Timing Moral Hazard under Deductibles in Health Insurance[*]

Véra Zabrodina

Abstract

This paper evaluates whether individuals strategically time their healthcare consumption to reduce out-of-pocket costs, and the frictions they face in doing so. I set up a dynamic model of healthcare consumption in which individuals suffer a large health shock, exceed their deductible, and have an incentive to advance care from the following year. The model provides a sufficient statistic for timing moral hazard by comparing the consumption of individuals with shocks at random times within the coverage year. It also shows that advancing care mitigates classical moral hazard and adverse selection the following year. This insight highlights important trade-offs for insurance market design. In the context of mandatory health insurance in Switzerland, I find substantial timing moral hazard, though with strong dynamic frictions. The shorter the time horizon, the less care is advanced. The timing of health risk realizations has important implications for cost-sharing and insurance premiums.

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

This version: December 8, 2023.
Latest version

[*] University of Basel. Correspondence to: vera.zabrodina@unibas.ch. Website: verazb.github.io. 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, 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 on the timing of risk realizations has important implications for insurance markets. In healthcare, certain medical procedures are urgent, but many are not. This can generate a time lapse between the moment individuals realize a health need, and the moment they address it. The insurer cannot perfectly observe (or contract on) the date the health need realizes. It can only reimburse medical procedures based on the actual date of treatment. Individuals can use this private information to strategically time their healthcare consumption to reduce out-of-pocket costs (Cabral, 2017). For instance, under mandatory health insurance with a deductible, individuals have an incentive to time healthcare towards a year where they expect to exceed their deductible. Understanding the extent of *timing moral hazard*, its drivers, and its interactions with other sources of asymmetric information is crucial to design health insurance markets that sustainably balance risk protection and incentive costs.

In this paper, I develop a new approach to identifying timing moral hazard in health insurance. I also provide insights on how it relates to classical moral hazard and selection.[1] Timing moral hazard matters for the allocation of healthcare costs. Although it leaves the amount of healthcare consumption unchanged, it increases costs in the risk pool by shifting consumption across coverage periods with different marginal prices.[2] This can generate externalities by affecting insurance premiums. In contrast, classical moral hazard constitutes a price-response within a given coverage period. It raises costs because individuals increase their healthcare consumption when faced with a lower marginal price in this period (e.g. by getting further diagnostic tests).

The literature has largely focused on classical moral hazard, and has suggested that it can be limited by offering contracts with declining cost-sharing (Einav and Finkelstein, 2018). Motivated by this insight and growing healthcare spending, health insurance contracts with deductibles have become widespread, and a relevant case study. 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 1000 (Kaiser Family Foundation, 2021).

[1] I follow the conventional (ab)use of terminology in the context of health insurance and consider both behaviors to be types of *ex post* moral hazard, as I exclude any feedback effect on health risk. As the insured's actions (consumption) are observable, the information asymmetry stems from the insured having private information about their own price sensitivity and timing of procedures. See Einav et al. (2013) for a discussion of this terminology. In the context of healthcare, consumption (or demand) is measured one for one by spending, so I use these terms interchangeably (Kowalski, 2015).

[2] If the marginal price is constant across years, there is no incentive to retime.

While deductibles might limit classical moral hazard, they generate salient strategic timing incentives which evolve dynamically. Decisions about the amount and timing of consumption depend on expectations about future consumption within and across years. The insured have an incentive to advance care from next year once their deductible is exceeded, or to postpone care in anticipation of exceeding it next year. Importantly, timing moral hazard can occur even under mandatory insurance without the possibility to choose deductible levels. In practice, the insured can often switch deductibles every year. Timing moral hazard then becomes a driver of *ex post* selection based private information on realized risks. The timing of healthcare consumption determines spending, and thus coverage choices in given periods.

I begin by formulating a dynamic model of healthcare consumption, with the possibility to time planned healthcare consumption. A rational, forward-looking individual chooses their monthly healthcare consumption and yearly deductible. I focus on individuals with a high-deductible plan who suffer a large, unanticipated health shock, and are pushed into free care. After the shock, individuals have a potential incentive to advance care planned for next year to the shock year to reduce out-of-pocket costs.

The model provides a new sufficient statistic for pure timing moral hazard that circumvents *ex ante* selection issues and differences out classical moral hazard. Specifically, I show that the differences in healthcare spending across comparable individuals with shocks at random times within the calendar year measure differences in the amount of care advanced from the next year. Shock timing exogenously varies the incentive to advance care after the shock via cross-year differences in the marginal price. 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 spills over into the year after. Later shocks thus weaken the incentive and the time available to advance care towards the shock year. However, they are otherwise comparable in their shock-related health needs and preferences. If shock timing is random, the consumption mandated by the shock and classical moral hazard can be differenced out.

This result motivates a reduced-form identification strategy based on random variation in shock timing. I implement this strategy in the context of mandatory health insurance in Switzerland. I use individual-level claims data from the largest health insurance firm for the years 2012 to 2019, that are roughly representative of the Swiss population. Switzerland offers an attractive setting for this analysis. Health insurance contracts are highly regulated, and bear a single, yearly deductible, which the insured can freely choose without risk classification. They cover a broad scope of medical procedures due to illness. I run an event study of healthcare spending, where treatment effects are allowed to vary over time and across treatment groups defined by the calendar month of the first observed hospitalization. Using alternative shock definitions leaves the key qualitative conclusions unchanged.

I document substantial timing moral hazard. Individuals with shocks early in the year advance about CHF 1,000 worth of consumption (CHF 1 $\approx$ \$1). This represents approximately 10% of their consumption in the shock year, and a price-elasticity of -0.14 which is close to the benchmark for classical moral hazard of -0.2 (Keeler and Rolph, 1988). However, these results also suggest that the price-elasticity of healthcare consumption may overestimate classical moral hazard because it includes timing responses.

The less time individuals have, the less care they advance. In other words, timing moral hazard is subject to dynamic frictions. These may stem from supply-side constraints (e.g. scheduling of appointments, need for referrals), hassle costs, or behavioral and cognitive biases (e.g. wrong expectations about future health needs). The timing response is mainly driven by individuals with temporary shocks. These results point to coverage length and differentiating co-payment schedules across types of care as relevant policy tools to address timing moral hazard.

I then extend the model to shed light on how the incentives for timing moral hazard relate to other sources of asymmetric information. First, timing moral hazard can limit classical moral hazard if it reduces the probability of exceeding the deductible in years from which care was shifted. In my setup, advancing care means foregoing utility from classical moral hazard the year after. A higher utility from classical moral hazard decreases the propensity to advance care. This highlights an important policy trade-off between the two behaviors. Second, advancing care decreases selection into lower deductibles, as it decreases expected spending in the following year. This result mirrors the one in Cabral (2017), where deferring care generates *ex post* adverse selection. This comes on top of classical adverse selection on *ex ante* risk, as even *ex ante* comparable individuals may become heterogeneous over time due to differences in risk realizations, and the timing thereof. A large leeway for synchronizing healthcare consumption with coverage purchase threatens the existence of insurance markets. Empirically however, I do not find a link between advancing care and deductible choice. This suggests that the reduction in timing moral hazard over shorter horizons is mainly driven by a smaller amount, rather than share of strategic timers.

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.[3] These studies have adopted a static perspective by assuming that individuals only consider current or year-end prices, and implicitly rule out strategic timing across years. A recent body of papers have found evidence for individuals dynamically respond to evolving price incentives by, e.g. anticipating deductible resets, or increasing consumption after becoming eligible for coverage (Gerfin et al., 2015; Simonsen et al., 2021; Card et al., 2009). These studies do not separate the consumption response into timing and classical moral hazard.

Several papers provide direct evidence 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). They find no such responses among those who spend largely past the gap and have weaker incentives to shift. 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, where they compare individuals who hit the deductible with individuals under free care plans in the last month of the coverage year. Using data from the RAND experiment, they find that individuals with high deductibles have higher spending in the last month, suggesting that those individuals who hit 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. Using a structural modelling approach, the author finds that about 40% of individuals postpone deferrable dental care when incentivized to do so. The resulting *ex post* adverse selection, whereby individuals can easily delay care to purchase coverage, explains the largely-missing market for dental insurance.

This paper makes several contributions to this literature. First, it proposes a novel strategy to quantify timing moral hazard in reduced-form, in a literature that has mainly adopted fully-structural methods (Einav et al., 2015; Cabral, 2017). My sufficient statistics approach provides a clear characterization of the quantities estimated in reduced form, while imposing fewer conceptual and distributional assumptions. It does not require specifying all the primitives underlying healthcare consumption and timing decisions. Second, my theoretical model explicitly incorporates classical moral hazard and endogenous deductible choice, which allows studying how these behaviors relate to timing moral hazard.

[3] See Finkelstein (2014), Einav and Finkelstein (2018), and Gerfin (2019) for reviews and discussions of the literature on moral hazard in health insurance. Among recent reduced-form analyses, Kowalski (2016) uses injuries to family members in family-level insurance plans as an instrument for individual prices to estimate price elasticities across quantiles of the annual expenditure distribution. Ellis et al. (2017) instrument individual prices with employer-year-plan-month average cost shares to estimate price elasticities by type of medical service on a monthly basis.

Although important for insurance contract design, these links have received little attention, and complicate the separate identification of pure timing moral hazard. My approach overcomes this challenge, and adds to the literature on the links between behavioral responses to insurance stemming from different sources of asymmetric information (Einav et al., 2015; Cabral, 2017; Hendren et al., 2021).

Third, it focuses on a setting with a deductible where incentives are to advance rather than defer care. My findings suggest that individuals are not fully myopic and respond to future price incentives. Advancing indeed requires sophisticated agents who can anticipate and advance future non-emergent or planned procedures. Postponing may be less demanding as individuals can defer care as needs realize over time. This insight relates my study to the literature on dynamic price responses in healthcare consumption (Aron-Dine et al., 2015; Brot-Goldberg et al., 2017; Abaluck et al., 2018; Dalton et al., 2020; Klein et al., 2022). My dynamic framework also allows uncovering substantial dynamic frictions to advancing care. These results broaden our understanding of the role of dynamic, cross-year incentives in shaping healthcare consumption decisions. Finally, 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). I uncover heterogeneity in the amenability to retiming across types of care, with outpatient and drugs representing a large share of the response.

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 and robustness checks. It also discusses the magnitude and cost-sharing implications. Section 6 further explores the microfoundations of timing moral hazard and dynamic frictions. The final section concludes.

2 Institutional Setting

Mandatory health insurance in Switzerland is regulated at the federal level and offers several compelling features to analyze strategic timing behavior. First, 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. These regulations limit the possibility for insurers to underwrite strategically timed claims, or design contracts with constant cost-sharing (i.e. no timing incentives). Second, all mandatory health insurance contracts bear a single deductible chosen every year between CHF 300 (the default), 500, 1,000, 1,500, 2,000 and 2,500. The financial stakes are high (the median monthly household disposable income was at about CHF 4,500 in 2019), and strategic timing incentives salient. Third, mandatory health insurance covers a comprehensive range of medical services received due to illness, which allows studying timing moral hazard across different types of care. The theoretical model presented in the next section is tailored to these features.[4]

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 characteristics of the insurance plan. Figure 1 sketches this function for the lowest CHF 300 and highest CHF 2,500 deductibles. The maximum annual out-of-pocket spending 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 spending for healthcare spending 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.

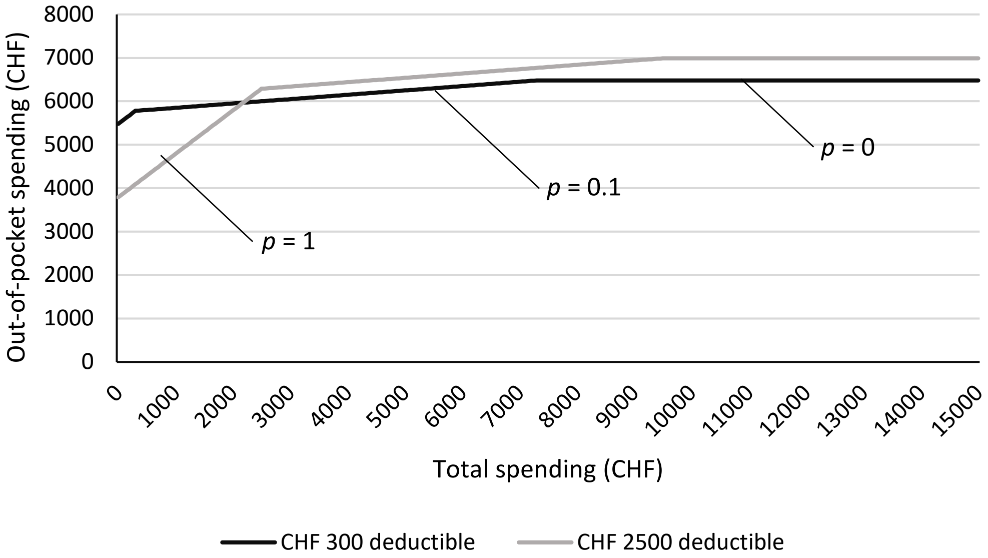
This setting creates incentives to strategically time consumption towards years where individuals expect to exceed the deductible. In particular, individuals can advance care from the following coverage year if they hit the deductible in a given year. Then might then choose a higher deductible the following year as expected consumption is decreased. Conversely, they may delay care if they expect to exceed the deductible in the next year. In that case, individuals can also purchase additional coverage by choosing a lower deductible.

