Klein, Tobias J.; Salm, Martin; Upadhyay, Suraj

Working Paper

Patient Cost-Sharing and Redistribution in Health Insurance

IZA Discussion Papers, No. 16778

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Klein, Tobias J.; Salm, Martin; Upadhyay, Suraj (2024) : Patient Cost-Sharing and Redistribution in Health Insurance, IZA Discussion Papers, No. 16778, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/295801

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16778

Patient Cost-Sharing and Redistribution in Health Insurance

Tobias J. Klein
Martin Salm
Suraj Upadhyay

JANUARY 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16778

Patient Cost-Sharing and Redistribution in Health Insurance

Tobias J. Klein
Tilburg University and IZA

Martin Salm
Tilburg University and IZA

Suraj Upadhyay
Tilburg University

JANUARY 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Patient Cost-Sharing and Redistribution in Health Insurance*

Health insurance premiums often do not reflect individual health risks, implying redistribution from individuals with low health risks to individuals with high health risks. This paper studies whether more cost-sharing leads to less redistribution and to lower welfare of high-risk individuals. This could be the case because more cost-sharing increases out-of-pocket payments especially for high-risk individuals. We estimate a structural model of healthcare consumption using administrative data from a Dutch health insurer. We use the model to simulate the effects of a host of counterfactual policies. The policy that was in place was a 350 euro deductible. Our counterfactual experiments show that redistribution would decrease when the deductible would increase. Nonetheless, high-risk individuals can benefit from higher levels of cost-sharing. The reason is that this leads to lower premiums because both high-risk and low-risk individuals strongly react to the financial incentives cost-sharing provides.

JEL Classification: I13

Keywords: health insurance, moral hazard, patient cost-sharing, redistribution

Corresponding author:
Tobias J. Klein
Tilburg University
Department of Econometrics and OR
PO Box 90153
5000 LE Tilburg
The Netherlands

E-mail: T.J.Klein@uvt.nl

* We are grateful for numerous insightful discussions we have had with Johan Polder and Arthur Hayen, and for comments received from Jaap Abbring, Angie Acquatella, Jan Boone, Amitabh Chandra, Rudy Douven, Keith Ericson, Sebastian Fleitas, Benjamin Handel, Annika Herr, Richard van Kleef, Misja Mikkers, Wanda Mimra, Hendrik Schmitz, Lan Zou, and participants of an internal seminar at Tilburg University, the 2022 German Health Economics Association conference in Tilburg, the 2022 German Economic Association Health Committee (Mannheim), the 2022 CEAR/MRIC behavioral insurance workshop in Munich, the 2022 Essen Health Conference, the 1st CEPR Conference in Health Economics in Toulouse, and participants of seminars at the University of Augsburg and the University of Hamburg. Funding from the RIVM is gratefully acknowledged.

1 Introduction

In addition to providing insurance against unexpected healthcare expenditures, health insurance is frequently also used to redistribute resources from individuals with low health risks to individuals with high health risks. Such redistribution is enacted through various institutional arrangements in healthcare systems throughout the world. For example, prices and terms of employer-provided insuranceusually do not depend on an individual employee's health, the Affordable Care Act does not allow insurance providers to reject individuals or base premiums on pre-existing conditions, and many countries provide free public health insurance for vulnerable groups in the population, such as Medicare and Medicaid in the United States.

At the same time, it is well understood that full insurance can lead to moral hazard and inefficiently high healthcare costs (Pauly, 1968). One way to address this concern is to use some form of patient cost-sharing, such as deductibles, co-insurance, or co-payments. While this is an effective way to reduce healthcare expenditures (Manning et al., 1987; Zweifel and Manning, 2000; McGuire, 2011; Aron-Dine et al., 2013; Brot-Goldberg et al., 2017), patient cost-sharing is a contentious topic in the public debate. Related concerns are that cost-sharing leads to a higher financial burden for sicker individuals and to less redistribution towards individuals with high health risks.

In general, cost-sharing has two effects. The first effect is that it increases out-of-pocket payments for a given level of care consumption. This results in out-of-pocket risk born by individuals. It also leads to less redistribution, because high-risk individuals have on average higher healthcare expenditures, so that their out-of-pocket expenditures increase more when there is more cost-sharing. The second effect is that it lowers care consumption of all groups. That can lead to more or less redistribution depending on which group reacts stronger to it. It generally reduces total healthcare costs and thereby leads to lower premiums. The welfare effects are generally ambiguous.[1]

Our research question is whether higher levels of cost-sharing lead to more or less redistri-

[1]For example, it could be that high-risk individuals are better off with more cost-sharing even if it leads to less redistribution. This is the case if they value the premium reduction net of the additional out-of-pocket payments and net of the risk premium associated with increased out-of-pocket risk more than they value the care they do not consume anymore.

