can be consumed at any time within the two years, regardless of any spot consumption.[19]

The framework allows for multidimensional heterogeneity at the individual level, including key objects such as ω, $g(\cdot)$ and F. Random shock timing implies that individuals are comparable on average, leaving the identification result unchanged. While the framework does not allow identifying classical moral hazard, it also shows that this is unnecessary for estimating timing moral hazard after a large health shock.

4 Empirical Implementation

4.1 Insurance Claims Data

The analysis uses mandatory health insurance claims for all enrollees of CSS Insurance, the largest health insurer in Switzerland, in the years 2012 to 2019. CSS Insurance operates across the whole country, with approximately 800,000 customers yearly and a stable market share of 10%. It is subject to standard federal law on mandatory health insurance. Its enrollees are roughly representative of the Swiss resident population (Appendix Table B.1).

The data contain daily information on the costs of care covered by mandatory health insurance based on the actual date of treatment. Importantly, they cover all claims,

[19] If nondiscretionary consumption in a given period is associated with higher planned care, my approach identifies an upper bound on timing moral hazard.

including those below the deductible, as well as individuals who do not have any claims. I thus observe all costs covered by the insurer, and those paid out of pocket. Healthcare providers send claims directly to CSS, which then invoices the patient for any costs below the deductible. This happens by default unless the patient specifically requests to pay the healthcare provider and send the claim to the insurer themselves.[20] Costs are disaggregated by type of provider (e.g., physician, specialist, outpatient surgey clinic, hospital, imaging clinic, laboratory), as well as type of care (e.g. outpatient care, inpatient care, imaging, diagnostic tests, mental health treatment). Specific treatment codes are not available. For hospitalizations however, I observe the diagnosis-related group (DRG) that determines the reimbursement rate and is based on the patient's diagnoses, treatments received, age and length of hospital stay. With this, I construct an individual-level monthly panel of healthcare spending.[21]

The data further include individual demographic information (gender, age, nationality), insurance plan characteristics (premiums, deductible, type of plan, start and end dates of enrolment). I observe all determinants of premiums. They also include an indicator for whether the individual subscribes accident insurance at CSS. They also contain diagnoses inferred from claims for physician-prescribed drugs (e.g. cardiovascular, gastrointestinal, or mental health diseases). Finally, I observe the municipality of residence, which enters in the determination of premiums.

The baseline sample includes individuals aged between 19 (minors have different contracts, and switch in the calendar year where they become 18), and 90 (to limit the influence of end-of-life spending). I exclude insured-years with maternity or nursing home care, as these fall under different cost-sharing rules. I exclude incomplete insured-years with plan changes or interruptions due to, e.g., emigration, military service, or death. I also exclude temporary attriters, e.g. those who switch away and back to CSS. I however keep individuals who move during the year without changing the other features of their plan, although they may face a change in premiums. I do not observe individuals before and after they enroll with CSS. The yearly insurer switching rate is of 10%, which corresponds to the Swiss average.

[20] The deductible structure incentivizes the filing of all claims, as opposed to, e.g. a maximum benefit. Since the analysis focuses on individuals who exceed their deductible, unfiled claims are unlikely to be a concern.

[21] The month as a time unit balances the trade-off between statistical power and the granularity of the elicited dynamics, while smoothing out within-month variation in consumption and billing effects. I censor total monthly spending at CHF 20,000 (i.e. at approximately the 99th percentile of the distribution in the baseline sample) to avoid extreme outliers. All spending is nominal, as the available deductibles are constant.

Table 1: Summary statistics by level of deductible

	(1) High deductible	(2) High deductible and health shock	(3) Low deductible
Demographics			
Age	42.2 (13.4)	47.6 (15.2)	56.2 (18.7)
Female	0.41	0.50	0.58
Swiss	0.76	0.84	0.76
Insurance plan			
Monthly premiums	2,870 (654)	3,063 (843)	4,612 (838)
CHF 2500 deductible	1.00	0.84	0.00
Standard plan	0.35	0.39	0.30
Accident insurance	0.26	0.37	0.63
Spending and prices			
Total out-of-pocket spending	3,465 (1,194)	4,203 (1,453)	5,214 (969)
Total annual spending	1,039 (3,146)	2,955 (6,130)	5,896 (9,913)
Exceeded deductible	0.10	0.35	0.84
Cost sharing	0.55	0.60	0.29
Insured-years	1310801	101343	3790419

Notes: The table presents means and standard deviations (in parentheses) for samples of insured-years. Column (1) presents figures for insured-years with the high deductible of CHF 2,500. Column (2) contains the main analysis sample of high-deductible individuals with a health shock, defined as a hospitalization observed the first time in the observation window. Column (3) contains insured-years with low deductibles of CHF 300 and 500. All spending is in Swiss Francs (CHF). Cost-sharing is calculated as out-of-pocket spending (net of premiums) over total annual healthcare spending. Total out-of-pocket spending includes insurance premiums.

4.2 Reduced-Form Estimation

I now estimate timing moral hazard in reduced form based on the model. As there, identification relies on random shock timing. I adopt a multiple treatment framework, where the calendar month of the first shock $S_i \in \{2, ..., 11\}$ defines mutually-exclusive treatment groups.[22]

Main sample and shock definition. The main definition of a shock is the first hospitalization which occurs at least one year into the observation window. This definition ensures that individuals exceed their deductible for the rest of the year.

[22] 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, and the individuals do not switch treatment groups. I exclude individuals who have a shock in January and December in the empirical analysis to avoid turn-of-the-year spending specific to the winter holiday season, as well as billing effects. Healthcare providers often settle accounts and send out bills in December.

Over 90 percent of hospitalizations cost more than the highest CHF 2,500 deductible in my data.

I focus on individuals with a high deductible of CHF 2,500 to shut down selection at baseline, and incentives to advance care before (or in the absence of) the shock. For these individuals, a hospitalization is most likely unanticipated and therefore timed randomly. These individuals initially expect to end year 1 below the deductible.[23] Rational individuals who expect a costly hospitalization would choose a low deductible, especially if they are risk averse and prone to moral hazard (Einav et al., 2013). This sample restriction also decreases the likelihood that these individuals select into a specific treatment group by shifting expected spending towards the year of the shock. The main analysis thus relies on a selective but highly-relevant sample of high spenders with salient timing moral hazard incentives, as discussed further in Section 5.4, where I also present robustness checks with alternative definitions.

Table 1 provides descriptive statistics at the insured-year level. High-deductible individuals without a shock (column 1) represented 30 percent of the yearly population of insured in 2019. As expected, this sample is younger, and has a lower share of women than individuals who suffer the hospitalization with a high deductible (column 2) and those with a low deductible (column 3). These characteristics are strongly correlated with health status and healthcare consumption. The main analysis sample in column (2) is on average 48 years old, 50 percent female, and 84 percent Swiss. It generally lies in-between the other two samples in terms of average characteristics, insurance plan choices, and prices. Its premiums are slightly higher than for high-deductible insured-years without a shock. These figures support that the main sample includes individuals with relatively low baseline risk who suffer a severe health event. The low-deductible sample has characteristics associated with worse health and higher spending. They exceed their deductible 84 percent of years and face much weaker timing incentives.

Event study of monthly healthcare consumption. I set up an event study to evaluate how spending dynamics vary with shock timing and to estimate $\Delta h_k(s, s')$ in reduced form. If shock timing is random, any differences in spending between the shock and the reset capture timing moral hazard. 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 outcome of interest is healthcare consumption h_{it} for individual

[23] The shock itself does not reflect whether the high deductible has been chosen optimally, but rather captures that the individual suffered an unfavourable realization of health needs. It is possible that healthcare consumption increases before the hospitalization itself, which signals an worsening of the health status. Consistent with this, the results show that anticipatory spending is similar across treatment groups, and mainly consists of outpatient spending and diagnostic tests.

i in calendar month t. I estimate the following event study at the insured-month level:

$$h_{it} = \sum_{s=2}^{11} \mathbf{1}\{S_i = s\} \left(\sum_{k=-12}^{23} \gamma_k^s \mathbf{1}\{t - E_i = k\} + \gamma_{23+}^s \mathbf{1}\{t - E_i > 23\} \right) + \sigma_t + \nu_i + \zeta X_{it} + \varepsilon_{it}. \tag{10}$$

The coefficients γ_k^s capture the effect of the shock on healthcare consumption in relative month $k \in \{-12, ..., 23\}$, and so for all treatment groups s relative to those with a shock in February. Relative shock time is normalized to the pre-shock period $k = -1$. Long-term level effects are captured by γ_{23+}^s. The results are not sensitive to setting this horizon to 12 or 36 months. Dynamic treatment effects inform on shock anticipation and persistence via the interaction of the relative period indicators with the treatment group indicators. They are assumed to be homogeneous across individuals within a treatment group.[24] This specification allows computing estimates for the differences in timing moral hazard as in (9) for 210 comparison pairs $\{s, s'\}$ for $s, s' \in S$ and $s' > s$:

$$\Delta\hat{m}_k(s, s') = \hat{\gamma}_k^s - \hat{\gamma}_k^{s'} \tag{11}$$

For instance, $\Delta\hat{m}_4(2, 3)$ provides an estimate of the timing moral hazard of the March relative to the February group in the fourth month after the shock, holding all else equal.

Seasonality in nondiscretionary healthcare consumption is controlled for by σ_t. This includes calendar month dummies to take out differences in seasonal healthcare spending that would occur even in the absence of the shock. That is, for e.g. all relative months corresponding to December, they take out baseline spending on seasonal flu (assumed to be homogeneous across treatment groups), but the differential timing moral hazard responses (heterogeneous across treatment groups) remain identified. Furthermore, σ_t also includes a quadratic polynomial trend to account for secular trends (e.g. changes in insurance premiums, technology, ageing of the sample).[25] Individual fixed effects ν_i subsume any time-invariant individual characteristics (e.g. gender, preferences, education, baseline health) that determine baseline healthcare

[24] The random shock timing assumption implies that all individuals with $S_i = s$ experience the same average effect γ_k^s in any given relative month. In other words, the cohort of individuals with a shock in March 2012 has the same (conditional) dynamic spending patterns as the March 2013 cohort.

[25] Calendar month dummies and time trend 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 for the same reasons.

consumption and potentially correlate with shock timing. Still, identification relies on between-individual variation. X_{it} controls for time-varying individual characteristics (age, type of insurance plan, accident insurance, and canton). Finally, ε_{it} is random noise. Estimations are performed using linear least squares with standard errors clustered to allow for arbitrary correlation at the individual level.

From monthly differences to total timing moral hazard. Comparing monthly healthcare consumption across all shock groups pairs from February to November in relative months after the shock as in (11) yields 210 data points for $\Delta\hat{m}(s,s')$. I now integrate up the estimates of the difference to obtain the total timing moral hazard response. A parametrization of m_t is required as the system in differences is otherwise under-determined. I make this functional form assumption in a data-driven manner based on the source of variation. The negative linear relationship between the estimates $\Delta\hat{m}_k(s,s')$ and shock timing (see Appendix Figure B.10) suggests that the integral $m(s)$ is a quadratic function of s, $m(s) = \alpha + \beta s + \delta s^2$, such that the derivative is $m'(s) = \beta + 2\delta s$. This functional form allows $m(s)$ to vary across groups due to, e.g. frictions in timing, but the amount of planned care consumption is assumed constant over year 1 through a consumption smoothing argument. I run the following regression to estimate $(\hat{\beta}, \hat{\delta})$ that most closely fit the difference estimates rescaled by the distance between shock groups $\Delta s \equiv s' - s$:[26]

$$\frac{\Delta\hat{m}_k(s,s')}{\Delta s} = \beta + 2\delta s + \epsilon. \tag{12}$$

To compute a lower bound for $\alpha = \underline{\alpha}$, I use $(\hat{\beta}, \hat{\delta})$ and the constraint that planned care consumption is weakly positive $m(s) \geq 0$. Finally, I calculate total timing moral hazard $M(s)$ in the shock year as follows:

$$M(s) \equiv \sum_{t=s}^{12} m(s) = (13 - s)\left(\underline{\alpha} + \hat{\beta}s + \hat{\delta}s^2\right). \tag{13}$$

This quantity measures the difference in planned care consumption in year 1 across shock groups, where the lower bound is set by the group with the lowest prediction.