The insured can switch plans or insurers until November 30th, with the new contract taking effect on January 1st.[5] Individuals can freely choose any insurance plan from

[4] The laws regulating mandatory health insurance have remained broadly unchanged since 2005. The setting closest to the Swiss one is the Netherlands, where health insurance is also mandatory and the choice of deductible lies between EUR 385 and 885.

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

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



Notes: The figure presents out-of-pocket healthcare spending as a function of total healthcare spending within a calendar year for the CHF 300 and CHF 2,500 deductibles. 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.

insurers operating within their canton (region) of residence without risk classification. Cantons provide means-tested insurance premium subsidies for low-income households. Apart from the deductible, individuals choose between the standard plan, which offers free choice of authorized healthcare provider, and alternative plans, which restrict this choice (i.e. health maintenance organizations, gatekeeping family physician, or telemedicine), but come with lower premiums.

Mandatory health insurance covers a broad range of ambulatory and inpatient services, as well as all drugs prescribed by a physician due to illness. It does not cover dental care, accidents or care reimbursed under supplementary insurance. Insurance for these categories is subscribed on top under specific regulations. 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. The prices of medical services covered by mandatory health insurance are fixed. Ambulatory services are reimbursed through a fee-for-service system, while hospitalizations are reimbursed through a prospective payment system with diagnosis-related groups (DRGs).

3 Theoretical Model

In this section, I present a dynamic model of individual healthcare consumption and deductible choice. I build on models that incorporate classical moral hazard in health insurance (Einav et al., 2013; Abaluck et al., 2018; Klein et al., 2022), and add the possibility for timing moral hazard by allowing individuals to shift the consumption of a fixed amount of planned care in time. The model serves three key purposes. First, it describes the timing incentives for individuals under mandatory health insurance with deductibles, as in the Swiss system (I discuss possible extensions below). Second, it provides a sufficient statistic for timing moral hazard that can be estimated in reduced form while retaining a clear interpretation. Third, it sheds light on how timing relates to classical moral hazard and deductible choice (Section 6).

3.1 A Model of Healthcare Consumption with Strategic Timing

Setup and timing. Take a rational, forward-looking individual who lives in two calendar (coverage) years, split into months $t = 1, \ldots, 24$.[6] Every month, the individual chooses their spot healthcare consumption c_t given their maginal year-end price P_t^e and health needs.[7] They do so by trading-off utility from health measured in monetary units, and money:

$$\max_{c_t \geq 0} u_j(c_t; \lambda_t, \omega, m_t, R_{j,t}) = v_c(c_t; \lambda_t, \omega) - v_m(m_t; \mu_t) - C_j(c_t; m_t, R_{j,t}) \tag{2}$$

Utility is quasilinear in money and additively-separable across periods. Utility from health $v_c(c_t; \lambda_t, \omega) = (c_t - \lambda_t) - \frac{1}{2\omega}(c_t - \lambda_t)^2$ is concave.[8] Individuals take as given nondiscretionary health needs λ_t, their price sensitivity ω, the consumption of planned care possibly shifted in time m_t, and the utility cost of strategic timing $v_m(\cdot)$. I discuss the rationale behind these key elements further below. The out-of-pocket cost function is $C_j(c_t; m_t, R_{j,t}) \equiv \min\{c_t + m_t, R_{j,t}\} + n_j$. It depends on total consumption in that period, and the deductible $D_j \in \{D_L, D_H\}$. The remaining deductible at the beginning of period t is denoted by $R_{j,t} \equiv \max\{0, R_{j,t-1} - c_{t-1} - m_{t-1}\}$. Individuals pay the full price of care out of pocket up to the deductible, and enter free care above.[9] They pay monthly premiums n_j.

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

[7] The marginal price in period t is the derivative of the out-of-pocket cost function with respect to consumption. For forward-looking individuals, only the marginal price in the last period of the coverage year (i.e. one minus the probability of exceeding the deductible) matters for consumption decisions. See Klein et al. (2022) for a detailed exposition.

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

[9] This formulation follows existing literature in assuming an exogenous income, and no saving and borrowing, see Klein et al. (2022) for a discussion. It gives rise to a non-convex annual budget set in

Types of healthcare consumption. The model classifies healthcare consumption into three categories based on the individuals' discretion over the amount and timing. Each bears specific implications for cost-sharing. First, *shock-induced needs* λ_t capture a minimal set of nondiscretionary medical services that have to be consumed and cannot be timed. Consider a patient who suffers a heart attack and requires emergency care and a series of follow-up treatments to survive. I allow for the possibility of serial correlation over time. Expectations about λ_t measure the individual's risk type, which is the source of classical, *ex ante* adverse selection in deductible choices.

Second, the time-constant preference parameter $\omega > 0$ determines *classical moral hazard*. This measures the additional consumption induced by a lower year-end marginal price (i.e. a higher probability of exceeding the deductible, with health modelled as a normal good, see Einav et al., 2013 for a discussion). In other words, it drives discretionary consumption that occurs because individuals do not cover the cost out of pocket (e.g. additional diagnostic tests). Importantly, this type of consumption is determined by the year-end price P_t^e in the current year only, and is independent of prices in other coverage years. It raises costs in the risk pool. Together, λ_t and ω determine monthly spot consumption, before any timing response.

Third, individuals have a given amount of *planned care* μ_t that can be shifted in time (e.g. a check-up).[10] The amount actually consumed m_t is endogenous and the key object of interest. It is determined based on relative prices across years (which depend on the shifting). There is an incentive to shift planned care towards years where the year-end price is lower. Importantly, the total over the two years is fixed, such that $\sum_{t=1}^{24} m_t = \sum_{t=1}^{24} \mu_t$. At any time t, the individual can consume no or at most all planned care.[11] Hence, timing moral hazard does not affect total healthcare spending, but shifts costs in time and onto the risk pool. The utility gain from timing moral hazard stems from savings in out-of-pocket costs. However, it possibly induces a utility cost $v_m(\cdot)$ as it requires active action from the individual. The utility cost is incurred in the period where care is consumed, while out-of-pocket cost savings occur in the period where consumption was initially planned (the individual has higher consumption of non-medical goods in that period, as there is no saving in this model).

health and residual income (consumption of other goods), which introduces the possibility of multiple solutions, but excludes bunching at the kink as in nonlinear price schedules with maximum benefits (as in e.g. Abaluck et al., 2018; Cabral, 2017; Einav et al., 2015).

[10] Individuals can only shift care that they know they (will) need. This planned care may or may not be related to the shock. Another way to rationalize intertemporal substitution in healthcare is with health capital (in the spirit of Grossman, 1972), whereby individuals invest into a durable 'stock' of health when prices are lower. Here, as in Cabral (2017), healthcare consumption does not translate into future benefits through greater health capital, i.e. it does not impact its marginal utility in other periods.

[11] Ignoring discounting and assuming that planned care consumption yields a fixed utility, it drops out of the optimization problem.

3.2 Identifying Timing Moral Hazard Using Random Shock Timing

I now explain how random variation in the timing of a large health shock serves to identify timing moral hazard in this framework.

Decisions after a large health shock. Assume all individuals suffer an unanticipated, exogenous shock (e.g. a hospitalization) in a random period $s \in S = \{1, \ldots, 12\}$ of year 1. This shock pushes their cumulated spending above the deductible, so that their price drops from 1 to 0 for the rest of the year. Furthermore, assume that they have a given high deductible in year 1 to circumvent selection effects at baseline. The deductible resets in $t = 13$, and the individual can freely choose a low or a high deductible for year 2. In the wake of the shock, the individual makes the following choices by backwards induction. They choose

1. Their monthly healthcare consumption for the rest of year 1;
2. When to consume care planned for year 2, and the deductible for year 2;
3. Their monthly healthcare consumption for year 2.

Consider the optimal healthcare consumption decision for the rest of year 1. Since the deductible is exceeded after the shock, all additional care is free for $t = s, \ldots, 12$. Individuals can all engage in classical moral hazard by increasing their consumption. Additionally, they have a possible incentive to engage in timing moral hazard by advancing care planned for the next year to the year of the shock. There is no incentive to postpone care.[12]

Figure 2 illustrates that the timing of the shock s within year 1 varies the incentive to advance care via two channels. First, it varies the time left until the year-end deductible reset: The later the shock, the less time to engage in timing moral hazard. Individuals with a shock in February (Panel a) have more time before the reset than those with a shock in June (Panel b). Second, shock timing shifts shock-related health needs (black dashed line) in calendar time, and onto year 2: The later the shock, the more likely are shock-induced needs to persist into the year after.[13] The June group suffers larger spillovers into year 2 than the February group. Moving forward, I emphasize where s

[12] By focusing on individuals with a high deductible, I shut down incentives to advance care before (or in the absence of) the shock, as individuals initially expect to end year 1 below the deductible. 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. However, it remains possible that these individuals choose a low deductible in year 2 if total planned spending is high enough, even without the shock. Under random shock timing, individuals are comparable on average in all of these dimensions.

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

generates heterogeneity across individuals.

Taking first-order conditions, optimal spot consumption then equals nondiscretionary needs plus any classical moral hazard spending $c_t^*(s) = \lambda_t(s) + \omega$. Importantly, any decisions regarding year 2 do not affect c_t in year 1. This comes from the additively-separable structure of the utility, where spot and planned consumption only influence each other through out-of-pocket costs. This assumption enables identification, and is reasonable if one sees planned care as fixed spending that is not tied to any spot consumption, and can be consumed at any time within the two years. Any advanced care comes on top, such that total healthcare spending in $t = s, \dots, 12$ is given by

$$h_t^*(s) = \lambda_t(s) + \omega + m_t^*(s) \tag{3}$$

The insurer (and the researcher) only observes $h_t^*(s)$, and not the single components.

I incorporate dynamic frictions in timing moral hazard by allowing the utility cost of timing $v_m(\cdot)$ to depend flexibly on shock timing. This captures any drivers of heterogeneity in the magnitude of the timing response as a function of s (e.g. healthcare supply constraints, hassle costs). I do not impose further assumptions on its shape. I discuss possible microfoundations for such frictions in Section 6.1.

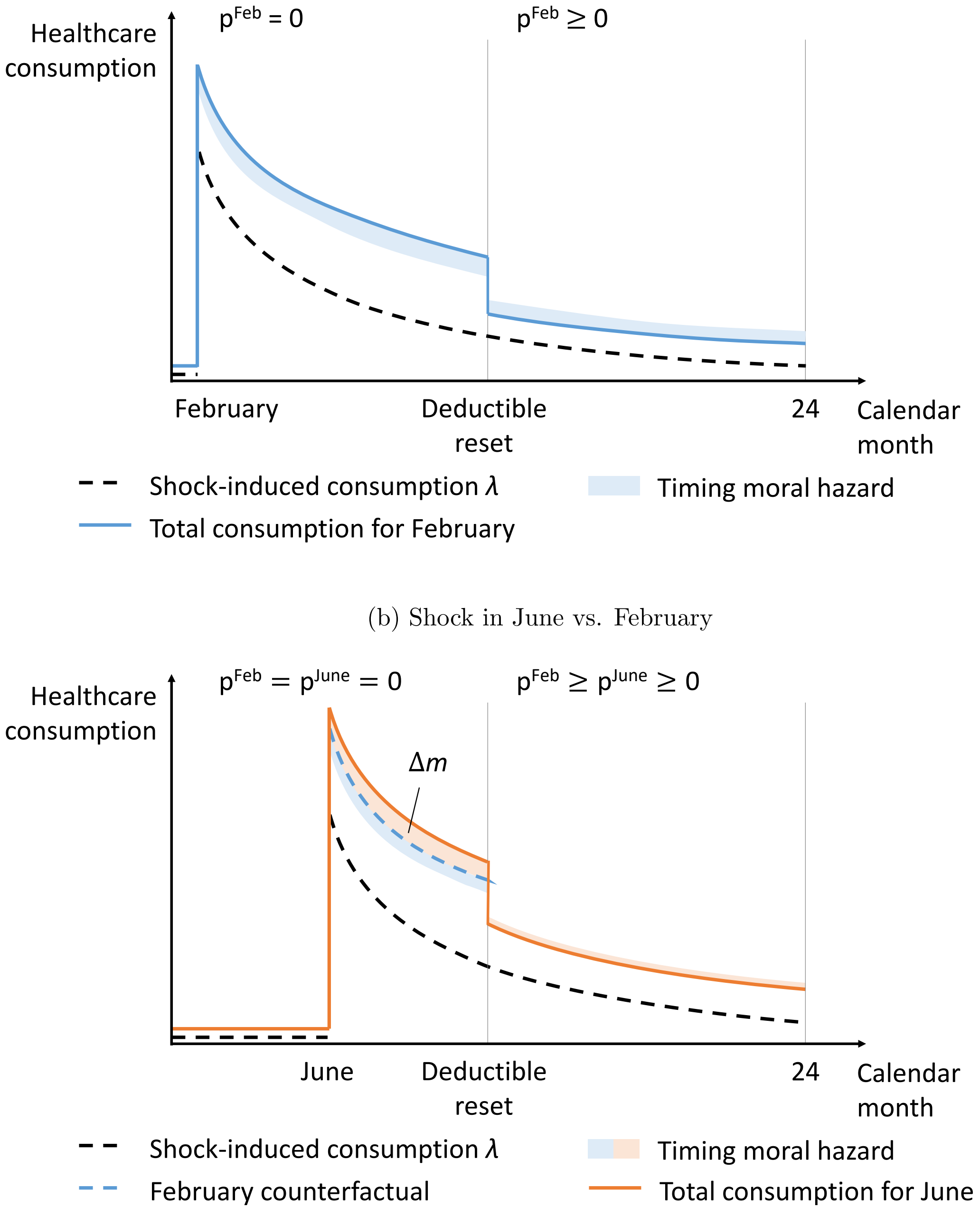
Differencing out λ_t and ω. If shock timing is random, it is independent of any individual characteristics, e.g. in preferences (in particular ω), baseline health, and other determinants of insurance plan choice. In other words, individuals differ only in the timing of their risk realization. Empirically, this makes for a more credible comparison than between individuals with and without a shock. This assumption implies that shock-induced consumption is the same on average, i.e. $\lambda_k(s) = \lambda_k(s')$ for all $s, s' \in S$.[14] Classical moral hazard consumption ω is also the same every month until the deductible resets, since individuals are all in free care. This means that individuals across s have the same spot consumption in time relative to the shock, and so until the year-end deductible reset. Hence, comparing healthcare consumption across shock groups allows teasing out differences in pure timing moral hazard, while differencing out shock-induced needs and classical moral hazard.[15]

Specifically, take the differences in healthcare consumption month-by-month in relative time between the shock and the deductible reset. This quantity identifies the difference

[14] I focus on the incentives triggered by the first shock, but the framework allows for repeated shocks in relative time.