bution from low to high risks and to higher or lower welfare of the most vulnerable persons with high health risks. To answer this question, we need to quantify the effects of counterfactual cost-sharing policies on spending and welfare. In the absence of an experiment we exploit variation in observational data, and we develop a structural model to examine the effects of patient cost-sharing on spending, redistribution between risk groups, and the welfare of persons with high and low health risks. We use data from a large health insurance company in the Netherlands to estimate the parameters of the model. Then, we evaluate the effects of a number of counterfactual cost-sharing policies. The Netherlands provides an almost ideal setting for this exercise. It combines community-rated premiums, an individual mandate to purchase health insurance, and a mandatory annual deductible that applies to a broad range of healthcare services. The level of the deductible is set by the national government. In the year 2013, the focus of our study, the mandatory deductible was €350.[2] Our sample is representative of the national population. This makes it very suitable to examine the distributional effects of alternative cost-sharing schemes.

Developing and estimating a structural model for our purpose requires overcoming several challenges. First, health insurance contracts generally feature nonlinear prices; for example, the price of healthcare with a deductible contract depends on whether spending up to a given point in time has exceeded the deductible limit or not. A model thus needs to allow for patients to respond to dynamic incentives generated by these nonlinear prices, as shown in Keeler et al. (1977). Second, when patients decide whether to visit a doctor, they are uncertain about the size of the healthcare need they have. Third, healthcare needs are generally unobserved. We only observe healthcare spending. Observed healthcare spending can both be higher than healthcare needs (if insured patients use more care than they would in the absence of insurance) or lower than healthcare needs (if patients do not seek care despite having a healthcare need). Fourth, the model should be general enough so that we can simulate the effects of a variety of counterfactual cost-sharing policies.

We overcome these challenges by modeling an individual's healthcare consumption as the solution to a finite-horizon dynamic programming problem within a single year. Our model cap-

[2]Very few individuals choose an additional voluntary deductible. We exclude them from the analysis. Handel et al. (2024) show that risk aversion does not explain why only very few individuals choose an additional voluntary deductible.

tures the dynamics introduced by non-linear prices. Consuming more healthcare today makes it more likely to exceed the cost-sharing limit before the end of the year and thus to receive "free" care in later periods. Patients in our model take this into account. They make choices at the extensive and the intensive margin. Patients have incomplete information. In each month, individuals face a probability of having a healthcare need, which is modeled as a first-order Markov process. They have incomplete information about the healthcare needs. They know whether they have a healthcare need. When they have a healthcare need, they form expectations over its size and decide whether or not to visit a doctor. If they decide to visit a doctor, the size of their healthcare need is revealed and they then choose how much healthcare to consume based on their healthcare need and cost-sharing incentives. We allow all our model parameters to vary across risk score quartiles. To do so, we calculate risk scores using background characteristics and data on healthcare use from previous years. These risk scores are the higher the higher expected spending is for each individual in our data.

We discuss identification and then estimate the parameters of our model by matching model predictions to moments in the data. Our model fits the data well. We perform counterfactual simulations for a host of counterfactual cost-sharing policies that are either part of the academic discussion or have actually been proposed in the Dutch policy debate. These include several levels of the annual deductible up to €500, donut hole contracts, a bi-annual deductible, co-insurance, co-pays, and a recently proposed option where deductible payments are capped.

Our main finding is that policies that lower spending relative to no cost-sharing do not necessarily led to less redistribution from low-risk to high-risk individuals; and even if they do, policies that lead to lower spending generally lead to higher welfare for all risk score groups. The reason for this is that patients of all groups, including those with the highest predicted spending, value premium reductions more than the additional care they consume under less cost-sharing. More specifically, for one-year deductible contracts, redistribution is biggest for the €350 deductible that was in place. It is similar for lower values of the deductible or no cost-sharing and lower when the deductible is increased to €500. In the latter case, redistribution decreases since care consumption by individuals in the fourth risk score quartile decreases the most when the deductible changes from €350 to €500. Yet, increasing the deductible from the

€350 that were in place to €500 also leads to a large reduction in the insurance premium. This reduction in the premium is valued by all individuals. The transfer patients in the fourth risk score group receive is smaller than before, but nonetheless, their welfare increases because they value the premium reduction more than they valued the additional transfer. Similar patterns arise for the other policies we consider.

In our model, patients are risk-neutral. We made this modeling choice because the out-of-pocket risk is generally small for the policies we consider. To assess the robustness of our results, we follow Handel et al. (2024) and quantify the effect of out-of-pocket risk by calculating a risk premium for each of the policy options. We find that risk premiums are generally small and do not affect the ranking of policies in terms of welfare. We also carefully discuss why additional factors that are not captured by our model, such as liquidity constraints or behavioral hazard, are unlikely to overturn our main finding.