4.3 Identifying Assumption of Random Shock Timing

As in the theoretical model, the identification of timing moral hazard relies on the timing of the health shock being exogenous. Defining the health shock requires particular

[26] The difference estimates correspond to approximations of the derivative, as $\partial m_k(s)/\partial s = \lim_{\Delta s \to 0} \Delta m_k(s,s')/\Delta s$. I assume that the time distance between groups Δs does not systematically bias the estimate.

care, as it determines sample inclusion and treatment group assignment, and affects the plausibility of shock timing being random. Under the ideal shock, the composition of the population remains stable throughout the year, such that treatment groups provide valid counterfactuals for each other.[27] In other words, there should be no time-varying unobserved factors that simultaneously influence healthcare consumption and the probability of experiencing a shock in a given calendar month (possibly conditional on observable characteristics). There may still be seasonality in the probability of occurrence; for example, even heart attacks, which are generally unanticipated, are more likely in winter (Kurihara et al., 2020). Defining the shock as the first hospitalization has the advantage of maintaining generality and statistical power,[28] as hospitalizations occur throughout the year, at all ages and deductible levels.

Importantly, high-deductible individuals in my sample are unlikely to have strategically timed a hospitalization, otherwise they would have taken a low deductible.[29] They also have a low *ex ante* probability of exceeding the deductible, which reduces the likelihood that the hospitalization itself results from a classical moral hazard response.

In Appendix B.2, I conduct multiple checks to support that there is no systematic selection into the month of the first hospitalization in the high-deductible group. First, I assess balance in average characteristics across treatment groups and find no significant differences in observable factors, such as demographic characteristics, drug-based diagnoses, and insurance plans. Additionally, shock groups exhibit similar probabilities of attrition. Second, I examine whether pre-shock characteristics predict the timing of the shock by calculating propensity scores for each possible treatment group and comparing these scores across actual treatment groups. The close overlap in propensity score distributions indicates that observable factors do not jointly predict shock timing. Third, I analyze healthcare consumption patterns before and at the shock

[27] Identification also requires that the stable unit treatment value assumption holds, i.e. the outcome of an individual in treatment group s does not affect the outcomes of those that suffer the shock in another calendar month. This assumption is plausible in the Swiss context, given the relatively low supply-side constraints.

[28] Shock definitions using specific types of hospitalizations lead to small sample issues, even for common ones such as strokes or heart attacks. These are also more likely to occur among older, chronically-ill individuals with low deductibles and no timing incentives. Comparability is not guaranteed by knowing the exact nature of the shock, as any shock may unobservably differ in their severity.

[29] Consider an individual with a high deductible who learns that they need a costly elective hospitalization. They have an incentive to delay it and take up a low deductible next year, so as to reduce out-of-pocket costs, and benefit from free care. By doing so, they would enter a shock group in the next calendar year, which may then yield a selected group of individuals prone to timing moral hazard. However, these individuals do not enter my sample with high deductibles in the year of the shock. Furthermore, my analysis excludes the earliest and latest months to avoid any turn-of-the-year effects. Hence, individuals in my sample are unlikely to have anticipated and timed a hospitalization (bearing high insurance plan switching costs).

to ensure comparability in terms of nature and severity. The shock mainly involves inpatient spending, while anticipatory spending includes outpatient physician and specialist visits, as well as diagnostic tests and imaging. Finally, the lack of differential pre-trends indicates that individuals do not adjust their healthcare consumption to enter specific months (see Figure 10 and Appendix B.5). This helps rule out scenarios where more severe patients might be hospitalized earlier in the year to consume more care. Controlling for compositional differences in time-varying individual characteristics does not affect the results. I present robustness checks with alternative shock and sample definitions in Section 5.4.

Appendix Figure B.5 shows the marginal price at end of the shock year across shock groups. Year-end prices in the shock year (panel a) are at 0.1 on average, i.e. in the 10 percent co-payment region, and not significantly different from each other. This result suggests that incentives between the shock and the reset are aligned across groups. It also supports that dichotomizing the price structure is a reasonable approximation in the Swiss setting.

5 Results

5.1 Dynamic Healthcare Consumption Patterns

This section examines how the shock shapes healthcare consumption dynamics. Figure 3 depicts the coefficients on dynamic treatment effects for selected shock groups from the event study. Several patterns emerge. First, consumption begins to increase about six months before the shock, though these increases are small and do not lead to exceeding the deductible on average. The dynamics support the assumption that treatment groups do not differ systematically in their pre-shock consumption.[30] The approach remains valid if the onset of health deterioration leading to the shock is random and its severity is comparable across groups.

Second, the close magnitude of the spikes at the shock support that shock groups are comparable in severity. Third, spending gradually decreases and stabilizes one year after the shock, with a long-term effect of around CHF 200 per month, i.e. CHF 2,400 per year, which is close but not greater than the highest deductible of CHF 2,500. This pattern supports that shocks persist on average. The persistence into the next year thus varies with shock timing, and induces differential timing incentives. Finally, future spending dynamics mainly differ between the shock and the deductible reset.

[30] While I refrain from plotting confidence intervals to emphasize the link to the theory, less than 10 percent of the monthly pre-shock differences are statistically significantly different from zero.

This provides a first indication of timing moral hazard.

I provide further evidence on how the timing of the shock influences cumulative healthcare consumption in both the year of the shock and the following year. Figure 4 shows differences in cumulated dynamic treatment effects in calendar time. I take intervals between the shock and the year-end reset, over the whole shock year, the following year, as well as in both years taken together (see Appendix B.4 for details). Panel (a) shows a significant negative relationship between shock timing and total consumption between the shock and the year-end reset. This happens partly mechanically as individuals with later shocks have lower nondiscretionary consumption in year 1. However, it can also be driven by timing moral hazard. The difference reaches CHF 3,000 for the November relative to the February 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 consumption prior to the shock. Panel (c) suggests that the later the shock occurs, the higher the total consumption in the year after. This difference amounts to CHF 1,400 between February and November groups. Taking both years together in Panel (d), larger consumption in the following year offsets lower consumption in the shock year for early shock groups, but not for later ones. In sum, these results indicate the presence of shock spillovers, but are also suggestive of timing moral hazard via shifts in consumption from year 2 to year 1.

5.2 Estimates of Timing Moral Hazard

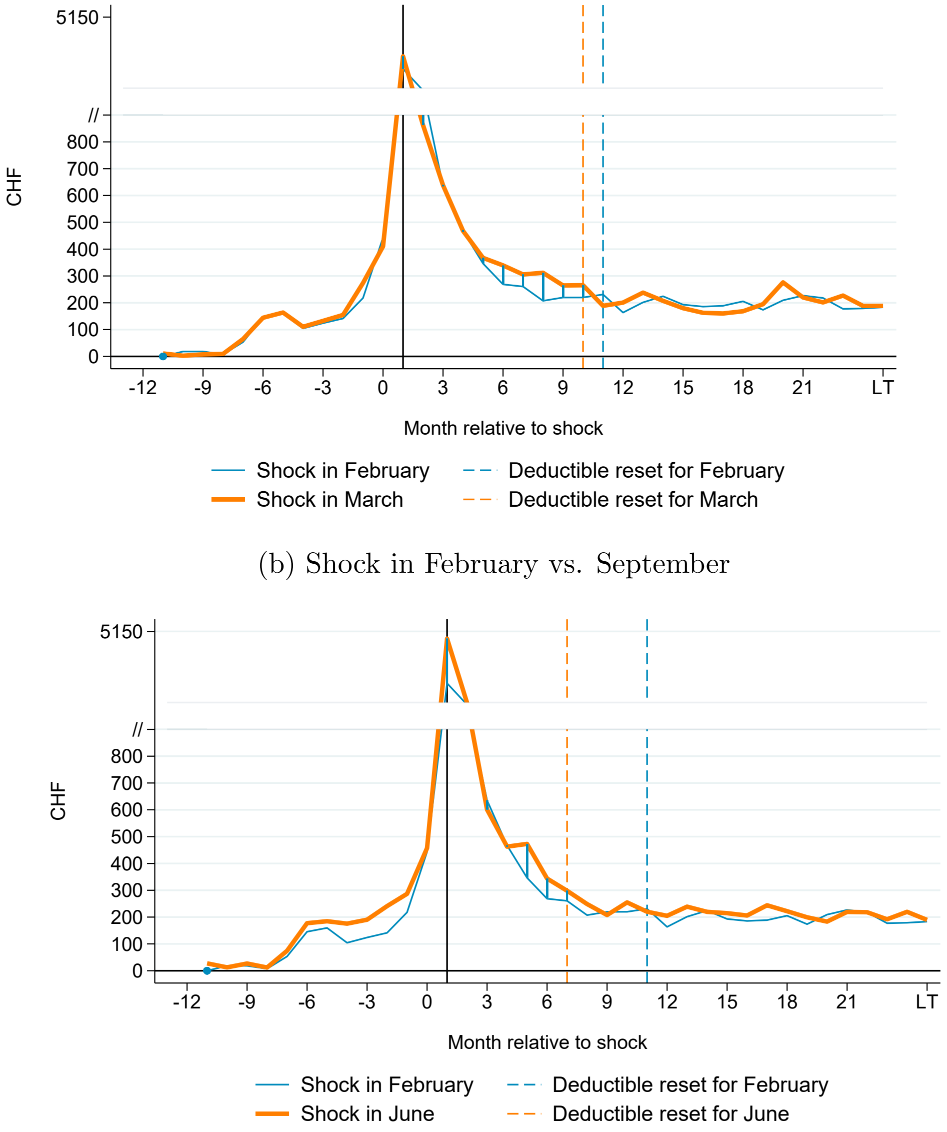
Monthly differences. The event study yields 210 point estimates for the monthly differences in total healthcare consumption across shock month comparison pairs, $\Delta\hat{h}_k(s, s')$. The average difference is CHF 22.30 (SD=88.20).[31] The differences decrease as a function of s and cross the zero line. This relationship serves to inform a data-driven choice of a quadratic functional form for $m_k(s)$ (see Section 4.2). Appendix Figure B.10 presents results from the regression in (12).[32] The estimate for β (i.e. the constant in the regression) is not significantly different from 0. The coefficient on shock timing, 2δ, is negative and statistically significant at -13.70. I use the parameters to predict the total timing moral hazard across shock months as in (13).

This negative relationship is consistent with early shock groups advancing, and later shock groups postponing care in the year of the shock. Importantly, it indicates a significant variation in the amount of care advanced as a function of shock timing.

[31] The event study specification unadjusted for covariates yields virtually identical results.

[32] This specification minimizes the Akaike information criterion, and higher order terms are not significant.