[15] In Figure 2, classical moral hazard is comprised in the unshaded area between the nondiscretionary spending path and the timing moral hazard, but cannot be identified separately without further assumptions. As soon as the deductible reset is crossed, individuals are no longer comparable, because consumption in year 2 is confounded by differences in the chosen deductible and classical moral hazard. In other words, individuals face different, endogenous prices in year 2.

Figure 2: Dynamics of healthcare spending 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 spending dynamics in relative time. It depicts exemplary healthcare spending 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 spending. The dashed blue line illustrates the use of the mid-year's group spending as a counterfactual. The area Δm marks the difference in timing moral hazard, i.e. care shifted from year 2 to year 1 following the shock realization.

in the amount of care advanced from year 2 to year 1 (shaded orange area in Panel (b) of Figure 2). Let $\Delta h_k(s, s') \equiv h_k(s') - h_k(s)$ for $s, s' \in S$ and $s' > s$ in relative month $k = 1, \ldots, 13 - s'$ after the shock, and $\Delta s \equiv s' - s$. Plugging in (3) gives a sufficient statistic that approximates the derivative of timing moral hazard as a function of shock timing $m_k^{*'}(s)$:

$$\Delta h_k^*(s, s') = \lambda_k(s') - \lambda_k(s) + \omega - \omega + m_k^*(s') - m_k^*(s) \tag{4}$$

$$= m_k^*(s') - m_k^*(s) \tag{5}$$

$$\Rightarrow \frac{\Delta h_k^*(s, s')}{\Delta s} = \frac{\Delta m_k^*(s, s')}{\Delta s} \tag{6}$$

Notice that the difference is sufficient but not necessary to establish the existence of this response, as a null difference may mask identical monthly $m_k^*(s)$. Furthermore, the approach does not yield a prediction about the sign of the derivative, i.e. how timing responses depend on shock timing.[16]

The final object of interest is total timing moral hazard over the shock year, rather than its monthly derivative. Making a functional form assumption for $m_k^*(s)$ allows integrating up the derivative and obtaining the total yearly response. This parametrization is required as the system in differences is otherwise under-determined. In Section 4, I specify it based on the data. Timing moral hazard can then be quantified without fully specifying the underlying primitives, or estimating shock-induced needs and classical moral hazard.

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, 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

[16] In the example in Panel (b) of Figure 2, the February group has lower consumption if the advanced care is the same as June but spread over a longer time interval between the shock and the reset. In that case, $\Delta m_k^*(2, 6) < 0$, 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, $\Delta m_k^*(2, 6) > 0$, i.e. the orange line lies below the blue.

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.[17] 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.[18]

The data further include individual demographic information (gender, age, nationality, language used in administrative correspondence), insurance plan characteristics (premiums, deductible, type of plan, start and end dates of enrolment), and 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.

4.2 Reduced-Form Estimation

I now estimate timing moral hazard in reduced form based on the model. Using the assumption of random shock timing, I adopt a multiple treatment framework, where the

[17] The deductible structure incentivizes the filing of all claims, as opposed to, e.g. a maximum benefit. Anecdotal evidence suggests that unfiled claims may happens for small amounts such as drugs. Since the analysis focuses on individuals who exceed their deductible due to large shocks, this is unlikely to be a concern.

[18] The month as a time unit balances the trade-off between statistical power and the informativeness 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)	(2)	(3)	(4)
	All	High deductible	High deductible and health shock	Low deductible
Demographics				
Age	50.97 (18.09)	42.22 (13.40)	47.65 (15.24)	56.23 (18.67)
Female	0.52	0.41	0.50	0.58
Swiss	0.77	0.76	0.84	0.76
Insurance plan				
Monthly premiums	3,988 (1,106)	2,870 (654)	3,063 (843)	4,612 (838)
CHF 2500 deductible	0.20	1.00	0.84	0.00
Standard plan	0.32	0.35	0.39	0.30
Accident insurance	0.49	0.26	0.37	0.63
Spending and prices				
Total out-of-pocket spending	4,594 (1,312)	3,465 (1,194)	4,203 (1,453)	5,214 (969)
Total annual spending	4,014 (8,311)	1,039 (3,146)	2,955 (6,130)	5,896 (9,913)
Exceeded deductible	0.56	0.10	0.35	0.84
Cost sharing	0.40	0.55	0.60	0.29
Insured-years	6450240	1310801	101343	3790419

Notes: The table presents means and standard deviations (in parentheses) for samples of insured-years. Column (1) presents figures for all insured-years in the baseline sample. Column (2) presents figures for insured-years with the high deductible of CHF 2,500. Column (3) 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 (4) 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.

calendar month of the first shock $S_i \in \{2, ..., 11\}$ defines mutually-exclusive treatment groups.[19]

Main sample and shock definition. The main definition of a shock is the first hospitalization, at least one year into the observation window. This definition ensures that individuals exceed their deductible and individuals enter free care (or the co-payment region) for the rest of the year. Over 90% of hospitalizations cost more than the highest CHF 2,500 deductible in my data. Following the model, the main analysis focuses on high-deductible individuals, for whom a hospitalization is most likely unanticipated and therefore timed randomly. Rational individuals who expect a costly hospitalization would choose a low deductible, especially if they are risk averse and prone to moral haz-

[19] I exclude individuals who have a shock in January and December in the empirical analysis to avoid turn-of-the-year spending and billing effects. 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.

ard (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. I present robustness checks with alternative definitions of the shock in Section 5.6.

Table 1 provides descriptive statistics for the baseline sample (column 1), as well as subsamples by deductible level at the insured-year level. High-deductible individuals (column 2) represented 30% of the yearly population of insured in 2019. As expected, the high-deductible sample without a shock is younger, and has a lower share of women than the low-deductible sample (column 4). These characteristics are strongly correlated with health status and healthcare consumption. Column (3) contains the insured-years with a high deductible and the first hospitalization, i.e. the main shock sample. It generally lies in-between in terms of average characteristics, insurance plan choices, and prices. It is on average 48 years old, 50% female, and 84% Swiss. Its premiums and other spending outcomes are slightly higher than for high-deductible insured-years without a shock. These figures suggest that the main sample includes individuals with relatively low baseline risk who suffer an adverse health event. The analysis thus relies on a selective but highly-relevant sample of high spenders with strong timing moral hazard incentives, as discussed further in Section 5.6.

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 characterises the full treatment path. The outcome of interest is healthcare consumption h_{it} for individual i in calendar month t. I estimate the following event study at the insured-month level:

$$h_{it} = \sum_{s=2}^{11} \mathbf{1}\{S_i = s\} \left(\sum_{k=-11}^{24} \gamma_k^s \mathbf{1}\{t - E_i = k\} + \gamma_{24+}^s \mathbf{1}\{t - E_i > 24\} \right) + \sigma_t + \nu_i + \zeta X_{it} + \varepsilon_{it} \tag{7}$$

The coefficients γ_k^s flexibly capture the effect of shock timing s on spending in relative month $k \in \{-11, ..., 24\}$, and so for individuals with a shock in month s relative to those with a shock in February. Relative shock time is normalized to the pre-shock period $k = 0$. Long-term level effects are captured by γ_{24+}^s. The results are not sensitive to shortening these horizons. Dynamic treatment effects inform on shock anticipation and persistence, and so across treatment groups, 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.[20] This specification stems directly from (5), and allows computing estimates for the derivative of timing moral hazard as follows across 210 comparison pairs $\{s, s'\}$ for $s, s' \in S$ and $s' > s$:

$$\Delta \hat{h}_k(s, s') = \hat{\gamma}_k^s - \hat{\gamma}_k^{s'} \tag{8}$$

For instance, $\Delta \hat{h}_k(2, 3)$ provides an estimate of the difference in timing moral hazard between the February and March shock groups, 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 (homogeneous across treatment groups), 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).[21] Individual fixed effects ν_i subsume any time-invariant individual characteristics (e.g. gender, preferences, education, baseline health) that determine baseline healthcare 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 the derivative to the total response. I now show how to integrate up the derivative estimates to obtain the total timing moral hazard response. The negative linear relationship between the estimates $\Delta \hat{m}_k(s, s')$ and shock timing (see Appendix Figure B.11) suggests that the integral $m(s)$ is a quadratic function of s:

$$m(s) = \alpha + \beta s + \delta s^2 \tag{9}$$

$$\Rightarrow m'(s) = \beta + 2\delta s \tag{10}$$

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

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

where $m(s)$ is allowed to vary across groups due to heterogeneous retiming utility costs or dynamic frictions, but the amount of advanced care is assumed to be evenly allocated over year 1. Comparing monthly healthcare consumption across all shock groups pairs from February to November in relative months after the shock as in (8) yields 210 data points for $m'(s)$. I run the following regression to estimate $(\hat{\beta}, \hat{\delta})$ that most closely predict these data points:

$$\Delta \hat{m}(s) = \beta + 2\delta s + \epsilon \tag{11}$$

where regressors are demeaned. To compute a lower bound for $\alpha = \underline{\alpha}$, I use $(\hat{\beta}, \hat{\delta})$ and the fact that there is only an incentive to advance care in this setup, i.e. $m(s) \geq 0$. I use these parameter estimates to predict total timing moral hazard $M(s)$ over the shock year as follows:

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

In a frictionless case, individuals would advance the same total amount on average regardless of shock timing.

4.3 Identifying Assumptions

The identification of timing moral hazard relies on the timing of the health shock being exogenous. Defining the health shock thus requires particular care, as it determines who enters the sample, and which treatment group they belong to. Under the ideal shock, the composition of the population should remain stable throughout the year, such that treatment groups provide valid counterfactuals for each other.[22] In other words, there are no time-varying unobservable factors that jointly influence healthcare consumption and the probability of having a shock in a given calendar month (possibly conditional on observable characteristics and seasonality). The ideal shock may still display seasonality in its probability of occurring. Even heart attacks, which arguably cannot be timed, are more likely to occur in winter (Kurihara et al., 2020). Defining the shock as the first hospitalization has the advantage of maintaining generality and statistical power.

Individuals with high deductibles are unlikely to have strategically timed a hospitalization to the year where they suffer it (otherwise they would have taken a low deductible).[23]

[22] 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, which appears plausible in this setting.

[23] 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,

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. Comparability is not guaranteed by knowing the exact nature of the shock, as any shock may unobservably differ in their severity.[24]

In Appendix B.2, I conduct a battery of checks to support that there is no systematic selection into the month of the first hospitalization. Individuals do not observably differ in terms of demographic characteristics, drug-based diagnoses, as well as insurance plan. They also do not differ in the probability of attrition. Furthermore, I explore the composition of healthcare consumption at and before the shock to reassure that the shock is comparable across groups in terms of nature and severity. The shock mainly consists of inpatient spending, while anticipatory spending comprises outpatient physician or specialist visits, as well as tests and imaging. The absence of differential pre-trends under the main definition indicates that individuals do not (differentially) adjust their healthcare consumption to select into specific months (see Figure 7 and Appendix B.5). This excludes health deteriorations whereby more severe patients are more likely be hospitalized early in the year and consume more care. Controlling for compositional differences in time-varying individual characteristics does not alter the results. I present robustness checks with alternative shock and sample definitions in Section 5.6.