Figure 3: Event study of healthcare consumption around the health shock

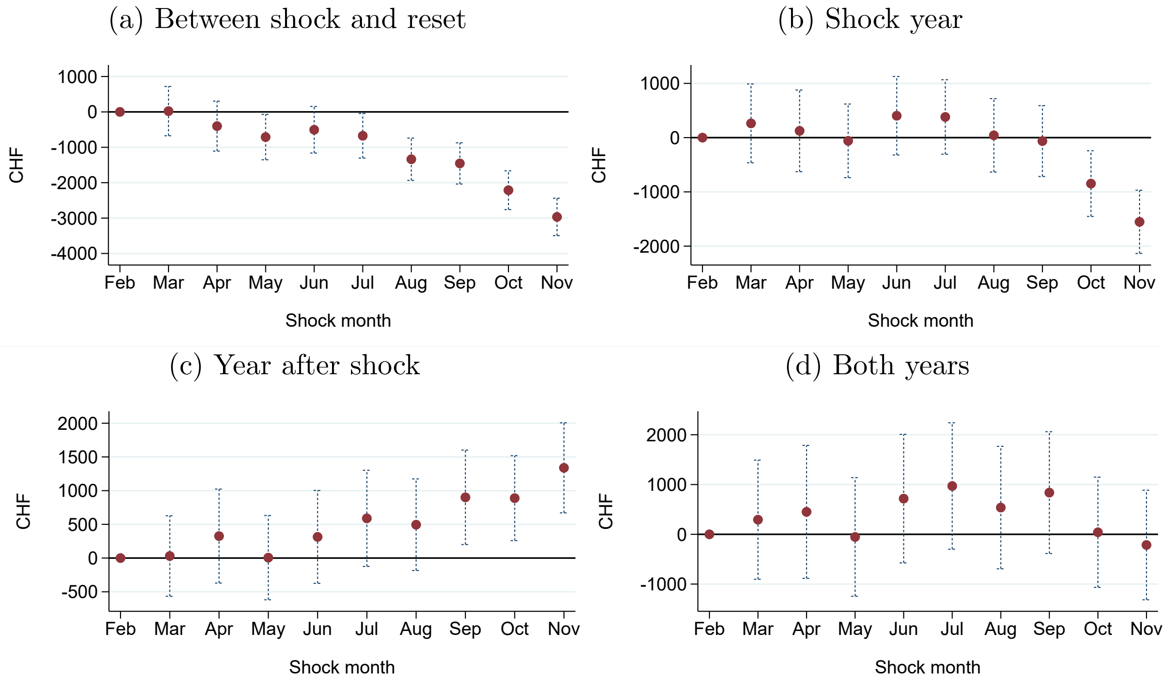


(a) Shock in February vs. March

(b) Shock in February vs. September

Notes: The figure depicts the coefficient estimates on monthly treatment effects from the event study of healthcare consumption (as measured by total spending). It compares individuals with shocks in February vs. March (panel a), and February vs. June (panel b). Additional comparisons are displayed in Appendix B.5. The estimation uses the main analysis sample of insured with a high deductible and the health shock defined as the first observed hospitalization. These effects are normalized to the average spending of the February group 12 months before the shock. The dashed lines indicate the last month before the year-end deductible reset 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 by shock month, relative to February



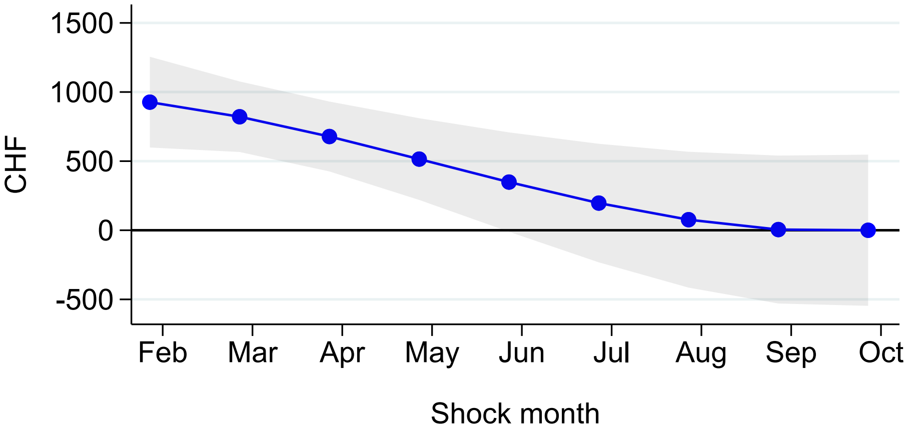
Notes: The figures depict cumulated differences in dynamic treatment effects in calendar time between (a) the shock and the year-end reset, (b) over the whole shock year 1, (c) the post-shock year 2, as well as (d) in both years taken together. Details are presented in Appendix B.4. All differences are in Swiss Francs (CHF), and taken relative to the February group. Confidence intervals at the 95 percent level are based on block-bootstrapped standard errors with 49 replications, clustered at the individual level.

Recall that significant differences between treatment groups are sufficient to reject the null of no timing moral hazard in this framework.

Total timing moral hazard. Figure 5 shows the main estimates of timing moral hazard, i.e. the difference in planned care consumption across months. Timing moral hazard is large and statistically significant for shocks groups until May. Individuals with a shock in February consume CHF 940 more of planned care than those with a shock in November. This represents 10.2 percent of their total consumption of CHF 9,200 in the shock year. This means that individuals are forward-looking and act on the timing incentive.

The amount of planned care consumption decreases the later the shock occurs, i.e. the less time individuals have left in the shock year. The estimates for individuals with a shock later than June are not significantly different from zero. A null estimate in this context provides a lower bound on planned care consumption if it is interpreted as a complete delay of all planned care. I explore the drivers of heterogeneity across groups in Section 6. Given $m'(s) < 0$, the last shock month provides the lower bound

Figure 5: Total timing moral hazard by shock month



Notes: The figure presents estimates of the total yearly timing moral hazard response across shock months, computed as in (13). The last shock month serves as a lower bound. Confidence intervals at the 95 percent level are based on bootstrapped standard errors with 49 replications, clustered at the individual level.

at zero to pin down the level of the timing moral hazard. Although the decrease is predicted by the model, it is not imposed *ex ante* in the estimation.

The magnitude of the timing response is in line with the differences in cumulated consumption in the year after the shock (Panel (c) of Figure 4). This comparison suggests that a large part of the differences in consumption in the year after the shock is due to timing (70 percent for February). The remainder is due to shock-induced consumption and classical moral hazard. My framework does not allow identifying these two components.

A back-of-the envelope calculation shows that timing and classical moral hazard are of similar magnitude, presenting an important trade-off for policy. I use the benchmark estimate of -0.2 (Keeler and Rolph, 1988) to compute bounds for classical moral hazard and compare it to timing for the February group. The February group provides a relevant reference for gauging the effect of a cost-sharing policy which would affect the marginal price over the whole coverage year. If the benchmark measures total moral hazard relative to nondiscretionary consumption, classical moral hazard is equal to CHF 593, which translates into elasticities of $\epsilon^c = -0.08$ for classical, and $\epsilon^m = -0.12$ for timing moral hazard. If the benchmark measures pure classical moral hazard (i.e. itis not confounded by timing and $\epsilon^c = -0.20$), classical moral hazard equals CHF 1337 and the timing elasticity is $\epsilon^m = -0.14$.[33] This exercise illustrates how existing

[33] In both scenarios, I use an extreme marginal price change of -1 within year 1 (i.e. exceeding the deductible, which also yields a price difference of -1 relative to year 2) and I back out nondiscretionary needs from the total consumption. I then infer the portion attributable to timing moral hazard using

estimates of classical moral hazard may be overestimated by twice the magnitude of the structural within-year price elasticity if they include the timing response. A similar point is made in Einav et al. (2015). However, timing moral hazard is inherently dynamic, and the timing elasticity decreases to 0 for shocks in the second half of the year.

For interpretation, it is important to note that the main estimates are based on a selected sample of individuals with high deductibles who become high spenders and face timing incentives due to hospitalization. Understanding the behavior of this group is particularly relevant for policy, as unanticipated hospitalizations have lasting negative effects on key economic outcomes, such as earnings, access to credit, and consumer borrowing (Dobkin et al., 2018). A distinctive feature of healthcare markets is that a small proportion of high-spending individuals accounts for a significant share of total costs. In my data, this main sample represents 14 percent of collective healthcare spending.

In terms of timing incentives, the costly shock makes the price change salient. Smaller changes may blur the incentives and hinder individuals from forming correct expectations about the year-end price (Brot-Goldberg et al., 2017).[34] In terms of preferences, it is plausible that the sample a lower price-sensitivity, as individuals who are less prone to classical moral hazard tend to select into lower coverage *ex ante* (Einav et al., 2013). In terms of plan switching, the sample only includes individuals who are observed for at least three years, that is, who keep the same insurer and may have high switching costs.[35]

5.3 Composition of Timing Moral Hazard

I now evaluate the composition of the timing response. I repeat the estimation procedure for categories of healthcare spending, and calculate the portion of the total timing response attributable to each category. Figure 6 presents the results for the February group. Other shock months show identical composition since planned care consumption decreases proportionally with s across all categories. In other words, the time-heterogeneity is homogeneous across categories.

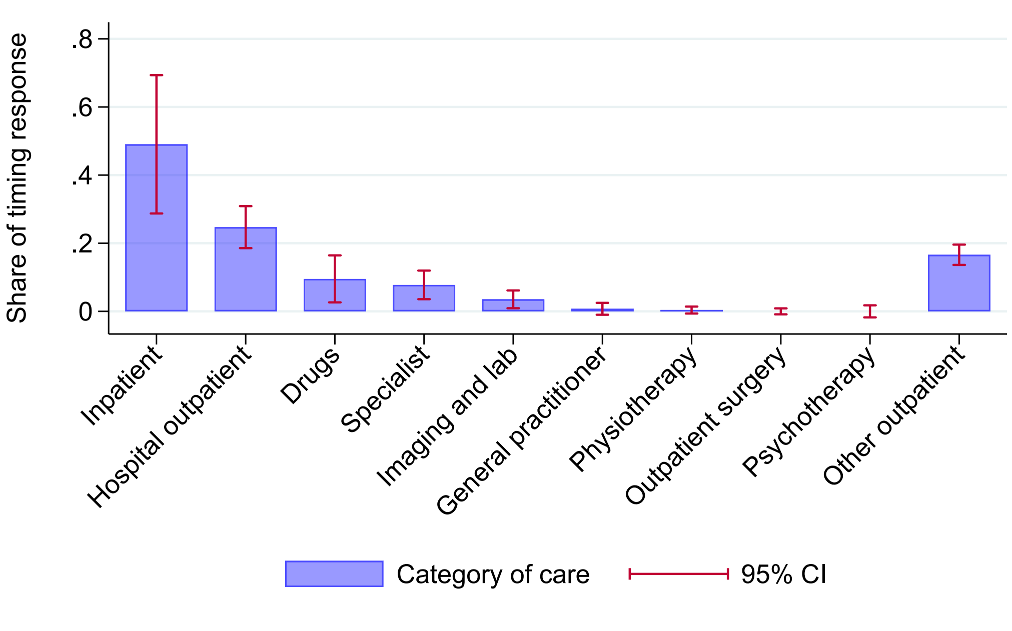
Timed care is roughly evenly split between inpatient care (hospitalizations with an overnight stay) and outpatient care (all other types). This contrasts with the overall

my main estimates.

[34] Some of the highest price-elasticity estimates have been found using exogenous variation in exceeding the deductible (e.g. Kowalski, 2016).

[35] Only 8% to 10% of insured in Switzerland switch insurers during the enrollment period. Individuals in the main sample are observed for 6.7 years on average.

Figure 6: Composition of timing moral hazard by category of care



Notes: The figure presents the share of the total timing response for each category of care for individuals who had a shock in February. Confidence intervals at the 95% level are based on bootstrapped standard errors with 49 replications, clustered at the individual level.

composition of healthcare consumption, where outpatient care accounts for only about 20 percent of total consumption in the shock year. This disparity supports the intuition that outpatient visits are easier to time than hospitalizations, thus driving the timing response. Prescription drugs play a significant role in timing, with estimates of CHF 110 for the February group. Unlike other medical procedures, drugs can be stocked for later use, as highlighted by Einav et al. (2015). Specialist visits, diagnostic imaging, and lab tests are also subject to timing adjustments, but they represent a smaller share of both timed and total consumption. In contrast, general practitioner visits, physiotherapy, and psychotherapy are less prone to timing adjustments, likely because they involve routine and acute visits, or treatments that span longer periods.

5.4 Heterogeneity and Alternative Shock Definitions

This section presents results for alternative samples and shock definitions. These analyses serve to gain further insights into heterogeneity in timing moral hazard, as well as to further assess the validity of the identifying assumption of random shock timing.

Persistent and temporary shocks. The timing response may vary depending on the persistence of the shock. Temporary shocks create a stronger incentive to advance care, as they are less likely to cause individuals to exceed their deductible again in the following year. The data support this intuition: Panel (a) of Figure 7 shows that planned care consumption is significantly higher for temporary shocks, while estimates for persistent shocks are close to zero and nonsignificant. Individuals with a persistent shock are defined as those whose annual spending increases by at least CHF 500 (roughly the median) 12 to 24 months after the shock, compared to pre-shock levels.[36] Similarly, Appendix Figure B.11 examines a subset of hospitalization codes that occur at least 1/7 of the time on weekends. Following the approach of Card et al. (2009), this subset likely includes a higher proportion of extremely emergent hospitalizations that cannot be delayed, even by a day or two to occur on a weekday, when typically more medical staff is available. The amount of planned care consumption in this group is much lower than in the main estimates at CHF 250.

Overall, the results suggest that more severe or urgent shocks limit individuals' ability to advance care. These individuals are more likely to postpone care to the year following the shock, potentially to reduce non-urgent healthcare consumption immediately after a severe event. Consistent with this, individuals experiencing persistent shocks are more likely to switch to a lower deductible (as discussed in Section 6).

Health insurance plan. Panel (b) of Figure 7 explores another interesting hypothesis: provider choice restrictions and gatekeeping may influence timing moral hazard. I divide the sample into individuals with standard plans offering free provider choice and those with restricted plans. Although individuals with standard plans typically have higher spending and may be selected based on their preference for consumption flexibility, the timing estimates for this group are four times larger than those for individuals with restricted plans. This suggests that choice restrictions can influence timing moral hazard. Appendix Figure B.12 supports this interpretation, showing that individuals with earlier shocks visit, on average, one additional category of healthcare providers or specialties.

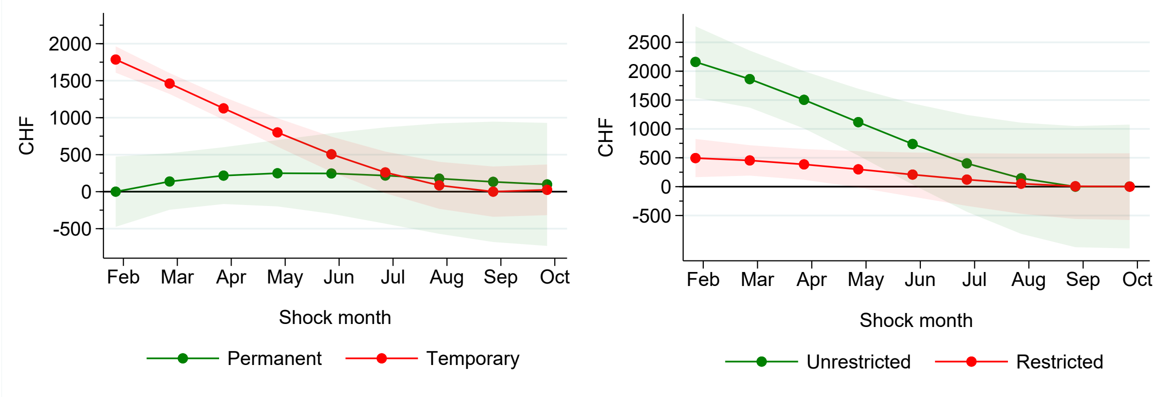
Appendix Figure B.13 shows that estimates for the sample of individuals with a CHF 1,500 deductible in the year of the shock are close to the ones for the main CHF 2,500 sample. Deductibles in the middle of the choice range in Switzerland are still relatively high and create timing incentives, though the middle deductible is cost-minimizing within a very narrow range of spending. The highest timing estimate is smaller than the middle deductible level.

[36] Although this sample split is based on endogenous outcomes post-shock, the timing incentives due to the shock are resolved by this point, and consumption stabilizes at its long-term level, as shown in Figure 10.

Figure 7: Heterogeneity in total timing moral hazard

(a) Persistent vs. temporary shock

(b) By freedom of choice of provider



Notes: The figure presents estimates of the total yearly timing moral hazard response across shock months, predicted as in (13). The last shock month serves as a lower bound. Confidence intervals at the 95% level are based on bootstrapped standard errors with 49 replications, clustered at the individual level. Panel (a) splits the sample into temporary and persistent shocks. Individuals with a persistent shock are those whose yearly spending increases by at least CHF 500 (i.e. roughly the median), 12 to 24 months after the shock compared to the pre-shock level. Panel (b) divides the sample into individuals with standard plans offering free provider choice and those with provider restrictions.

Demographic characteristics. Appendix Figure B.14 shows additional sample splits in terms of individual demographics. Beyond their propensity to respond to timing incentives, these groups systematically differ in health status and the nature of their shocks. The analysis reveals that planned care consumption in the shock year is higher among men and individuals under 40, both of whom have higher incidences of temporary shocks, as well as individuals living in a municipality with above-median average income. Despite these differences, the key patterns persist across all groups: estimates are significant for those experiencing early shocks but decrease for later shock groups.

6 Determinants of Timing Moral Hazard

Deductible choice. The model predicts that individuals are more likely to delay care and purchase more coverage to year 2 if the shock happens later in the year, and is persistent. I check for heterogeneity in deductible choice in the year across shock groups, and so after persistent and temporary shocks. Figure 8 shows the share of individuals who keep the CHF 2,500 deductible the following year. The levels differ significantly by shock persistence, at 92 percent for a temporary shock and 82 percent for a persistent shock. Overall, a small share of individuals switches to a low deductible, although half of those with a persistent shock exceed it the year after, and about 10 percent of those with a temporary shock do (especially those with a late shock).

Interestingly, the switching rate does not vary by shock month.[37] A possible explanation for this flatness is plan switching costs. These may be higher for individuals who experience the peak of the shock around the time of deductible choice in November, but also have the strongest incentives to switch. Furthermore, 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.[38] Recall that my panel contains high-deductible individuals who are negatively selected on the propensity to switch plans. This result yields two insights. First, there is little evidence for dynamic selection based on shock timing in my sample. Second, the heterogeneity in $M(s)$ is driven by the amount of timed care, rather than the share of strategic timers. To see this, think of the timing moral hazard estimates as a weighted average of the response of individuals who time, and those who do not.

Microfoundations for the utility cost of timing moral hazard. Given the lack of heterogeneity in deductible switching rates, frictions to timing are likely important determinants of heterogeneity in timing moral hazard across shock months. Although the model remains agnostic about the exact sources of these frictions, various constraints specific to healthcare markets can restrict individuals from retiming care flexibly and explain the low planned care consumption for later shock groups.

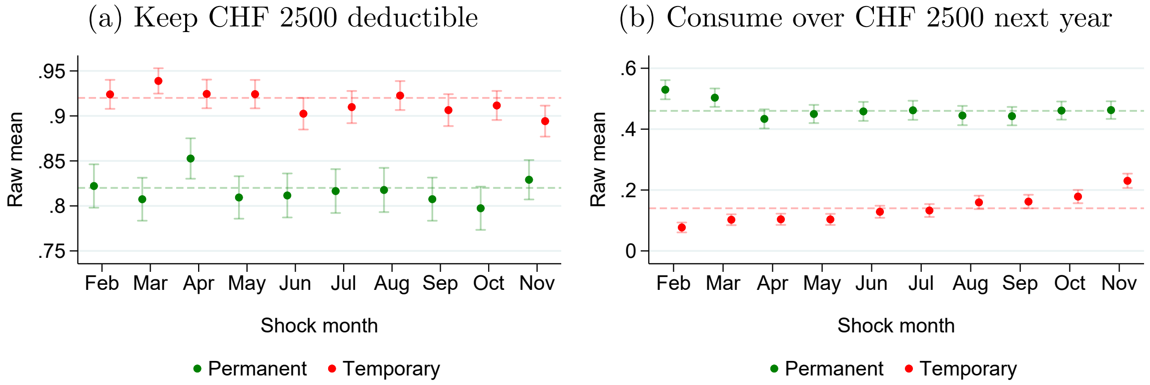
Healthcare supply can impose time constraints, such as limited patient control over appointment scheduling or the need for referrals. Additionally, capacity constraints may emerge toward the end of the year, as many patients increase their healthcare consumption after exceeding their deductibles (Lin and Sacks, 2019; Gerfin et al., 2015). This surge in demand can restrict the ability of late shock groups to advance their care. Another key characteristic of healthcare consumption is its lumpiness—patients typically cannot adjust just any continuous amount of care. Medical treatments are often bundled and follow a sequence that cannot always be compressed into a short period, making it easier to fit these bundles within a longer time horizon.

The utility cost of timing moral hazard may also be influenced by effort costs, such as scheduling doctor's appointments. The costs associated with finding a healthcare provider who can accommodate necessary treatments might also deter individuals from

[37] The share of individuals choosing a standard plan is also constant across shock groups, such that there is no differential selection into alternative managed care, family doctor or telemedicine plans that limit the choice of healthcare providers.

[38] See e.g. Abaluck and Gruber (2011); Handel (2013); Abaluck and Gruber (2016); Handel and Kolstad (2015); Heiss et al. (2016), as well as Winter and Wuppermann (2019) for a review of the recent literature.

Figure 8: Deductible switching and exceeding rates in the year after the shock


(a) Keep CHF 2500 deductible

(b) Consume over CHF 2500 next year

Notes: Panel (a) displays raw share of individuals who keep the CHF 2,500 deductible in the year after the shock, and so across shock months. Panel (b) displays shares of individuals who consume over CHF 2,500 worth of care in the year after the shock. The sample is split into temporary and persistent shocks. Individuals with a persistent shock are defined as those whose annual spending increases by at least CHF 500 (roughly the median) 12 to 24 months after the shock, compared to pre-shock levels. The dotted horizontal lines denote the corresponding subsample average. Confidence intervals are at the 95% level, based on robust standard errors.

advancing care, especially on short notice. These transaction costs may be higher during the acute phase of a shock. Individuals who experience a shock close to the deductible reset might be less inclined to schedule additional appointments within a short timeframe. In terms of health, it may be that deviating from the default or recommended treatment path μ_t reduces its effectiveness. Finally, a recent literature has identified several behavioral biases on the demand side that prevent the insured from achieving their optimal consumption.

Discounting. One behavioral factor that may influence healthcare consumption under dynamic price incentives is myopia. The model above describes the behavior of a rational, forward-looking individual. It can be extended to incorporate various modes of discounting, including (quasi-)hyperbolic discounting as in Klein et al. (2022); Abaluck et al. (2018) and Dalton et al. (2020). Healthcare consumption then depends on a weighted average of the current and the year-end prices. In my shock setup, spot consumption in the first year would remain unchanged as both the current and year-end prices are zero. In general, myopia works against observing timing responses as individuals would not fully anticipate these incentives, not only in terms of future health needs, but also price incentives (Brot-Goldberg et al., 2017; Abaluck et al., 2018; Dalton et al., 2020). However, the significant timing moral hazard I observe suggests that individuals are not entirely myopic. They adapt in the face of a bad risk realization by shifting large amounts of planned healthcare consumption in time. Furthermore, early shocks provide more time to act on timing incentives, although late shocks make the impending deductible reset more salient. The framework allows for individual heterogeneity in discount factors under random shock timing which average out across groups.

7 Implications for Costs and Insurance Markets

The theoretical and empirical results in this paper offer relevant insights on how dynamic and multidimensional asymmetric information shapes individual behavior under nonlinear cost-sharing. The framework clarifies that responses have different implications depending on whether they affect the *amount* or the *timing* of healthcare consumption. Healthcare consumed on a rolling basis as health risks realize within coverage years is subject to classical moral hazard. This response generates additional costs that are covered by the insurer, and that would not have occurred under higher cost-sharing for consumers. Timing moral hazard results in the insurer covering consumption that would have occurred anyway—and in that sense, no longer a pure risk—but would have been paid out of pocket by the insured. Taken in isolation, both behaviors increase costs in the risk pool and can affect premiums. Interestingly, their interaction makes their overall effect ambiguous. Advancing care this year reduces classical moral hazard next year. Delaying care this year increases classical moral hazard next year. The key parameters determining this choice are individual preferences for classical moral hazard and cost of retiming planned care.

The interaction between classical and timing moral hazard occurs even without deductible choice (or less frequent enrollment periods). When deductible choice is introduced, it creates an additional source of *ex post* dynamic selection. Even individuals with *ex ante* identical risk can become different over time due to the timing of shock realizations within the coverage period. They vary in the amount of care they reschedule and the degree of classical moral hazard they exhibit later. However, I find little evidence of dynamic adverse selection based on shock timing alone in terms of deductible choice in the next year.