5 Results

5.1 Dynamic Healthcare Consumption Patterns

This section provides evidence on 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 starts increasing about 6 months before the shock. The increases are however small and would not lead to exceeding the deductible on average. The patterns support the assumption that treatment groups do not differ systematically in their pre-shock consumption.[25] Importantly, the approach is still valid if the beginning of the health deterioration that leads to the shock is random and its magnitude 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 stabilises roughly one

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).

[24] Specific shock definitions lead to small sample issues, and are typically more likely to affect older individuals, most of whom have a low deductible and no incentives to retime.

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

year after the shock, with a long-term effect of around CHF 200 per month, i.e. CHF 2,400 per year, which close but not above to 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 dynamic incentives. Finally, these spending dynamics support the binning choice for lags and leads in the event study, as past and future spending dynamics mainly differ between the shock and the deductible reset. This provides a first indication of timing moral hazard.

I provide additional evidence of how the shock timing drives differences in cumulated consumption in the year of the shock and the one after. 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 post-shock 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 post-shock 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 1100 between February and November groups. Taking both years together in Panel (d), larger consumption in the post-shock 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 spillovers, but are also suggestive of timing moral hazard via shifts in consumption from year 2 to year 1.

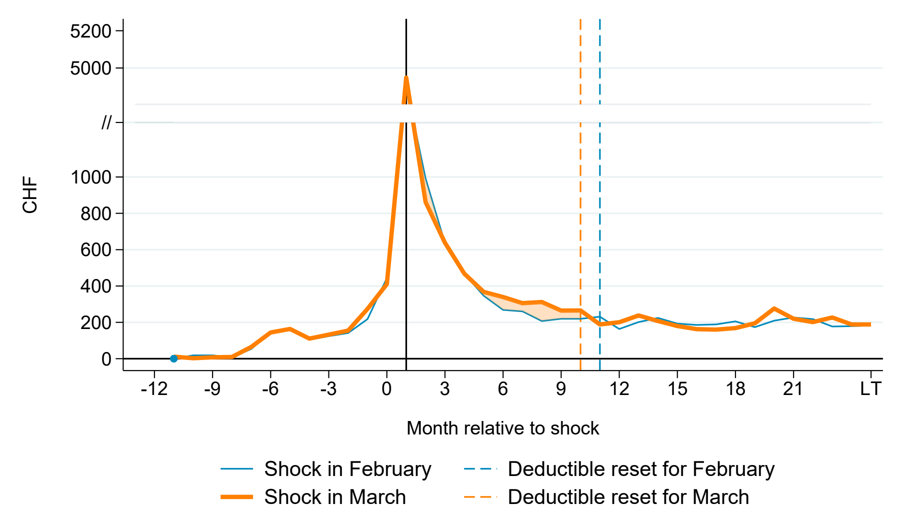
Appendix Figure B.6 shows the marginal price at end of the shock year across shock groups. Year-end prices in the shock year (panel a) are at 0.1 on average, i.e. in the 10% co-payment region, and not significantly different. This result suggests that incentives between the shock and the reset are aligned across groups. It also supports that dichotomizing the price structure is a reasonable approximation for the Swiss setting.

5.2 Estimates of Timing Moral Hazard

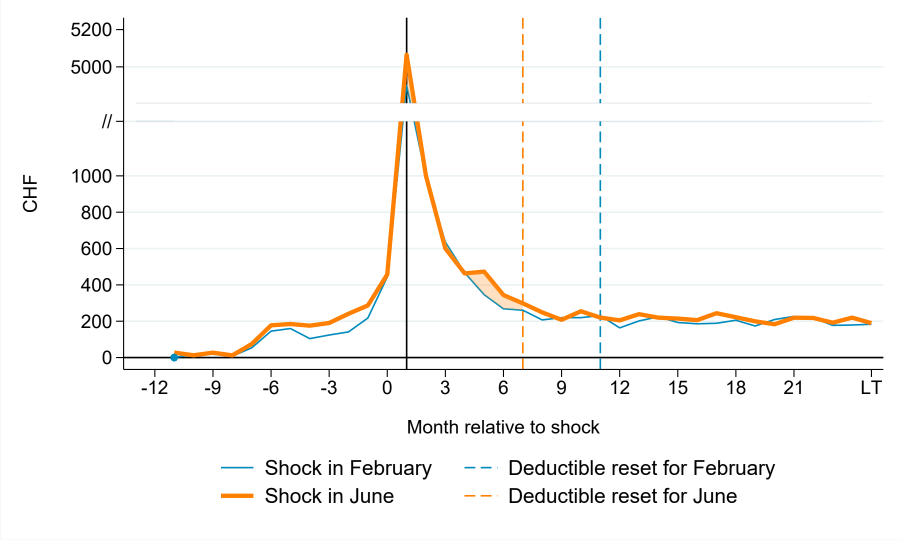
I begin by presenting figures for the estimates of monthly differences in total healthcare spending $\Delta\hat{h}_k(s, s')$ from the event study. These differences across shock comparison pairs yield 210 point estimates for the derivative of timing moral hazard $m'(s)$. The

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

(a) Shock in February vs. March

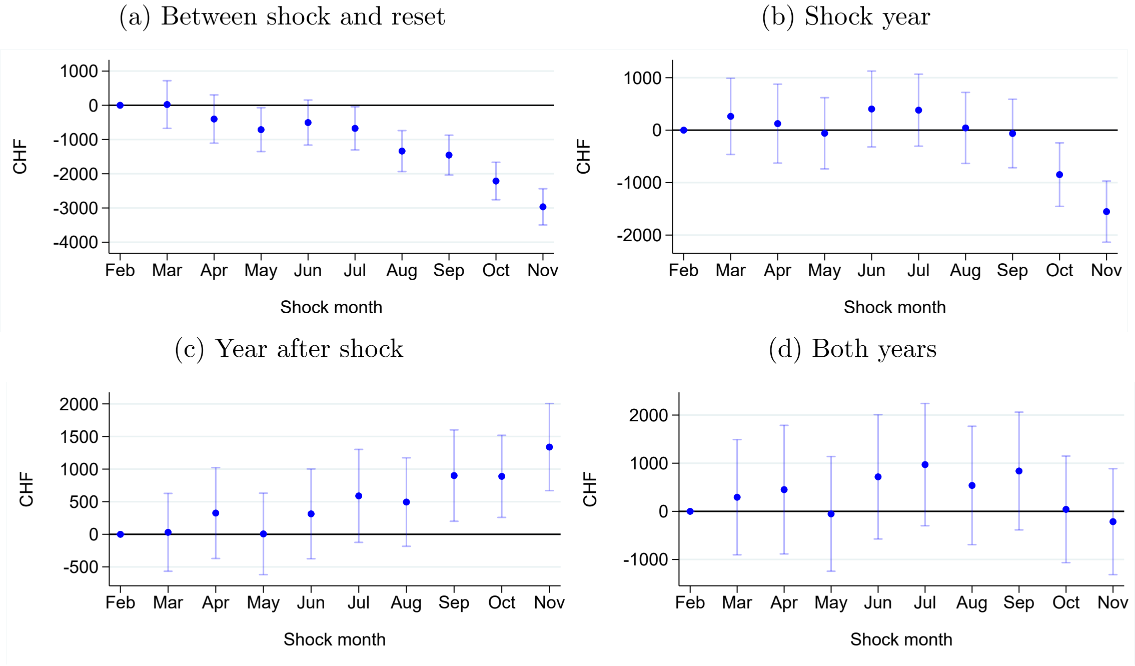


(b) Shock in February vs. September



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



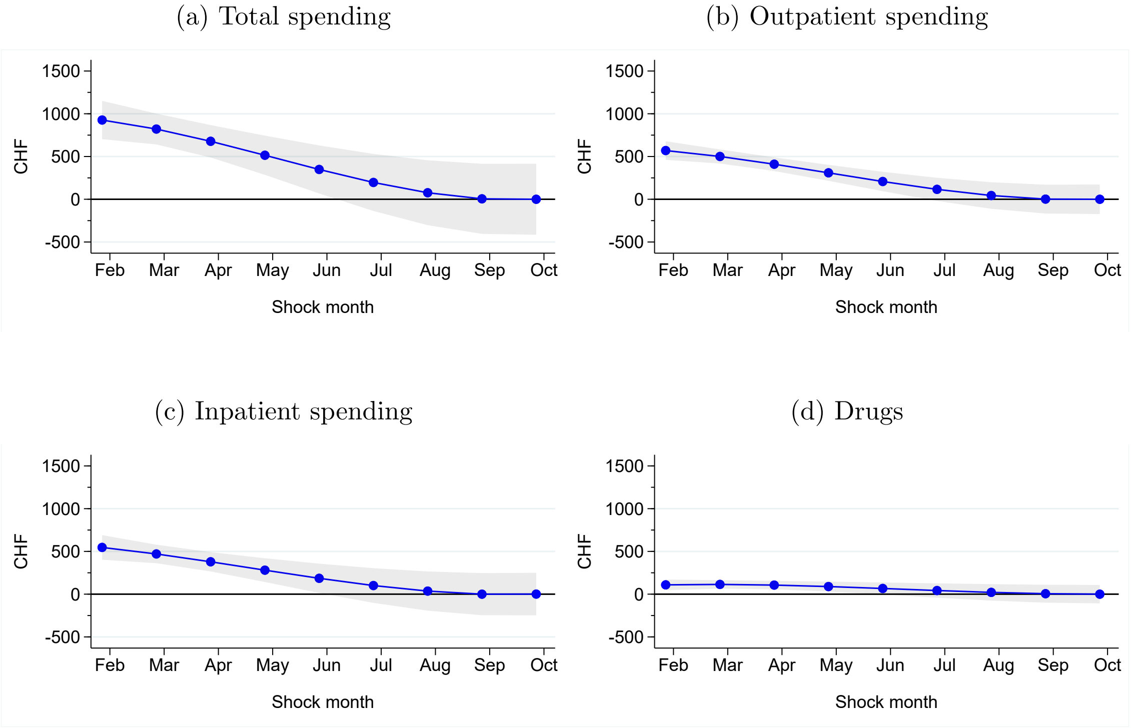
Notes: The figures depict cumulated differences in dynamic treatment effects in calendar time between (a) the shock and the year-end reset, (b) over the whole shock year 1, (c) the post-shock year 2, as well as (d) in both years taken together. Details are presented in Appendix B.4. All differences are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

average estimate is CHF 22.30 (SD=88.20).[26] The share of estimates that are statistically significantly different from 0 at the 10% level is 22%. Recall that significant differences are sufficient to reject the null of no timing moral hazard in this framework.

I then investigate the relationship between the estimated monthly differences and shock timing. This step serves to inform a data-driven choice of the functional form for $m(s)$ in (9). Appendix Figure B.11 presents results from the regression in (11). The estimate for β (i.e. the constant in the regression) is not significantly different from 0. The coefficient on shock timing, i.e. 2δ, is negative and statistically significant at CHF -13.70. These results support the choice of a quadratic functional form for $m(s)$. This specification minimizes the Akaike information criterion, and higher order terms are not significant. Importantly, this relationship indicates a significant variation in the amount of care advanced as a function of shock timing. I use the parameters to predict the total timing moral hazard across shock months as in (12). Given the negative relationship with s, the last shock month provides the lower bound to pin down the level $\underline{\alpha}$.

[26] The differences based on the event study specification that does not adjust for covariates yield virtually identical results.

Figure 5: Timing moral hazard by shock month



Notes: The figure presents estimates of the total yearly timing moral hazard response across shock months, predicted as in (12). 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 5 shows corresponding estimates. It suggests that timing moral hazard is statistically significant and substantial in magnitude for early shock groups. It reaches nearly CHF 1,000 in total for individuals with a shock in February. This means that individuals are forward-looking and act on the incentive to advance care. However, the amount decreases substantially, the later the shock occurs. The estimates for individuals with a shock later than June are not significantly different from zero. This indicates that timing moral hazard is subject to dynamic frictions. The less time individuals have, the less care they advance.

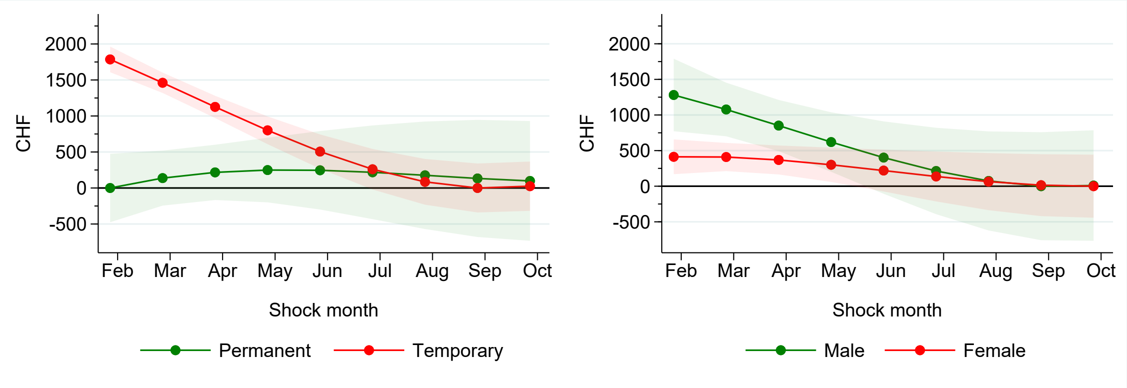
5.3 Heterogeneity

Heterogeneity across types of care. I repeat the estimation steps for three categories of healthcare spending: outpatient, inpatient, and prescription drugs.[27] Figure 5 presents the results. Timing responses decrease with s across all categories. Dynamic frictions are thus relevant in all three cases. Notice that adding the timing responses across types

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

Figure 6: Heterogeneity in timing moral hazard – Total consumption

(a) Permanent vs. temporary shock (b) By gender



Notes: The figure presents estimates of the total yearly timing moral hazard response across shock months, predicted as in (12). 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 permanent shocks. Individuals with a permanent 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) splits the sample by gender.

of care roughly matches the total response, up to an estimation error. Outpatient care amounts to about half of the total timing response. Inpatient care amounts to most of the remaining timing response, with up to 548 for shocks in February. This indicates that outpatient doctor visits are easier to time than hospitalizations, as the latter represent a larger share of spending in the shock year. As for drugs, the response goes up to CHF 110. Drugs can be stocked for later actual consumption, compared to other medical procedures. They have been the focus of, e.g. Einav et al. (2015).

The composition of the timing response constrasts with that of total consumption in the shock year, where inpatient care amounts to over 80% of consumption. These findings support that the estimated spending differences are driven by medical procedures that can be shifted in time. Conceptually, they also support the modellization of three types of healthcare consumption, based on individuals' discretion over the amount and timing. An implication of this heterogeneity is that medical procedures differ in terms of their propensity for strategic timing. In the Swiss context, mandatory health insurance covers emergent and non-emergent procedures under a single deductible. Whether this pooling is desirable from a welfare perspective is an important question for policy.

Permanent and temporary shocks. Temporary shocks create a stronger incentive to advance care, as they are less likely to push the individual above the deductible in the year after (i.e. cross-year prices are different). This intuition is confirmed in the data. Panel (a) of Figure 6 shows that timing moral hazard is much larger for temporary shocks, while all estimates for permanent shocks are close to zero and nonsignificant. Individuals with a permanent 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.[28] Permanent shocks lead to a deterioration in health status that is more likely to justify a low deductible. I explore this in Section 6. Panel (b) shows that women have a smaller timing response than men. This is partly driven by women being more prone to permanent shocks.

5.4 Magnitude and Cost-Sharing Implications

To take stock of their magnitude, I relate the estimates to total and cumulated differences in spending. The pattern and magnitude of the timing response align with the cumulated spending differences in Panel (c) of Figure 4. This comparison suggests that a large part of the differences in spending in the year after the shock is due to timing. Some of the care planned for year 2 may be related to the shock (the model does not exclude that), but can be consumed more or less early.

The February group advances CHF 940 worth of care, which equals a substantial share of 10.2% of their total consumption at CHF 9,200 in the shock year. This translates into a price-elasticity of -0.14.[29] The elasticity decreases to 0 for the last shock group. These estimates are a lower bound, as the latest shock month pinpoints the minimum.

How does timing moral hazard compare to classical moral hazard? My framework does not allow estimating classical moral hazard, so I rely on the benchmark elasticity of -0.2 (Keeler and Rolph, 1988). This is very close to the timing elasticity for the February group. This suggests that there is an important policy trade-off between timing and classical moral hazard in designing health insurance contracts. However, the existing estimates for classical moral hazard may be overestimated, as they have ignored that part of the consumption response is due to timing. A similar point is made in Einav et al. (2015). Also, dynamic frictions should be considered in relating the two behaviors.

How does timing moral hazard affect costs in the risk pool? The cost increase due to shifting amounts to 1% of the premiums in the high-deductible plan. This is based on a back-of-the-envelope calculation, where I assume that the costs of advanced would have been fully borne out-of-pocket by the insured in the absence of timing moral hazard. I shut down any deductible selection mechanism, and assume that this behavior only affects the high-deductible plan. I then apply the month-specific timing elasticity to individuals

[28] Although this sample split is based on endogenous outcomes after the shock, the timing incentives due to the shock are resolved by then, and consumption stabilizes to its long-term level after the shock, as seen in Figure 7.

[29] This estimate relates the additional consumption from timing moral hazard amount to a change in the marginal price from 1 to 0. To approximate baseline consumption, I first take out the share due to classical moral hazard by using the benchmark price-elasticity of -0.2 (Keeler and Rolph, 1988).

who exceed their deductible.

5.5 Characteristics of the Sample

The characteristics of the sample should be noted when interpreting my estimates. The analysis relies on a selected sample of individuals with high deductibles, who become high spenders due to a hospitalization. Understanding the behavior of this population is particularly relevant for policy. Unanticipated hospitalizations have lasting negative consequences on key economic outcomes, e.g. on earnings, access to credit, and consumer borrowing (Dobkin et al., 2018). A particularity of healthcare markets is that a small share of high-spending individuals generate a large share of costs. In my data, the main sample accounted for 14% of collective healthcare spending.

In terms of incentives, the costly shock makes the price change salient relative to smaller price changes. These may blur the incentives and prevent individuals from forming correct expectations about the year-end price (Brot-Goldberg et al., 2017).[30] Individuals with low deductibles exceed them 84% of years and leave little room for cross-year price variation and timing incentives. In terms of preferences, it is possible that the sample a lower price-sensitivity, as individuals who are more prone to classical moral hazard tend to select into higher coverage *ex ante* (Einav et al., 2013). In terms of switching behavior, the sample only includes individuals who are observed for at least three years, i.e. are prone to keeping the same insurer.[31]

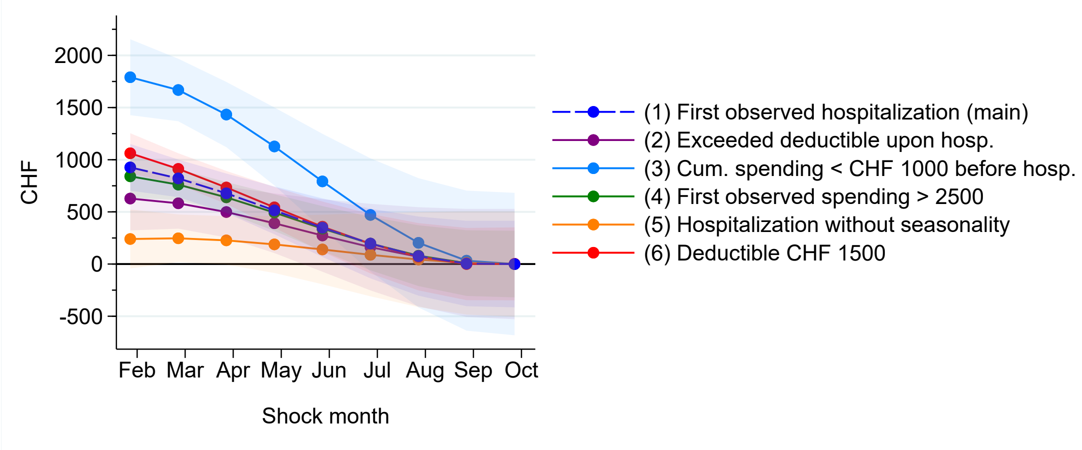
5.6 Alternative Shock and Sample Definitions

Figure 7 presents results for alternative samples and shock definitions. It serves to explore the validity of the identifying assumption of random shock timing, as well as gain further insights into heterogeneity in timing moral hazard. The key qualitative conclusions hold across versions. Timing moral hazard is significant for early shock groups, but decreases for later shock groups due to dynamic frictions. Version (1) is the main one, where I use the main high-deductible sample with the first observed hospitalization as the shock. In version (2), I keep only those individuals who exceeded the deductible in the month of the hospitalization. As individuals mechanically accumulate spending throughout the year, they may exceed the deductible prior to the shock itself. This is more likely for those with hospitalizations later in the year. The estimates are slightly smaller, which may reflect that part of the individuals in the main definition already exceeded the deductible and

[30] Some of the highest price-elasticity estimates for classical moral hazard were found using exogenous variation in exceeding the deductible (e.g. Kowalski, 2016).

[31] 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. Appendix Figure B.3 shows that there is no differential attrition across shock groups.

Figure 7: Timing moral hazard for alternative shock and sample definitions



Notes: The figure presents estimates of the total yearly timing moral hazard response across shock months, predicted as in (12). 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. The sample and shock definitions for different versions are described in Section 5.6.

started advancing care before the hospitalization. In version (3), I focus on individuals whose healthcare consumption in the 12 months before the shock was below CHF 1000. These have a larger response, as shocks occur more suddenly, and a larger proportion of them are temporary. Notice that samples across versions may differ in their underlying severity.

In version (4), I define the shock as the first observed spending over CHF 2,500 in a given month. Since over three quarters of this spending is due to inpatient care, the results closely match those of version (1). In version (5), I restrict the shock definition to hospitalizations types that do not display seasonality patterns,[32] i.e. that occur evenly throughout the year. These are more likely to be elective procedures scheduled regularly (e.g. hip replacements), depending on healthcare providers' capacity constraints. Emergent shocks due to illness exacerbations are more prone to seasonality. Version (6) differs from (1) in that I use individuals with a CHF 1,500 deductible. These have a very similar response, as the maximum amount of advanced care is also below the deductible.

6 Determinants of Timing Moral Hazard

So far, the model has remained agnostic about the determinants of $m(s)$. In the spirit of the sufficient statistics approach, this allows quantifying the magnitude of the response without fully characterizing the underlying primitives. In this section, I first discuss the possible sources for frictions that might drive (heterogeneity in) the utility cost $v_m(\cdot)$. I

[32] Formally, I test whether calendar month dummies significantly jointly predict the probability of observing a specific DRG code.

then impose more structure on the choice of deductible for year 2, as well as the amount of care to advance. This allows learning about how timing moral hazard relates to classical moral hazard and deductible choice. These links are key for understanding the policy trade-offs in designing health insurance contracts with cross-year price incentives.

6.1 Microfoundations for Dynamic Frictions

The decrease in timing moral hazard for late shock groups points to the existence of dynamic frictions to advancing care. Various constraints specific to healthcare markets can restrict individuals from retiming flexibly with little time left. Time constraints can stem from healthcare supply, through e.g. the imperfect control over the timing of appointments or the need for obtaining referrals. They can also come from capacity constraints, as many patients increase their consumption towards the end of the year after exceeding the deductible (Lin and Sacks, 2019; Gerfin et al., 2015). Another feature of healthcare consumption is its lumpiness. Patients can typically not shift just any continuous amount of consumption, as some medical treatments are bundled and come in a sequence of steps that cannot be compressed in time. These bundles might be easier to fit into a longer time horizon.

The utility cost of timing moral hazard may also be determined by, e.g. effort costs to schedule doctor's appointments. The costs of searching for a healthcare provider who can accommodate the necessary treatments might also prevent individuals from advancing care, especially on short notice. Such transaction costs may be higher during the acute phase of the shock. Then, individuals who experience the shock close to the deductible reset might be less inclined to schedule additional appointments within a short time horizon.

The model above describes the behavior of a rational, forward-looking individual. My findings provide support for individuals not being fully myopic. However, the recent literature has highlighted several behavioral biases on the demand side that prevent the insured from achieving their optimal consumption. A longer horizon until the reset leaves more time for individuals to internalize the consequences of the shock, and to respond to the incentive for strategic timing. Meanwhile, myopic individuals might not foresee these incentives (Brot-Goldberg et al., 2017; Abaluck et al., 2018; Dalton et al., 2020).

6.2 Timing Moral Hazard and Choice of Deductible

I now explore further how the incentive to advance care is linked to the choice of deductible in the coverage year after the shock. Assume the monthly healthcare consumption decision in year 2 follows the same optimization problem as in (2). Individuals choose their optimal consumption for all combinations of a low or a high deductible, and how much care they advanced from year 2 to year 1.[33] Given these optimal monthly consumption paths, they choose the combination of deductible and timing moral hazard that maximizes their utility value in year 2. In Appendix A, I show that only two options are rational in a deterministic setting where all health needs are known after the shock. An individual either (i) advances their optimal amount of planned care from year 2 to year 1 and keeps a high deductible; or (ii) keeps consumption as planned and switches to a low deductible in year 2. All other options are dominated. The optimal amount of timing moral hazard is determined by the utility cost of retiming planned care.[34]

The choice between the two options depends on the out-of-pocket cost savings from the timing moral hazard vs. the utility cost of retiming; the difference in premiums $n_L - n_H$; and the foregone utility gain from classical moral hazard above the deductible. Individuals choose option (i) if the reduction in out-of-pocket costs is larger than the utility cost of timing moral hazard plus the opportunity cost of classical moral hazard in year 2. Timing moral hazard might put the individual below the deductible and prevent them from drawing utility from classical moral hazard. Interestingly, a higher ω attenuates the incentive to advance care.

These results highlight two key policy trade-offs under deductibles, when incentives are to advance care. First, it reduces classical moral hazard in the next year. Second, advancing care reduces adverse selection the year after. There is an important asymmetry with the case where the incentive is to delay procedures in order to purchase additional coverage as in, e.g. Cabral (2017). An incentive to delay generates a complementarity between timing moral hazard and coverage purchase, instead of substitutability. It also increases classical moral hazard as it pushes individuals into free care.