Insurers cannot distinguish between the components of healthcare consumption at the individual level. However, they can observe the timing of health shocks within the coverage period, conditional on their realization. Based on my partial equilibrium framework, the timing of shocks provides information about the expected behavior in terms of deductible choice and moral hazard. The extent to which insurers can incorporate this information into premium pricing depends on the regulatory framework, particularly their flexibility in risk classification. In settings that allow some degree of risk classification, such as in the U.S., insurers may adjust individual premiums based on the timing of shocks to account for expected behavioral responses. In settings with community rating, such as Switzerland, moral hazard imposes an externality on premiums within the community. In general equilibrium, the effect on premiums may create a feedback effect on both types of moral hazard among consumers, and the deductible choice.

The analysis suggests several policy tools to address timing moral hazard. The first consideration is the coverage period. Timing incentives arise from the discrepancy between contracts, which are structured around calendar time, and the incentives generated by shocks, which operate on relative time. Introducing deductibles by illness episode can help realign these timelines. Additionally, breaking down insurance contracts into shorter periods with smaller deductibles may reduce the insured's ability to strategically shift planned care. However, this approach could also encourage care delays, as smaller deductibles are more likely to be exceeded, triggering classical moral hazard (Hong and Mommaerts, 2024). Secondly, differentiating deductibles by illness type or urgency could limit the ability to shift care that is unrelated to the initial shock. In Switzerland, for example, mandatory health insurance covers both emergent and non-emergent procedures under a single deductible. Whether this pooling is beneficial from a welfare perspective remains an important policy question. Lastly, restrictions on provider choice and the implementation of gatekeeping mechanisms could influence the insured's ability to time their care. The desirability of a given policy may depend on the value of covering planned compared to additional use, e.g. due to differences in health benefits and consumer willingness to pay.

8 Conclusion

In this paper, I introduce a new dynamic model of healthcare consumption with the possibility to time care under deductibles in health insurance. The model highlights that timing moral hazard influences classical moral hazard, as well as deductible choice. It yields a sufficient statistic for timing moral hazard net of the other two responses. The moment can be estimated in reduced form by exploiting the random timing of a large health shock within the coverage year. I consider high-deductible individuals who suffer an unexpected hospitalization, exceed their deductible, and have an incentive to advance care or delay care to the following year.

I find that the difference in planned care consumption in the year of the shock is nearly CHF 1,000 between individuals with shocks in February and November. This difference is substantial as it represents 10% of spending in the year of the shock for the February group. The response is mainly driven by individuals suffering temporary shocks, who do not expect to exceed their deductible again. My empirical estimates suggest that timing moral hazard is significant and of a similar magnitude to classical moral hazard. This affects the interpretation of the within-year price-elasticity of healthcare consumption. Policy tools such as coverage periods, differentiated deductibles, and restrictions on provider choice could influence timing moral hazard.

This paper emphasizes the importance of distinguishing between the intensive margin and the timing margin when analyzing moral hazard and evaluating insurance policies. Not only does the realization of risk (or its *ex ante* probability distribution) matter, but so does its timing, especially when individuals have the flexibility to decide when to address the risk and sufficient time to act. Dynamically evolving information asymmetries are particularly relevant when consumers face nonlinearities in their budget sets. Moreover, the interaction between multiple dimensions of asymmetric information can generate welfare-relevant effects.

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.

Albanese, Andrea, Matteo Picchio, and Corinna Ghirelli (2020). “Timed to Say Goodbye: Does Unemployment Benefit Eligibility Affect Worker Layoffs?”, *Labour Economics*.

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.

Brébion, Clément, Simon Briole, and Laura Khoury (2022). “Unemployment Insurance Eligibility and Employment Duration”, *Working paper.*

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.

Citino, Luca, Kilian Russ, and Vincenzo Scrutinio (2022). “Manipulation, Selection and the Design of Targeted Social Insurance”, *Working paper.*

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.

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.

Ericson, Keith Marzilli and Justin R. Sydnor (2018). “Liquidity Constraints and the Value of Insurance”.

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.

Gross, Tal, Timothy J. Layton, and Daniel Prinz (2022). “The Liquidity Sensitivity of Healthcare Consumption:Evidence from Social Security Payments”, *American Economic Review: Insights*, 4(2): 175–190.

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

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.*

Hendren, Nathaniel, Camille Landais, and Johannes Spinnewijn (2021). "Choice in Insurance Markets: A Pigouvian Approach to Social Insurance Design", *Annual Review of Economics*, 13(1): 457–486.

Hong, Long and Corina Mommaerts (2024). "Time Aggregation in Health Insurance Deductibles", *American Economic Journal: Economic Policy*, 16(2): 270–299.

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.

Kurihara, Osamu, Masamichi Takano, Erika Yamamoto, Taishi Yonetsu, Tsunekazu Kakuta, Tsunenari Soeda, Bryan P. Yan, Filippo Crea, Takumi Higuma, Shigeki Kimura, Yoshiyasu Minami, Tom Adriaenssens, Niklas F. Boeder, Holger M. Nef, Chong Jin Kim, Vikas Thondapu, Hyung Oh Kim, Michele Russo, Tomoyo Sugiyama, Francesco Fracassi, Hang Lee, Kyoichi Mizuno, and Ik-Kyung Jang (2020). "Seasonal Variations in the Pathogenesis of Acute Coronary Syndromes", *Journal of the American Heart Association*, 9(13) e015579.

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.

Manning, Willard G., Joseph P. Newhouse, Naihua Duan, Emmett B. Keeler, and Arleen Leibowitz (1987). "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment", *American Economic Review*, 77(3): 251–277.

Martínez, Isabel Z., Emmanuel Saez, and Michael Siegenthaler (2021). "Intertemporal Labor Supply Substitution? Evidence from the Swiss Income Tax Holidays", *American Economic Review*, 111(2): 506–546.

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

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

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 for Online Publication

Timing Moral Hazard under Deductibles in Health Insurance

Véra Zabrodina

This version: September 16, 2024.

A Theoretical Model

A.1 Marginal Price

The year-end price is defined as:

$$P_t = \begin{cases} 1 - \sum_{\tau=t+1}^{T} q_t(\tau) & \text{if } P_t^c = 1 \\ 0 & \text{if } P_t^c = 0, \end{cases} \tag{14}$$

where $P_t^c \equiv \frac{\partial C(c_t, m_t, R_t)}{\partial c_t}$ is the current marginal out-of-pocket price of the last unit of consumption in t and captures on which price segment the individual lies in the current period.

The probability of exceeding the deductible in period τ (i.e. the probability of the consumption in that period being larger than the positive remaining deductible), conditional on the information in period t is defined as:

$$q_t(\tau) \equiv \Pr(R_\tau > 0, c_\tau + m_\tau > R_\tau | R_t, c_t, m_t), \text{ where } \tau > t. \tag{15}$$

The intuition is that individuals who spend one CHF on healthcare in period t will spend one less CHF for care in period $\tau > t$. In other words, spending today decreases the expected price tomorrow. Once the deductible is exceeded, the price is zero, and the spot consumption optimization problem becomes static.

B Empirical Analysis

B.1 Data and Sample

Table B.1: Comparison between CSS Insurance enrollees and the Swiss population

	(1) Baseline sample	(2) Swiss population
Demographics[a]		
Age	50.3%	49.2%
Sex	52.5%	50.6%
Swiss nationality	76.8%	74.6%
Insurance plan[b]		
Monthly insurance premiums	4,300	4,360
Deductible CHF 300	18.3%	15.4%
Deductible between CHF 500 and 2000	8.6%	12.0%
Deductible CHF 2500	3.5%	4.2%
Other insurance plans[c]	69.6%	72.6%
Observations	982,003	

Notes: The table presents means for the baseline analysis sample from the CSS data, as well as population and health insurance statistics for the Swiss population for the year 2019. Premiums are in Swiss Francs (CHF).

[a] Figures condition on being aged between 19 and 90. Source for population statistics: Population statistics from the Federal Statistical Office.

[b] Figures condition on being over 19 years old. Source for population statistics: Statistics on Compulsory Health Insurance, Federal Office of Public Health (FOPH), Switzerland.

[c] Following the FOPH, the defining criterion for this category is to have an insurance plan with restricted choice, i.e. telemedicine, health maintenance organization, or family doctor (see Section 2).

B.2 Validity Checks for Shock Timing Randomness

This appendix provides evidence for the timing of the first shock being random in the sample of insured with a high deductible in the year of the shock. A shock is defined as the first hospitalization observed at least one year into the enrolment with CSS. The sample contains individuals who suffered this shock while holding a CHF 2,500 deductible.

The validity checks are based on the idea that, if shock timing is random, observable characteristics before the shock should not predict shock timing. Not rejecting the null of no differences provides support for no selection into treatment groups.

Single observable characteristics. I first check for differences in single observable characteristics across shock months. Figure B.1 shows that there are no systematic differences in observable demographic characteristics (gender, nationality, age), take up of accident insurance with CSS (which can be seen as a noisy measure of non-employment), as well as indicators for diagnoses inferred from claims for prescription drugs. There is only little seasonality in some characteristics (e.g. gender), but few point estimates are statistically significantly different from the pooled sample average.

Similarly for insurance plan characteristics, Figure B.2 shows no significant differences in monthly premiums and the share of standard plans without healthcare provider restrictions.

Attrition. I also test for systematic differences in attrition, which would capture either deaths or changes of insurer. Figure B.3 shows that the probability of leaving the sample two years after the shock, and the number of years observed do not vary across shock groups.

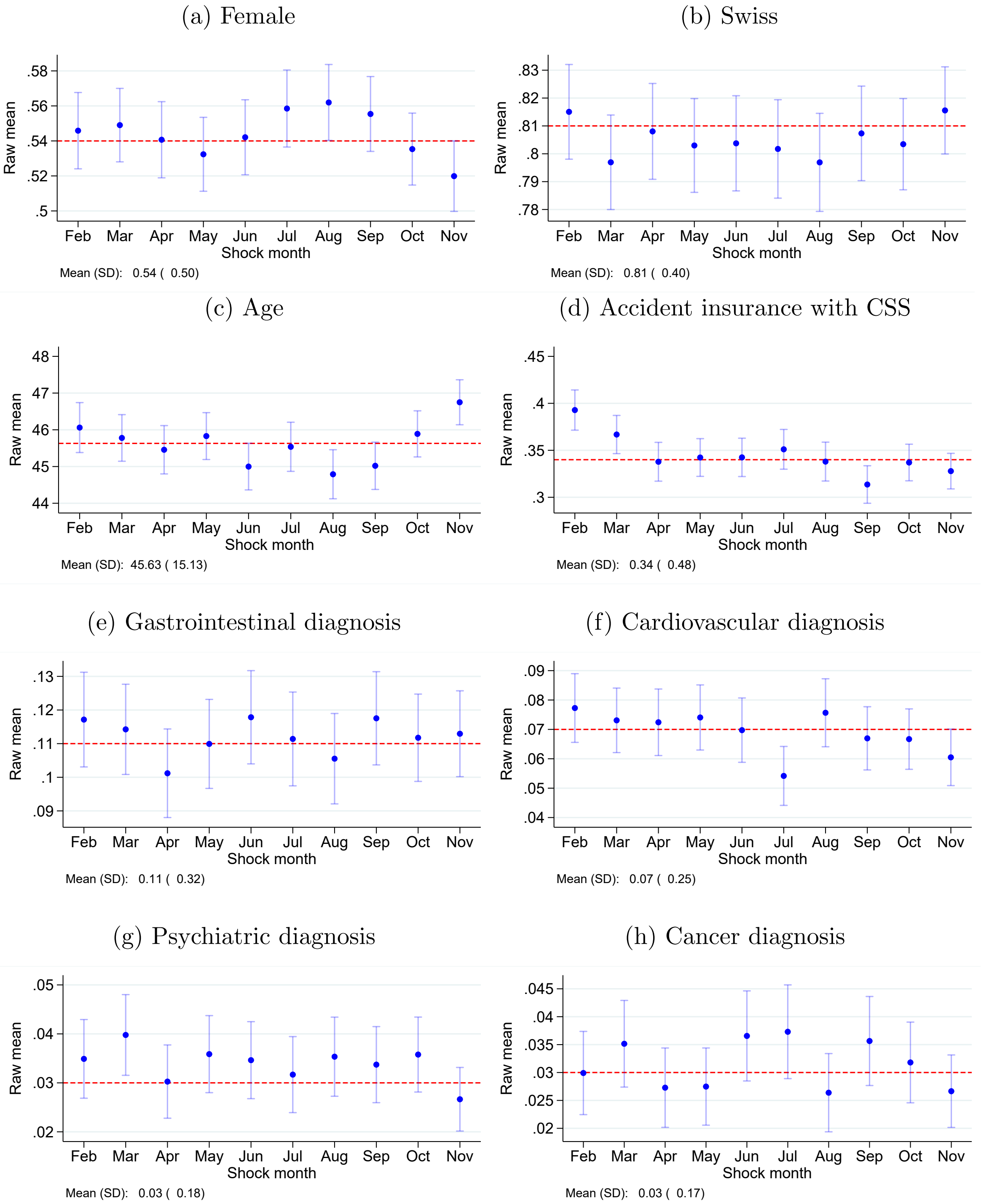
Propensity score for being in a given shock month. I then test for differences in combinations of characteristics, rather than single ones, in the spirit of propensity score matching with multiple treatment groups. For each shock month, I estimate a binomial probit model of the probability of being treated with a shock in a specific month versus any other month:

$$Pr(I_i(s)|X_i = x) = \Phi(\alpha + \sum_{k=1}^{K} \beta_k X_{k,i}) \tag{16}$$

where i denotes an individual, and $I_i(s) = \{1 \text{ if } S_i = s, 0 \text{ otherwise}\}$, $s \in \{2, \ldots, 11\}$. The covariates $X_{k,i}$ are measured at the end of the year before the shock and include age-gender group dummies, Swiss citizen, plan type, premiums, accident insurance, drug-based diagnoses, mover, year fixed effects, and past spending quartile. For each potential shock month, I compute and plot the estimated propensity scores separately for each actual shock group. I then perform Kolmogorov–Smirnov (KS) equality-of-distributions test using estimated propensity scores to check if individuals entering a given treatment group have selected sets of characteristics. I report the share of KS test that reject the null of equal distributions across all comparisons between actual shock months at the 5 percent level.

Figure B.1 reports the results of this exercise. The propensity score distributions have very strong overlap, with few outliers in the tails. The KS test is significant in 13 to 29 percent of cases. Generally, the group that actually has a shock in that month is slightly more likely to experience a shock in that month.

Figure B.1: Balancedness in observables across shock months



(a) Female

(b) Swiss

(c) Age

(d) Accident insurance with CSS

(e) Gastrointestinal diagnosis

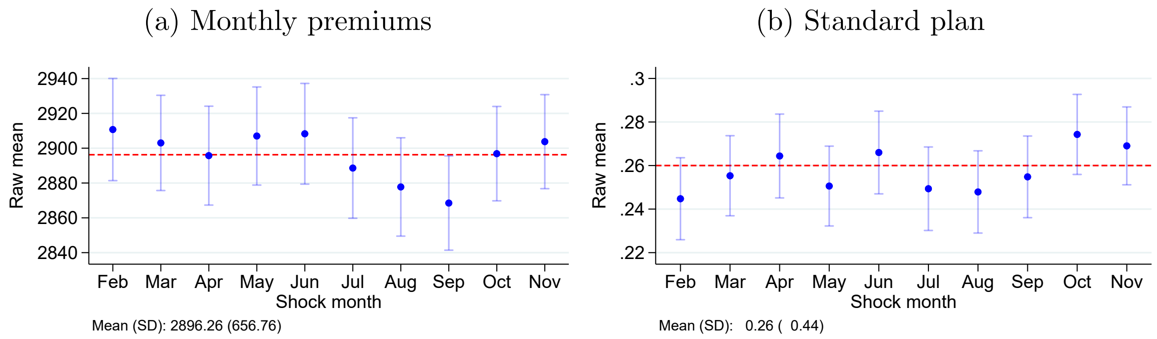
(f) Cardiovascular diagnosis

(g) Psychiatric diagnosis

(h) Cancer diagnosis

Notes: This figure presents raw averages across shock month groups. Observations are individuals from the main analysis sample. Diagnoses are inferred and aggregated into categories based on prescription drug claims. The red dashed line shows the sample average. Confidence intervals are at the 95% level based on robust standard errors.

Figure B.2: Premiums and spending across shock months

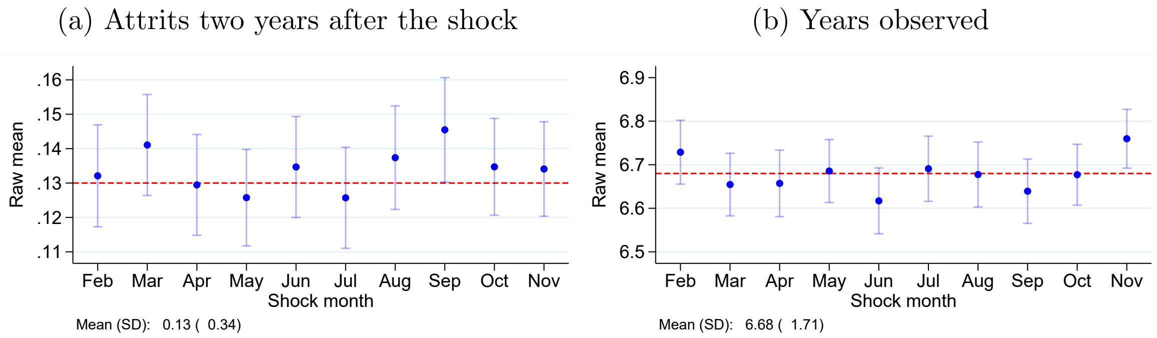


(a) Monthly premiums

(b) Standard plan

Notes: This figure presents raw averages across shock month groups. Observations are individuals from the main analysis sample. Total spending is in CHF. The red dashed line shows the sample average. Confidence intervals are at the 95% level based on robust standard errors.

Figure B.3: Attrition across shock months

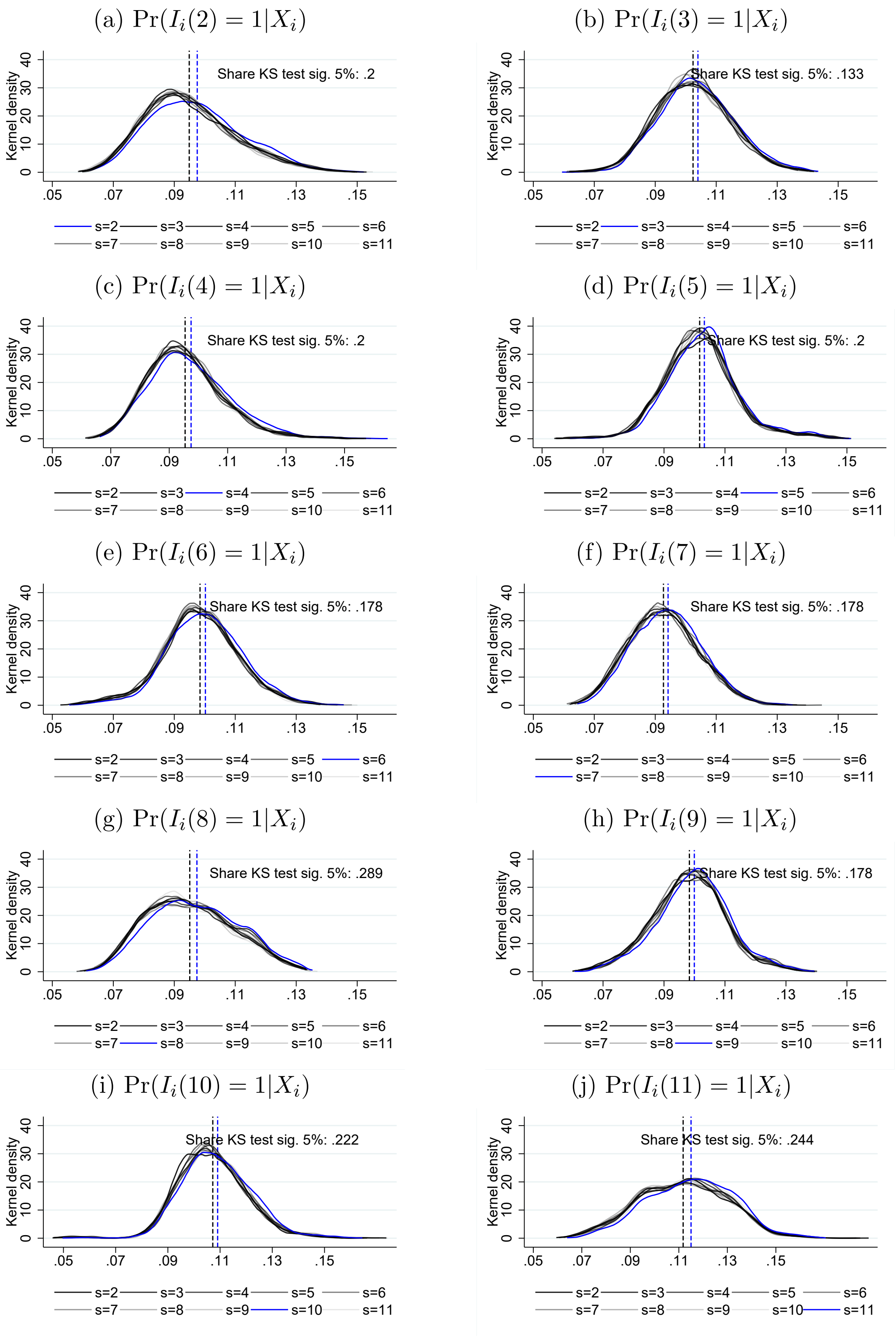


(a) Attrits two years after the shock

(b) Years observed

Notes: This figure presents raw averages across shock month groups. Observations are individuals from the main analysis sample. The red dashed line shows the sample average. Confidence intervals are at the 95% level based on robust standard errors.

Figure B.4: Distributions of the propensity to enter a given shock group



(a) $\Pr(I_i(2) = 1|X_i)$

(b) $\Pr(I_i(3) = 1|X_i)$

(c) $\Pr(I_i(4) = 1|X_i)$

(d) $\Pr(I_i(5) = 1|X_i)$

(e) $\Pr(I_i(6) = 1|X_i)$

(f) $\Pr(I_i(7) = 1|X_i)$

(g) $\Pr(I_i(8) = 1|X_i)$

(h) $\Pr(I_i(9) = 1|X_i)$

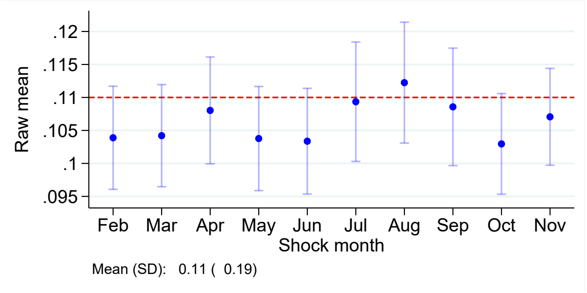
(i) $\Pr(I_i(10) = 1|X_i)$

(j) $\Pr(I_i(11) = 1|X_i)$

Notes: This figure presents propensity score distributions for the probability of entering a given shock month $Pr(I_i(s)|X_i = x)$, by actual shock month. Vertical lines denote the average for the group that actually had a shock in that month in blue, and all the other months in black. The distribution for the actual group is shown in blue.

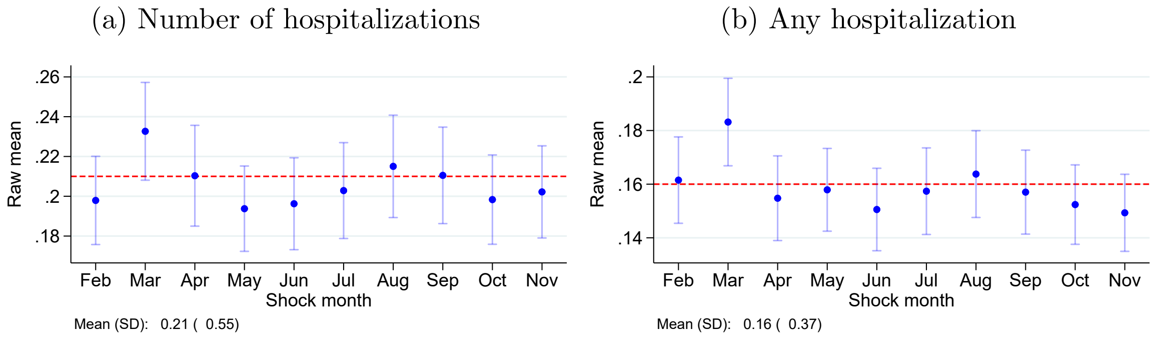
B.3 Prices, Deductible Choice, and Future Shocks

Figure B.5: Year-end marginal price in the shock year



Notes: This figure presents raw averages across shock month groups. Observations are individuals from the main analysis sample. The red dashed line shows the sample average. Confidence intervals are at the 95% level based on robust standard errors.

Figure B.6: Hospitalizations in the year after the shock


(a) Number of hospitalizations

(b) Any hospitalization

Notes: This figure presents raw averages across shock month groups. Observations are individuals from the main analysis sample. The red dashed line shows the sample average. Confidence intervals are at the 95% level based on robust standard errors.

B.4 Cumulated Differences in Spending

Cumulated difference between the shock and the deductible reset

$$\Delta \text{ShockReset}(s) = \sum_{k=0}^{11-s} \hat{\gamma}_k^s - \sum_{k=0}^{11} \hat{\gamma}_k^2 \tag{17}$$

Cumulated difference in year of the shock (year 1)

$$\Delta \text{Year1}(s) = \sum_{k=-s}^{11-s} \hat{\gamma}_k^s - \sum_{k=-2}^{10} \hat{\gamma}_k^2 \tag{18}$$

Cumulated difference in year after the shock (year 2)

$$\Delta \text{Year2}(s) = \sum_{k=12-s}^{23-s} \hat{\gamma}_k^s - \sum_{k=10}^{21} \hat{\gamma}_k^2 \tag{19}$$

Cumulated difference in both years

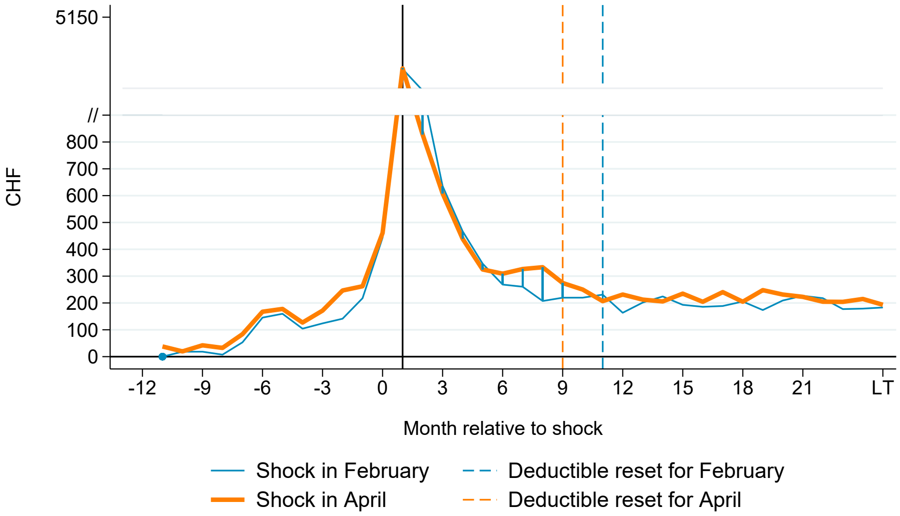
$$\Delta \text{BothYears}(s) = \sum_{k=-s}^{23-s} \hat{\gamma}_k^s - \sum_{k=-2}^{21} \hat{\gamma}_k^2 \tag{20}$$

Cumulated differences computed for all $s = 3, \ldots, 11$, relative to the February group $s = 2$.

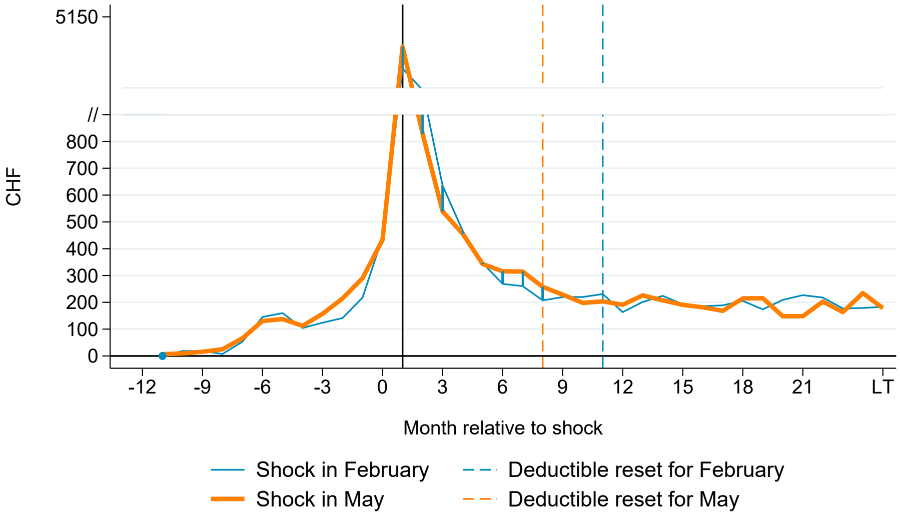
B.5 Additional Results

Figure B.7: Event study of healthcare consumption around the health shock – Additional treatment group comparisons

(a) Shock in February vs. April

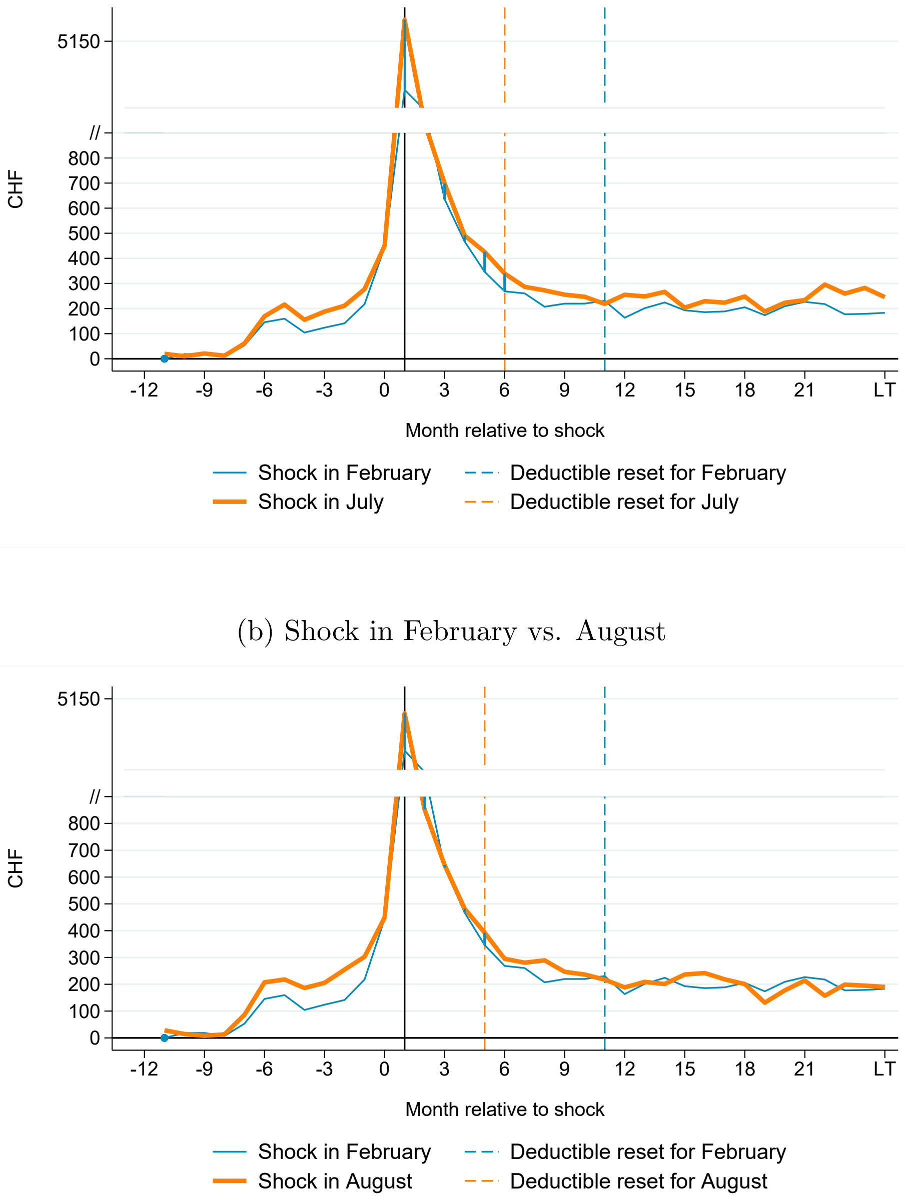


(b) Shock in February vs. May



Notes: The figure depicts the coefficient estimates on monthly treatment effects from the event study of healthcare consumption (as measured by total spending). It compares individuals with shocks in different months, for the main analysis sample of insured with a high deductible and the health shock defined as the first observed hospitalization. These effects are normalized to the average spending of the February group 12 months before the shock. The dashed lines indicate the last month before the year-end deductible reset after the shock. The last point estimate denotes the long-term effect (LT) of the shock, i.e. the average after 24 months.

Figure B.8: Event study of healthcare consumption around the health shock – Additional treatment group comparisons



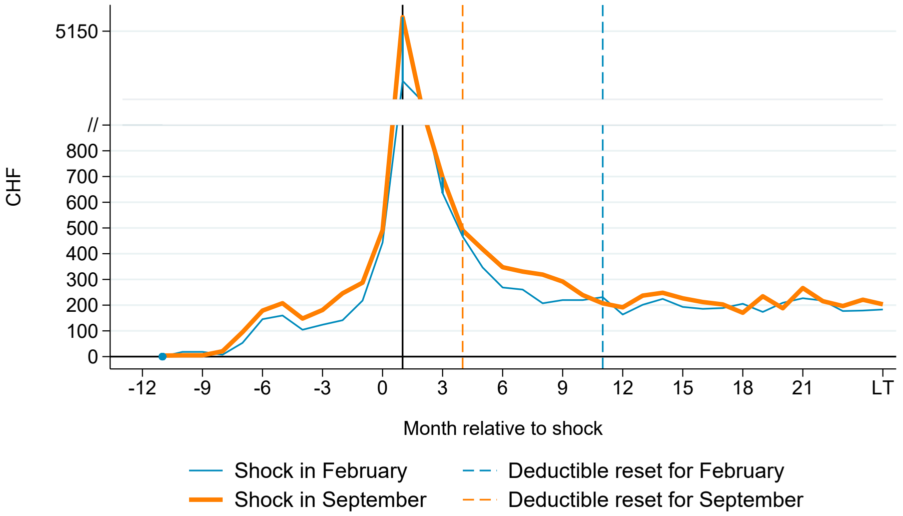
(a) Shock in February vs. July

(b) Shock in February vs. August

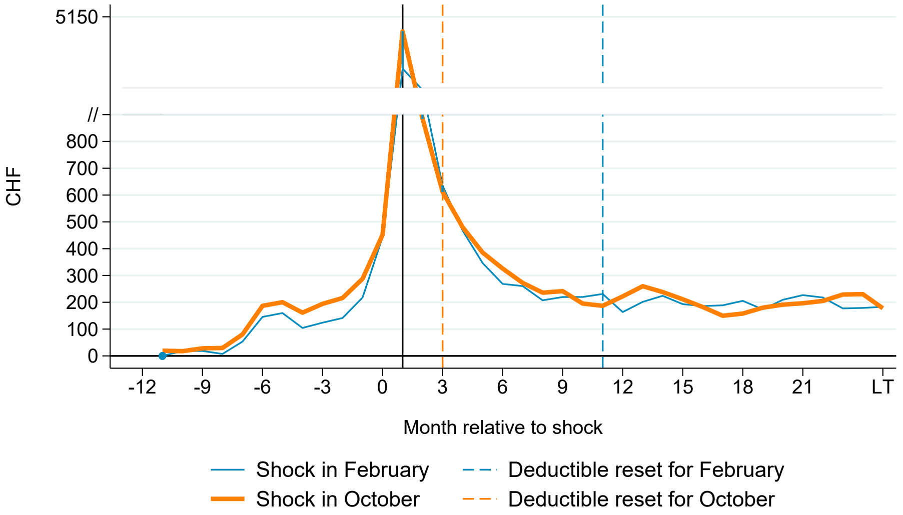
Notes: The figure depicts the coefficient estimates on monthly treatment effects from the event study of healthcare consumption (as measured by total spending). It compares individuals with shocks in different months, for the main analysis sample of insured with a high deductible and the health shock defined as the first observed hospitalization. These effects are normalized to the average spending of the February group 12 months before the shock. The dashed lines indicate the last month before the year-end deductible reset after the shock. The last point estimate denotes the long-term effect (LT) of the shock, i.e. the average after 24 months.