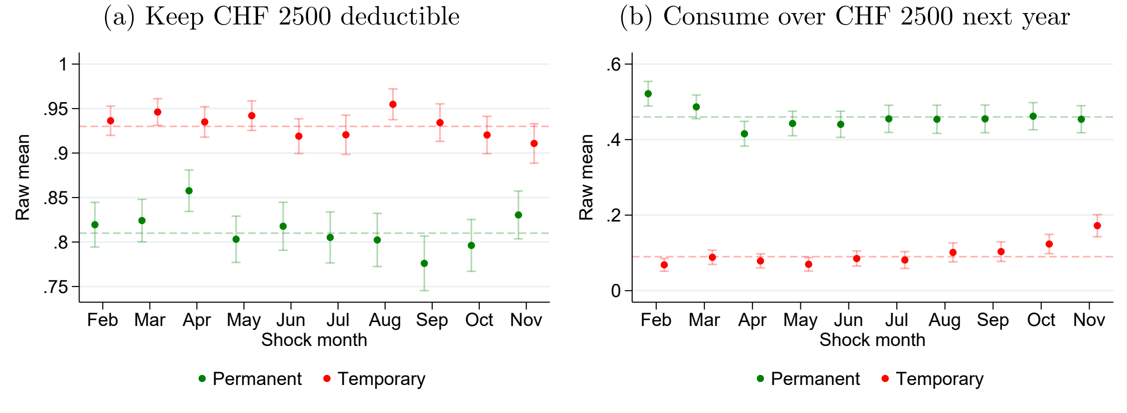
Although the intuition above holds regardless of s, shock timing affects the relative value of the two options via the time left to advance, and how strongly the shock persists into year 2. As a consequence, even individuals with comparable risk at the start of the coverage year may diverge in their timing moral hazard responses, depending solely on when their shock realizes. In other words, they can become unobservably differentiated by the amount of care they have advanced as a result of shock timing, as in Cabral (2017).

Extensive vs. intensive margin. Is the heterogeneity in timing moral hazard driven by the share of individuals who advance care, or the amount itself? The timing moral hazard estimates I compute are a weighted average of the response $\tilde{m}(s)$ of individuals

[33] For there to be an incentive to advance care, the shock should not be so persistent that the deductible is exceeded for everyone in year 2.

[34] Appendix A discusses extensions of the theoretical model, as well as their implications for the reduced-form estimates.

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 permanent shocks. The dotted horizontal lines denote the corresponding subsample average. Confidence intervals are at the 95% level, based on robust standard errors.

who advance care, denoted by $q(s)$, and those who do not:

$$m(s) = \tilde{m}(s)q(s) - 0 \cdot (1 - q(s)) \tag{13}$$

Based on the result above that strategic timers keep a high deductible, I check for heterogeneity in deductible switching rates after the shock. Figure 8 shows the raw shares of individuals who keep the CHF 2,500 deductible in the year after the shock, split by permanent and temporary shocks. The shares are roughly constant in both groups, at 92% for individuals with a temporary shock and 82% for those with a permanent shock. Hence, only a small share switches to a low deductible, although half of those with a permanent shock exceed it the year after, and about 10% of those with a temporary shock do (especially those with a late shock).[35] This may be driven by high switching costs or behavioral biases.[36] The main insight here is that heterogeneity in $m(s)$ is driven by the amount rather than the share of strategic timers.

[35] 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.

[36] Although this is beyond the scope of this paper, a large literature has shown that individuals do not choose their utility-maximizing health insurance plan, due to e.g. switching costs, inattention or inertia (see e.g. Abaluck and Gruber, 2011; Handel, 2013; Abaluck and Gruber, 2016; Handel and Kolstad, 2015; Heiss et al., 2016, and Winter and Wuppermann, 2019 for a review of the recent literature).

7 Conclusion

In this paper, I introduce a new approach to identifying timing moral hazard under deductibles in health insurance. My setup allows for classical moral hazard and deductible choice. I consider high-deductible individuals who suffer an unexpected hospitalization, exceed their deductible, and have an incentive to advance care from the following year. I provide a sufficient statistic approach that can be used to estimate timing moral hazard in reduced form by exploiting the random timing of the shock within the coverage year. Empirically, I find that the amount of care advanced after the shock is substantial, and reaches nearly CHF 1,000 when the shock occurs in February, i.e. 10% of spending in the year of the shock. This is mainly driven by individuals suffering temporary shocks. However, individuals who have a shock late in the calendar year face dynamic frictions and have a significantly lower timing moral hazard response.

This paper contributes several important insights into strategic timing behavior in health insurance with nonlinear cost-sharing, with implications for contract design. Individuals are not fully myopic. They adapt in the face of a bad risk realization by advancing significant amounts of planned healthcare consumption, so as to reduce out-of-pocket costs in the following coverage period. Not only the realization of the risk (or its *ex ante* probability), but also its timing matters for *ex post* consumption when individuals can choose when to address the risk, and have sufficient time to do so.

Distinguishing between the intensive and timing margins is key in analyzing price-reponses in consumption. Timing moral hazard affects our interpretation of existing estimates for the price-elasticity of healthcare consumption. My results underline that many of these include care shifted in time and onto the risk pool, and not only additional consumption driven by classical moral hazard. In other words, they capture consumption that would have happened anyway, but would have otherwise been paid for out of pocket.

Timing responses generate externalities via higher costs in the risk pool, with implications for premiums. The insurer can observe the realized (timed) consumption, and price plans accordingly. This can happen even without the possibility to choose coverage, e.g. the deductible level. When deductible choice is possible, I show that timing moral hazard is not only related to selection, but also to classical moral hazard. Shifting care away from a period and saving out-of-pocket costs may lead to foregoing utility from classical moral hazard, and conversely. This highlights important policy trade-offs for health insurance. Deductibles limit classical moral hazard but generate timing incentives, even without deductible choice. Additionally, deductible choice may create value if individual valuations for insurance are heterogeneous (Hendren et al., 2021). This value may be reduced not only by adverse selection on *ex ante* risk, but also by *ex post* selection based on the timing of risk realizations and planned consumption.

Based on my findings, coverage length and multiple deductibles for different types of care are relevant contract features for targeting timing moral hazard. Shorter coverage length could limit the advancing of care. However, there is likely to be an asymmetry between advancing and deferring planned care. Shorter coverage may facilitate the deferral of care. In turn, this may exacerbate *ex post* adverse selection and classical moral hazard in the next coverage period. This highlights that the welfare implications of timing moral hazard are generally ambiguous. They depend, among other things, on the time direction of the shift (advancing or deferring), the strength of the interaction with other sources of asymmetric information, the value of the timed care, and the potential effects of timing on health outcomes. These aspects are interesting avenues for future research.

References

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

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

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

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

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

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

Brot-Goldberg, Zarek C, Amitabh Chandra, Benjamin R Handel, and Jonathan T Kolstad (2017). “What does a deductible do? The impact of cost-sharing on health care prices, quantities, and spending dynamics”, *The Quarterly Journal of Economics*, 132(3): 1261–1318.

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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.

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

A Theoretical Model

A.1 Optimal Choice of Deductible and Timing Moral Hazard

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

$$V(D_j, \bar{m}(s)) = \begin{cases} -\bar{\lambda}(s) - \bar{v}_m(m,s) - \bar{m}(s) - \bar{n}_j & \text{if } \bar{\lambda}(s) + \bar{m}(s) < D_j \\ \frac{\bar{\omega}}{2} - D_j - \bar{n}_j & \text{if } \bar{\lambda}(s) + \bar{m}(s) \geq D_j \end{cases} \tag{14}$$

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

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

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

$$V(D_j, \bar{m}(s)) = -\bar{\lambda}(s) - \bar{v}_m(m,s) - \bar{m}(s) - \bar{n}_j \tag{15}$$

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

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

$$V(D_H, \bar{m}(s)) \geq V(D_L, \bar{\mu}(s)) \tag{16}$$

$$(\bar{n}_L - \bar{n}_H) + (D_L - \bar{\lambda}(s) - \rho(s)) - \bar{v}_m(m,s) - \frac{\bar{\omega}}{2} \geq 0 \quad \forall s \in S \tag{17}$$

They choose D_H and advance if the sum of the following terms is positive: the savings in premiums; the out-of-pocket cost difference from advancing care; the costs of planned care still consumed in year 2; the utility cost of retiming; and the opportunity cost in terms of foregone utility from classical moral hazard.

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

Hence, in all cases, the individual either chooses to advance the optimal amount and keep a high deductible, or not to advance anything and switch to a low deductible. The individual will choose either depending on the indifference condition (16).

By the functional form assumption in (9), shifted amounts are evenly allocated throughout the target year as $m(s)$ does not depend on t. Note that below the deductible j, conditional on advancing, the individual advances the amount that the marginal cost of retiming equals the out-of-pocket cost savings. Above the deductible, planned consumption is not shifted as there are no savings to be achieved. This yields two optimal planned care consumption paths under the two options. Strategic timers advance the optimal amount of planned care to year 1 and keep the high deductible. Switchers consume all care as planned, and do not advance.

A.2 Model Extensions

This appendix discusses the intuition behind possible extensions of the model, as well as the implications for the interpretation of empirical estimates of timing moral hazard.

Uncertainty. In this setup, the only source of uncertainty is in the timing of the shock, and is resolved once the shock is realized. This simple setup allows understanding the mechanisms and consequences of timing. Uncertainty with respect to shock-induced needs after the shock does not affect the general formula for optimal healthcare consumption in (3), and the key identification result holds. This is because all individuals face a zero price after the shock with certainty. They are also comparable in relative time, as their expected health needs in relative time after the shock are the same on average, $E[\lambda_k(s)] = E[\lambda_k(s')]$ for $s, s' \in S$. However, uncertainty affects the decisions in year 2 and the amount of timing moral hazard. Individuals now compare the expected utility value of combinations of timing moral hazard and deductible given their risk aversion.

Consider that individuals pay for their (known) planned healthcare consumption below the deductible in year 2. Advancing care increases the amount to cover out-of-pocket below the deductible, which is now uncertain and depends on realized shock-induced needs. Timing moral hazard then affects the variance of out-of-pocket spending in the year following the shock. More risk averse individuals tend to advance less care to reduce uncertainty. Importantly, this means that uncertainty works against me finding evidence for timing moral hazard in the empirical implementation, since it drives down differences in spending.

Discounting. The model can be extended to include exponential or quasi-hyperbolic discounting. This would affect the year-end price in the second year, but the structure of the decision in year 1 would remain unchanged. Individuals are comparable on average in their discounting factor if shock timing is random. Empirically, discounting works against me finding timing responses, as individuals do not foresee planned care or cross-year price incentives to advance care.

Coverage choice. First, the choice of coverage can be extended to multiple deductible levels (as in the Swiss or the Dutch setting), or to include the possibility to opt in and out of insurance (as in, e.g., employer-sponsored health insurance in the United States, Diamond et al. 2021). The latter would allow for choosing a zero deductible and premiums. Theoretically, an opt-out option would strengthen the incentive to advance care in my setup, as some individuals would have the incentive to advance care and not purchase any coverage the next year. This would reinforce the conclusions regarding the link between timing moral hazard and selection.

Time horizon. Given that I focus on the decision to advance care after a health shock, the model in two years reasonably captures the key short-term timing incentives. Two years is also a plausible horizon for advancing care. Extending the horizon further away from the shock would allow for the possibility that individuals advance care that is planned for over two years after the shock (which seems unlikely). It is still possible that the empirical estimates capture such a response. In general, the setup can be extended to a longer time horizon with repeated shocks.

B Empirical Analysis

B.1 Data and Sample

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

	(1)	(2)
	Baseline sample	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).