Figure B.9: Event study of healthcare consumption around the health shock – Additional treatment group comparisons

(a) Shock in February vs. September

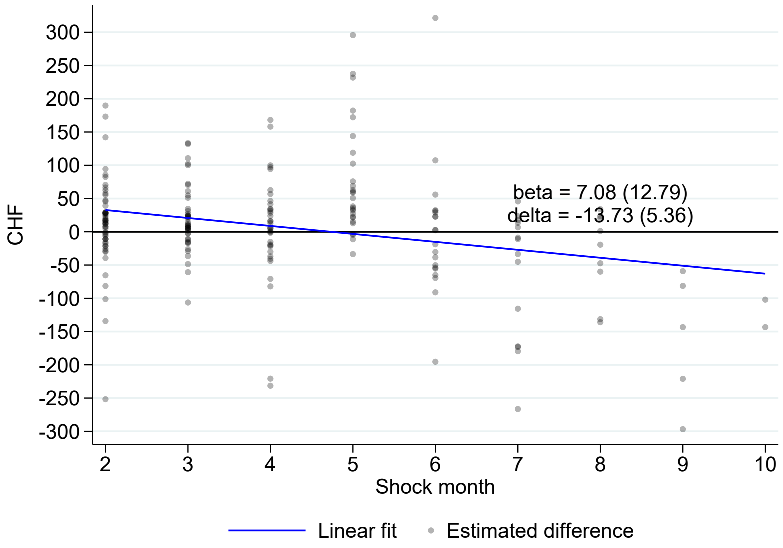


(b) Shock in February vs. October



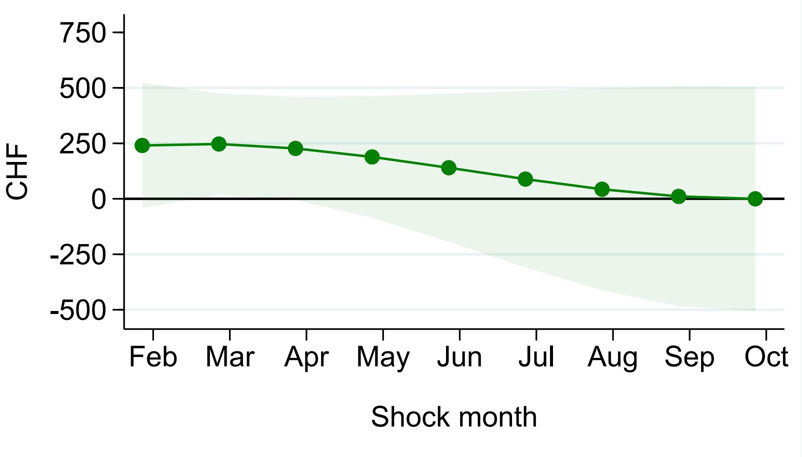
Notes: The figure depicts the coefficient estimates on monthly treatment effects from the event study of healthcare consumption (as measured by total spending). It compares individuals with shocks in different months, for the main analysis sample of insured with a high deductible and the health shock defined as the first observed hospitalization. These effects are normalized to the average spending of the February group 12 months before the shock. The dashed lines indicate the last month before the year-end deductible reset after the shock. The last point estimate denotes the long-term effect (LT) of the shock, i.e. the average after 24 months.

Figure B.10: Estimates of differences in total consumption



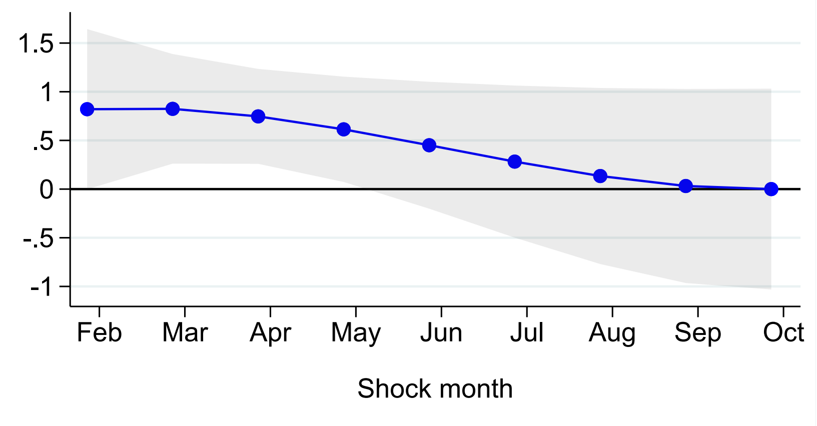
Notes: The figure displays estimates of $\Delta m(s, s')$ from the event study and the main analysis sample. The blue line shows the linear fit that serves to estimate the parameters for inferring the total timing moral hazard as in (12).

Figure B.11: Timing moral hazard for non-deferrable hospitalizations



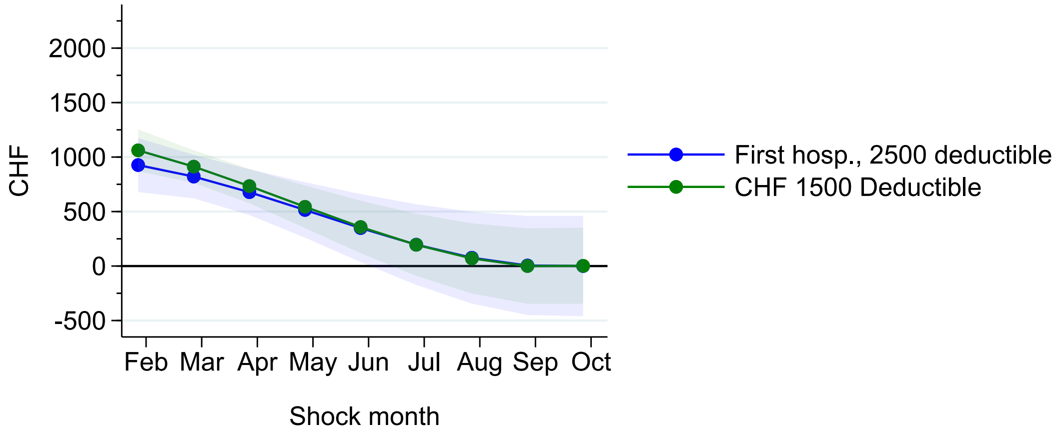
Notes: The figure presents estimates of the total yearly timing moral hazard response across shock months, predicted as in (13). The last shock month serves as a lower bound. The sample takes the subset of non-deferrable hospitalizations defined as invoice codes that occur at least 1/7 of the time on the weekend. Confidence intervals at the 95% level are based on bootstrapped standard errors with 49 replications, clustered at the individual level.

Figure B.12: Number of provider categories



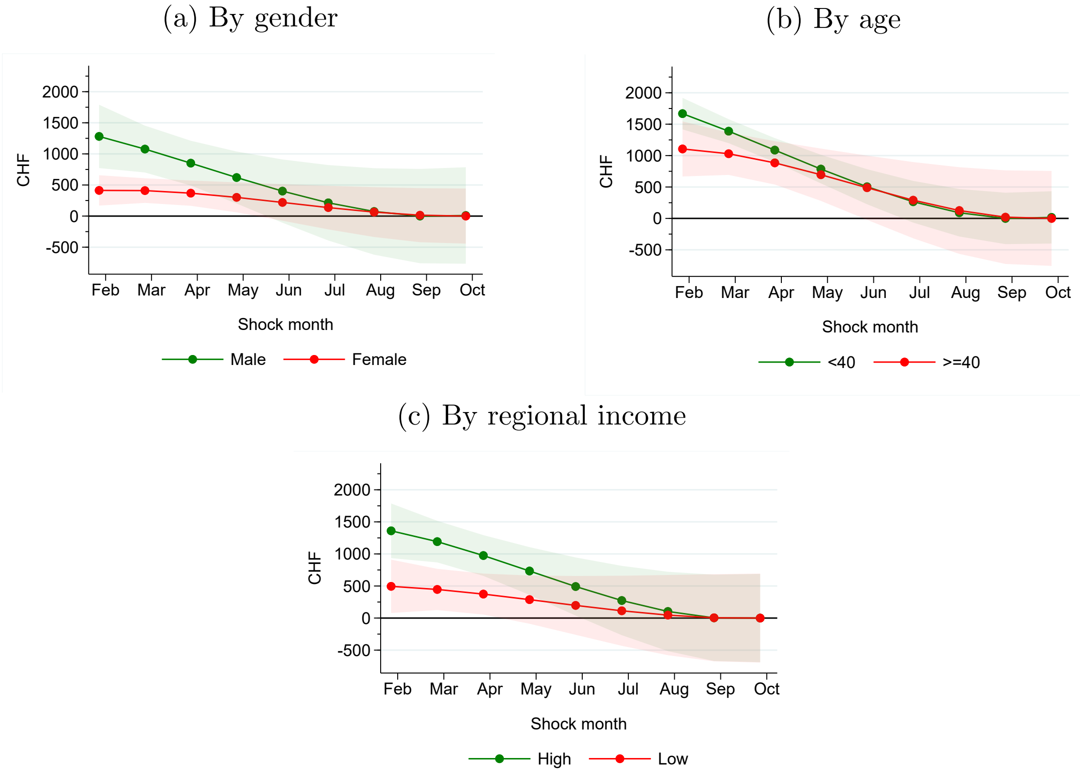
Notes: The figure presents estimates taking the number of different provider categories as the outcome of the timing moral hazard estimation procedure. The last shock month serves as a lower bound. Confidence intervals at the 95% level are based on bootstrapped standard errors with 49 replications, clustered at the individual level.

Figure B.13: Heterogeneity in total timing moral hazard by deductible level



Notes: The figure presents estimates of the total yearly timing moral hazard response across shock months for the samples of individuals with deductibles of CHF 1,500 and CHF 2,500 (the main definition) in the year of the shock. The last shock month serves as a lower bound. Confidence intervals at the 95% level are based on bootstrapped standard errors with 49 replications, clustered at the individual level.

Figure B.14: Heterogeneity in total timing moral hazard by individual characteristics


(a) By gender

(b) By age

(c) By regional income

Notes: The figure presents estimates of the total yearly timing moral hazard response across shock months, computed as in (13). The last shock month serves as a lower bound. Confidence intervals at the 95% level are based on bootstrapped standard errors with 49 replications, clustered at the individual level. Panel (a) splits the sample by gender. Panel (b) divides the sample by age. Panel (c) splits at median of income index at municipality level.