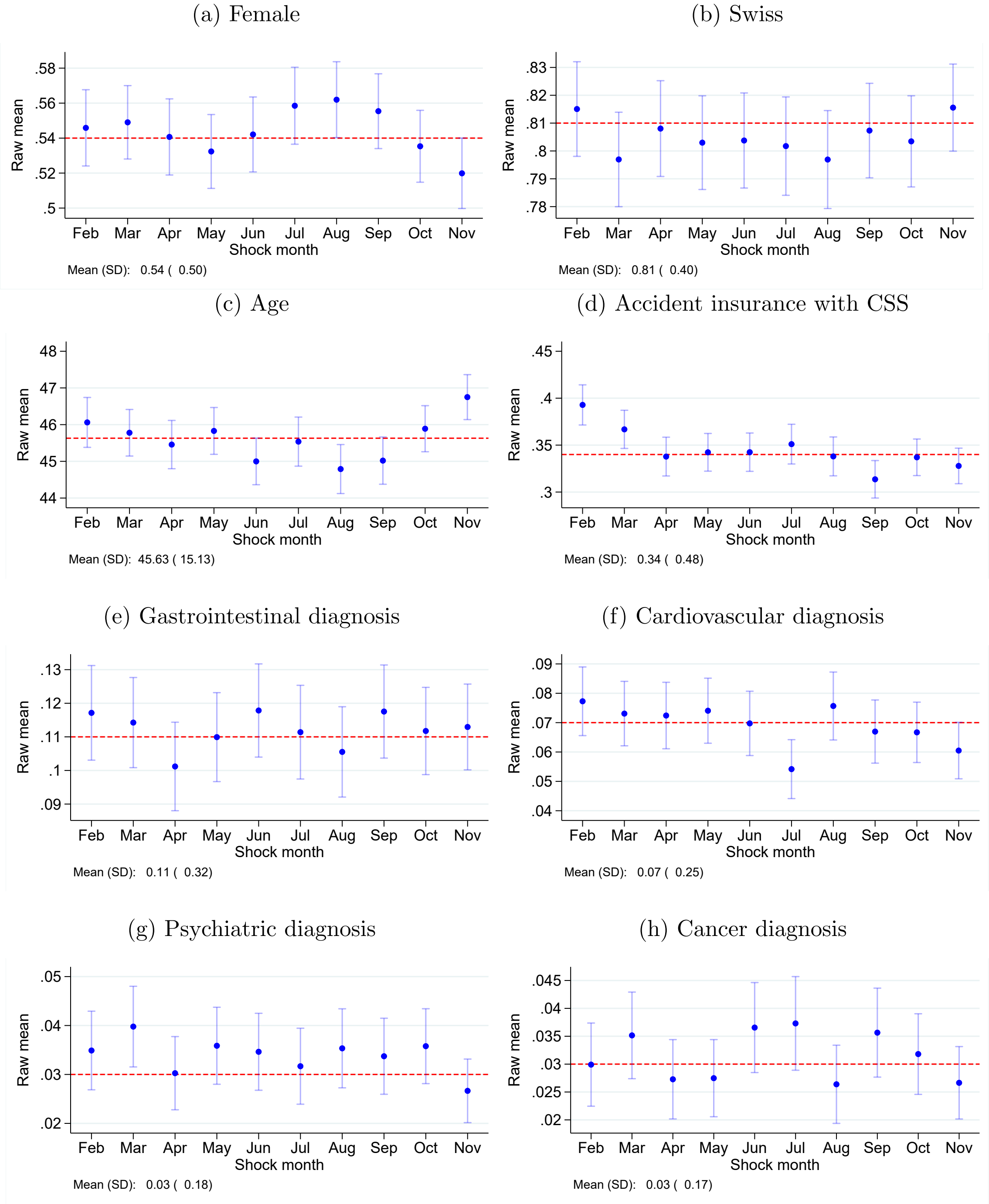
[a] Figures condition on being aged between 19 and 90. Source for population statistics: Population statistics from the Federal Statistical Office.

[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

Figure B.1: Balancedness in observables across shock months


(a) Female

(b) Swiss

(c) Age

(d) Accident insurance with CSS

(e) Gastrointestinal diagnosis

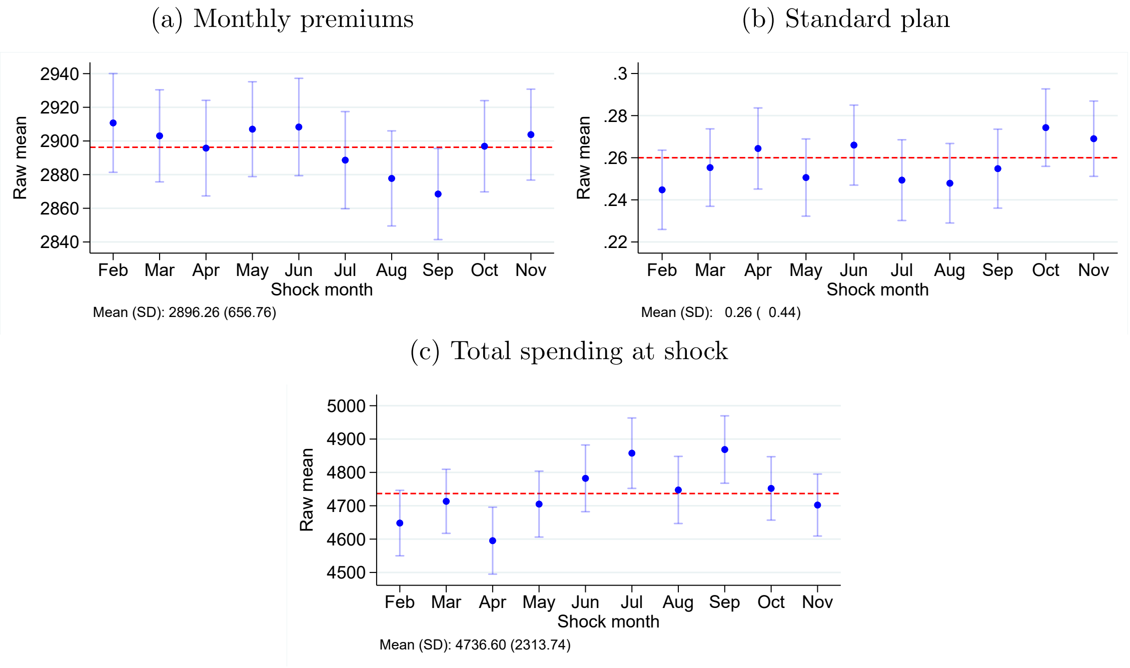
(f) Cardiovascular diagnosis

(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. 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

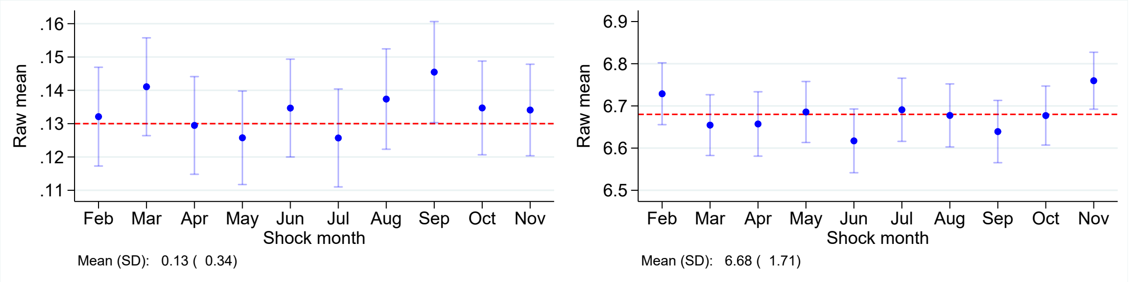
(c) Total spending at shock

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

Figure B.3: Attrition across shock months

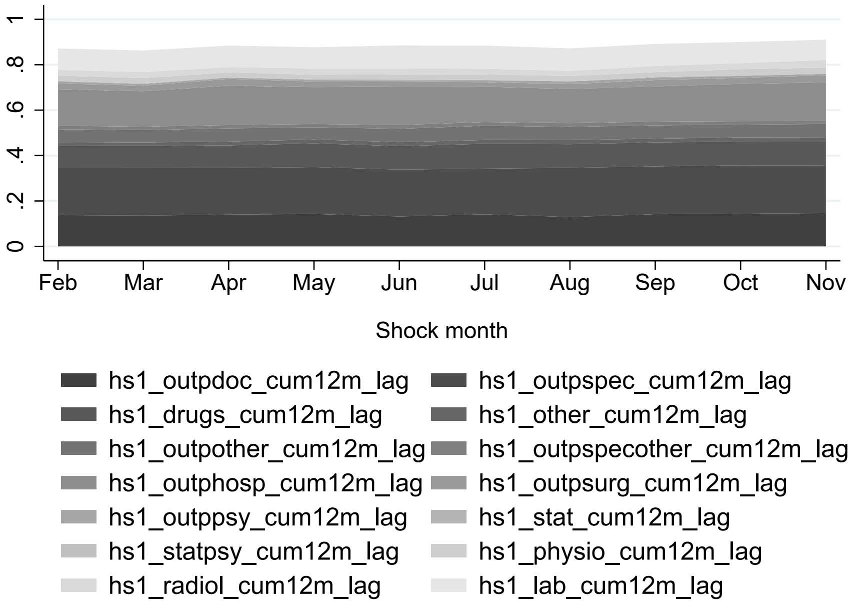
(a) Attrits the year after the shock

(b) Years observed



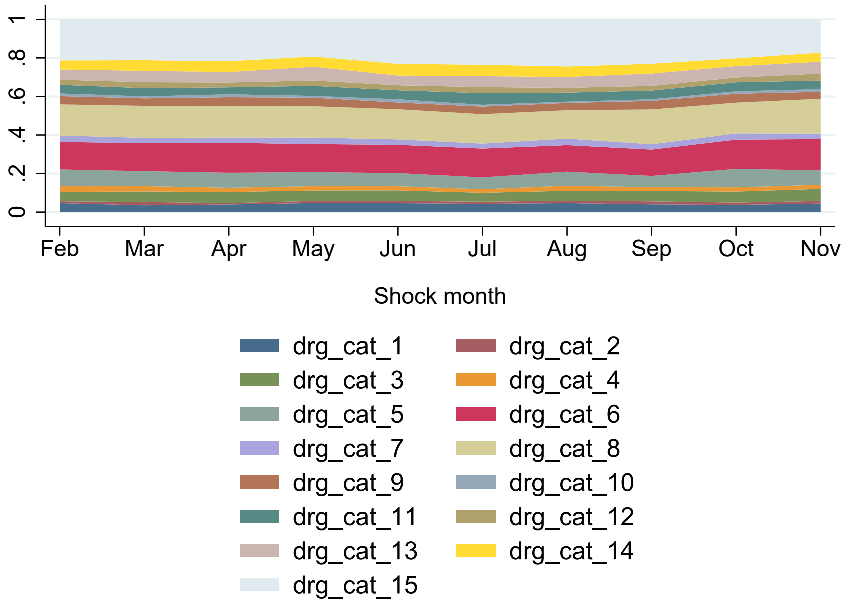
Notes: This figure presents raw averages across shock month groups. Observations are individuals from the main analysis sample. Confidence intervals are at the 95% level based on robust standard errors.

Figure B.4: Composition of cumulated spending 12 months before the shock



Notes: This figure presents the raw composition of cumulated healthcare consumption in the 12 months before the health shock (defined as the first hospitalization) across shock month groups.

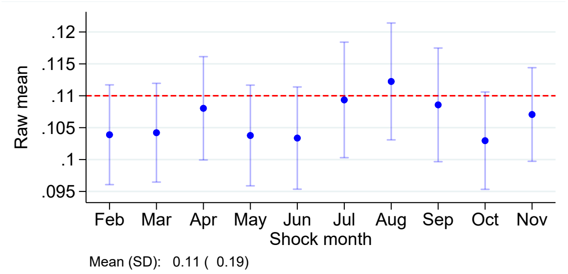
Figure B.5: Composition of DRG hospitalization categories at shock



Notes: This figure presents the raw composition of hospitalizations by category of diagnosis-related group (first letter) across shock month groups.

B.3 Prices, Deductible Choice, and Future Shocks

Figure B.6: Year-end marginal price in the shock year

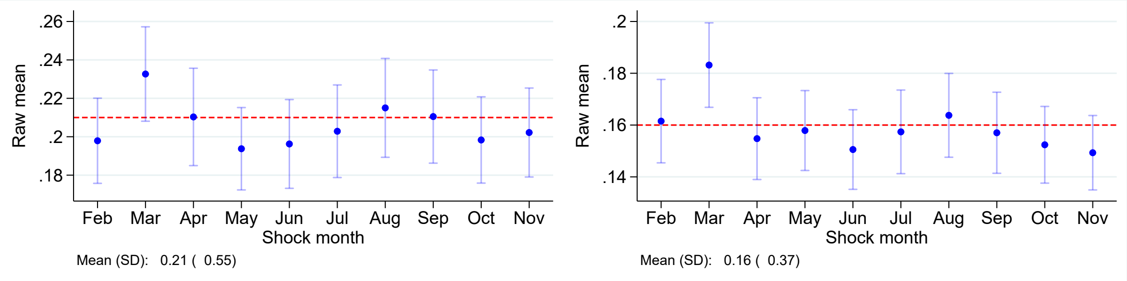


Notes: This figure presents raw averages across shock month groups. Observations are individuals from the main analysis sample. Confidence intervals are at the 95% level based on robust standard errors.

Figure B.7: 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. 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

$$\Delta\text{ShockReset}(s) = \sum_{m=1}^{12-s} \hat{\gamma}_m^s - \sum_{m=1}^{10} \hat{\gamma}_m^2 \tag{18}$$

Cumulated difference in year of the shock (year 1)

$$\Delta\text{Year1}(s) = \sum_{m=2-s}^{12-s} \hat{\gamma}_m^s - \sum_{m=-1}^{10} \hat{\gamma}_m^2 \tag{19}$$

Cumulated difference in year after the shock (year 2)

$$\Delta\text{Year2}(s) = \sum_{m=13-s}^{24-s} \hat{\gamma}_m^s - \sum_{m=11}^{22} \hat{\gamma}_m^2 \tag{20}$$

Cumulated difference in both years

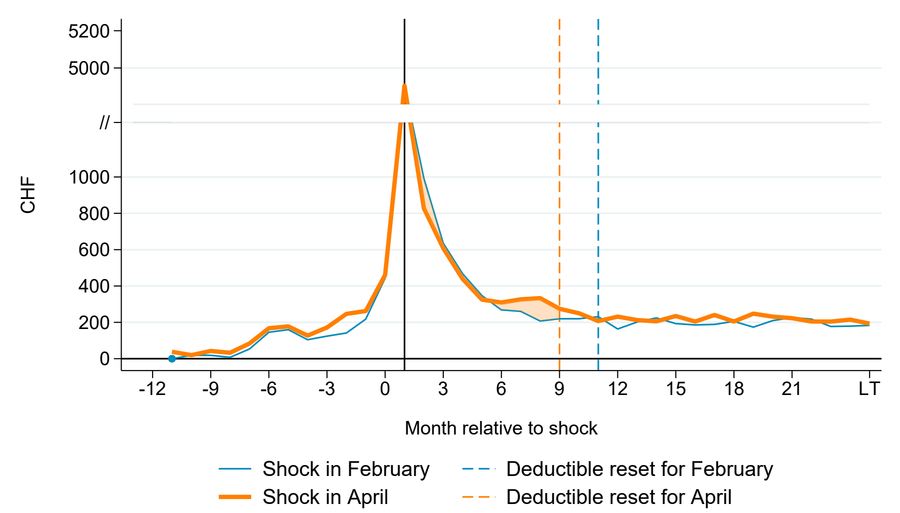
$$\Delta\text{BothYears}(s) = \sum_{m=2-s}^{24-s} \hat{\gamma}_m^s - \sum_{m=-1}^{22} \hat{\gamma}_m^2 \tag{21}$$

Cumulated differences computed for all $s = 3, \ldots, 10$, relative to the February group $s = 2$.

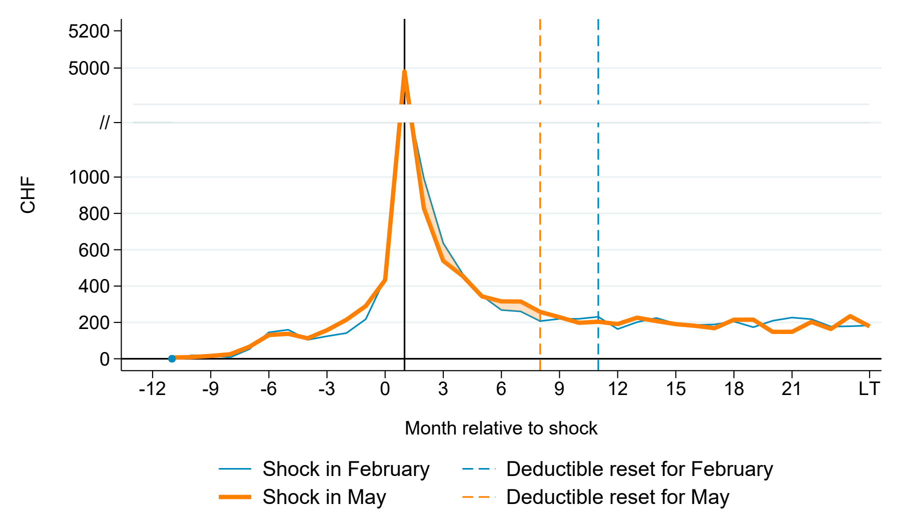
B.5 Additional Results

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

(a) Shock in February vs. April



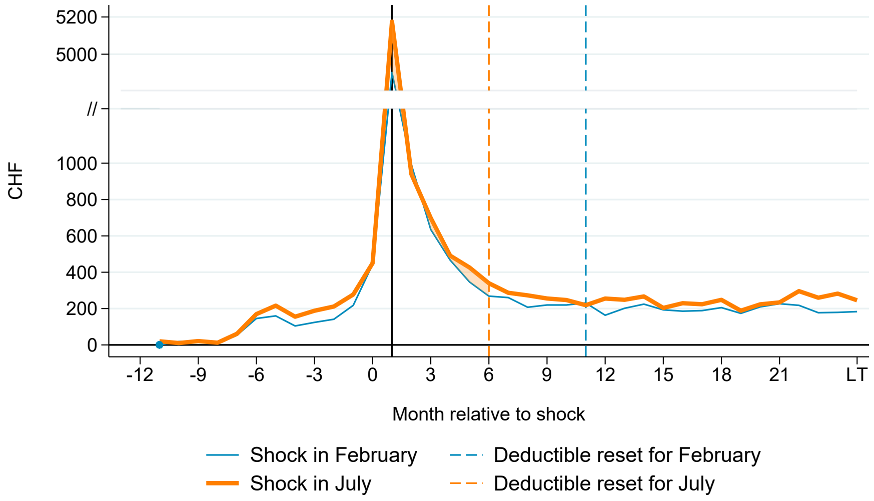
(b) Shock in February vs. May



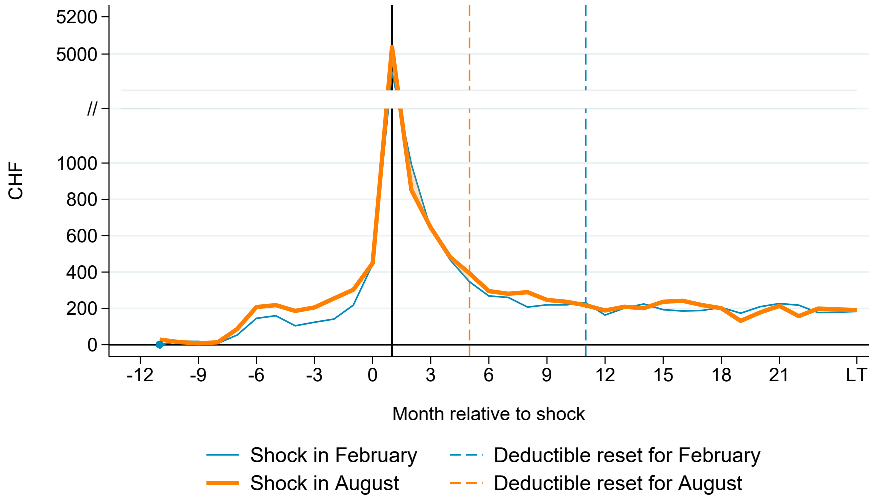
Notes: The figure depicts the coefficient estimates on monthly treatment effects from the event study of healthcare consumption (as measured by total spending). It compares individuals with shocks in different months, for the main analysis sample of insured with a high deductible and the health shock 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.9: Event study of healthcare consumption around the health shock – Additional treatment group comparisons

(a) Shock in February vs. July



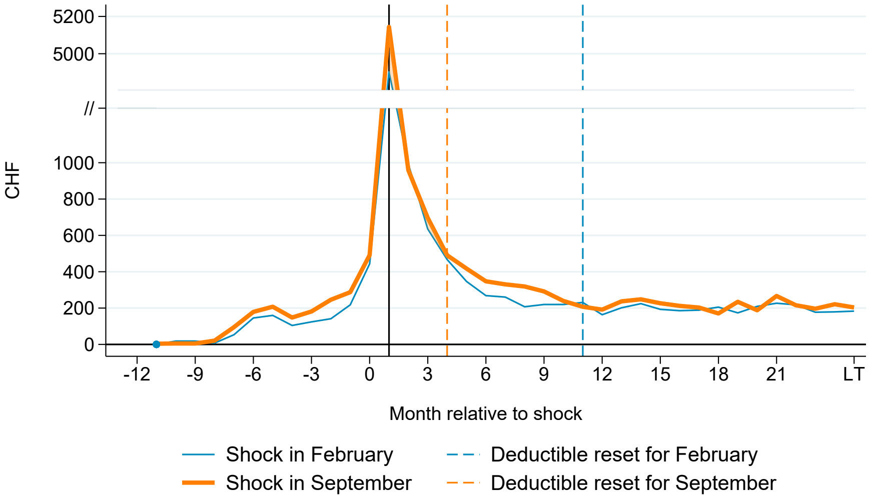
(b) Shock in February vs. August



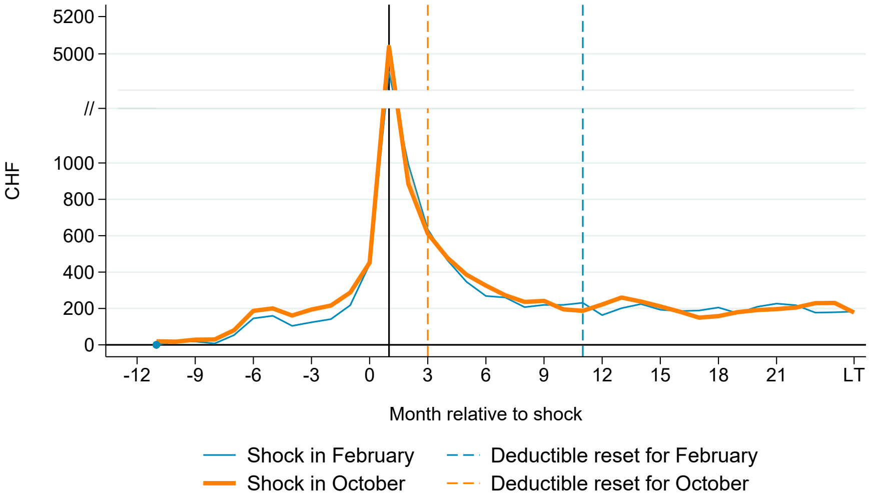
Notes: The figure depicts the coefficient estimates on monthly treatment effects from the event study of healthcare consumption (as measured by total spending). It compares individuals with shocks in different months, for the main analysis sample of insured with a high deductible and the health shock 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.10: Event study of healthcare consumption around the health shock – Additional treatment group comparisons

(a) Shock in February vs. September

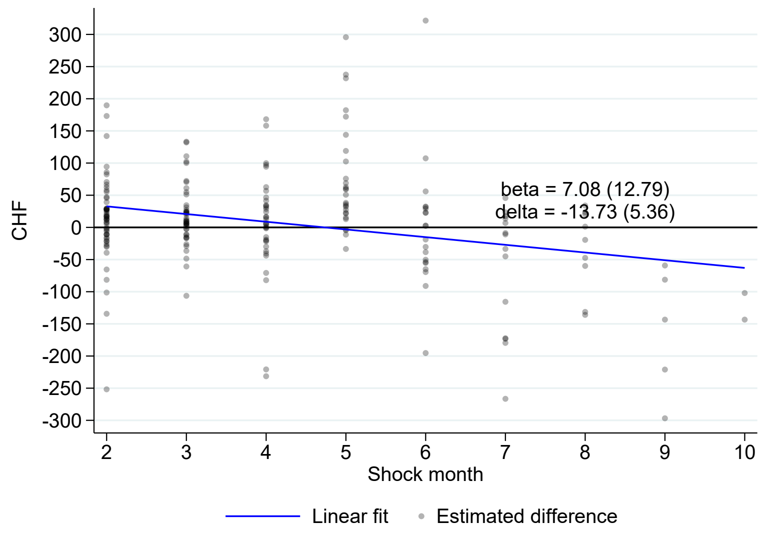


(b) Shock in February vs. October



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.11: 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 (11).

Table B.2: Summary statistics for alternative shock definitions

	(1)	(2)	(3)	(4)	(5)	(6)
	Main shock	Exceeded ded.	Cum. spending	First observed sp.	Hosp. without	CHF 1500
	= Hospitalization	upon hosp.	below 1000	over 2500	seasonality	deductible
Demographics						
Age	45.63 (15.13)	46.50 (15.17)	47.45 (15.02)	45.72 (15.42)	46.11 (15.12)	49.04 (16.11)
Female	0.54	0.51	0.43	0.58	0.53	0.52
Swiss	0.81	0.82	0.83	0.80	0.82	0.82
Insurance plan						
Monthly premiums	2,896 (657)	2,868 (647)	2,858 (641)	2,902 (661)	2,862 (647)	3,420 (776)
Family physician plan	0.26	0.26	0.26	0.26	0.26	0.32
Accident insurance	0.34	0.36	0.36	0.35	0.35	0.41
Spending and prices						
Total out-of-pocket spending	5,809 (765)	5,788 (705)	5,740 (707)	5,916 (716)	5,782 (706)	5,430 (834)
annual spending	9,073 (7,758)	8,432 (6,852)	8,213 (7,369)	10,422 (8,197)	8,273 (6,331)	9,677 (8,447)
Exceeded deductible	0.96	1.00	1.00	1.00	1.00	0.99
Cost-sharing	0.45	0.46	0.49	0.38	0.46	0.30
Insured	21080	15021	9484	17146	13786	19693

Notes: The table presents means and standard deviations (in parentheses) for samples of insured-years. Column (1) includes the main analysis sample—insured-year observations with annual deductibles of CHF 2,500 and the main shock definition of first observed hospitalization. Column (2) takes a subsample of individuals who exceeded their deductible at the hospitalization. Column (3) takes a subsample of individuals who consumed less than CHF 1,000 worth of care in the 12 months preceding the hospitalization. Column (4) defines the shock as a spending greater than CHF 2,500 observed for the first time. Column (5) takes a subset of hospitalizations without seasonality. Column (6) takes the CHF 1,500 deductible as sample restriction. Cost-sharing is calculated as out-of-pocket spending (net of premiums) over total yearly healthcare spending. Total out-of-pocket spending includes insurance premiums.