Our research relates and contributes to a large literature on the effects of patient cost-sharing on healthcare spending (for surveys see Cutler and Zeckhauser, 2000 and Einav and Finkelstein, 2018). The general finding in this literature is that patients strongly respond to cost-sharing incentives. This finding goes back to the RAND Health Insurance Experiment (Manning et al., 1987; Aron-Dine et al., 2013) and has more recently been confirmed (e.g. Chandra et al., 2010; Shigeoka, 2014; Brot-Goldberg et al., 2017). One strand of this literature examines the question to what extent patients respond to the incentives imposed by nonlinear prices in health insurance contracts. The empirical evidence is mixed. Some studies find evidence for substantial myopia (Abaluck et al., 2018; Guo and Zhang, 2019; Dalton et al., 2020), while other studies find that individuals are forward-looking and respond to dynamic incentives (Aron-Dine et al., 2015; Einav and Finkelstein, 2018; Klein et al., 2022). We contribute to this strand of the literature by proposing a model that features an intensive and an extensive margin and estimating it separately for four risk score groups. We find that all risk score groups are forward-looking. The degree of forward-lookingness increases in the risk score.

Another strand of the literature studies insurance design. A central theme in this literature is that cost-sharing leads to less moral hazard at the cost of more risk-taking by individuals. Another central theme is that better risks choose less coverage to benefit from lower premiums.

Cutler and Zeckhauser (2000) provide a survey of the older literature. Recent contributions to this literature include Einav et al. (2015), Kowalski (2015), and Ho and Lee (ming). We contribute to the literature on insurance design by studying the effects of alternative patient cost-sharing schemes on redistribution from low-risk to high-risk individuals with a special focus on the welfare of high-risk individuals. To the best of our knowledge, this has not been examined earlier. Redistribution between risk groups is important in the policy debate on health insurance, but has received little attention in the academic literature. We demonstrate that high-risk individuals can benefit from cost-sharing even if it leads to less redistribution towards high-risk individuals.

This study proceeds as follows: Section 2 describes the institutional setting. Section 3 describes our data. We develop our model in Section 4 and discuss identification and estimation of our model in Section 5. Section 6 presents results on parameter estimates and model fit. Section 7 presents the impact of counterfactual cost-sharing policies on redistribution between risk groups and welfare of high-risk insurees. Section 8 concludes.

2 Institutional Setting

The Netherlands has a system of managed competition in the health insurance market. Health insurance coverage is mandatory. All residents of the Netherlands have to purchase insurance from one of several competing insurance providers. Insurance is paid for by a combination of income-dependent employer contributions and premiums paid by insurees. There is a risk equalization scheme between health insurers. Premiums are community rated. Therefore, insurers cannot base premiums on individual health risk, and in addition, they cannot deny coverage for the basic health insurance package.[3]

The contents of the basic health insurance package are determined by law. Coverage is comprehensive. The basic package covers care by general practitioners (GPs), specialist and hospital care, prescription drugs, mental healthcare, and medical devices such as hearing aids and prostheses. In addition, individuals can buy supplementary health insurance for services

[3]Insurers are allowed to offer group discounts in premiums of up to 10%.

not included in the basic package, e.g. most types of dental care for adults.[4] Our study focuses on care included in the basic package.

Health insurance for the basic package features a mandatory annual deductible. Thus, individuals have to pay for a specific amount of care out-of-pocket before insurance coverage begins. The deductible resets automatically at the beginning of each calendar year, regardless of the amount of care consumed in the previous year. The amount of the deductible is set by the Dutch government. In the year 2013, the focus of our study, the amount of the deductible was €350. Individuals can choose a higher voluntary deductible of up to €500 above the amount of the mandatory deductible in return for a lower insurance premium. However, few individuals choose this option.[5] Knowledge of the deductible is almost universal in the Netherlands.[6] In general, the deductible applies to all care included in the basic package. However, some types of care are exempt from deductible payments, e.g. care provided by general practitioners, maternity care, medical equipment on rent (e.g. wheelchairs), and all care for children under the age of 18.

Healthcare charges are largely based on diagnosis-treatment-combinations (DTCs), which cover all care for an episode of treatment, including follow-up visits. Prices for DTCs are negotiated between health insurers and care providers. Patients do not pay healthcare providers directly. Providers first send bills to health insurers. Insurers then determine how much a patient has to pay out-of-pocket, depending on the remaining deductible. Hence, patients make deductible payments to insurers, not to providers.

3 Data

We use claims data from a large Dutch health insurer for the year 2013 (see Hayen et al., 2015 for details).[7] We restrict our analysis to care that is included in the basic package and that counts

[4] Supplementary health insurance is not allowed to cover deductible payments.

[5] In 2013, less than 4% of individuals in our data choose a higher voluntary deductible. We omit them from our sample.

[6] In 2010, 97% of the population knew about the deductible (Ecorys, 2011).

[7] We obtained the data to evaluate a field experiment on alternative payment forms for general practitioners. The experiment started in the year 2014, but the data also cover several years before the start of the experiment. To make sure that our results are not influenced by the experiment we use data from the year 2013.

towards the deductible. This excludes, for example, visits to the GP. We also exclude children under the age of 18 from our sample, since they are exempted from deductible payments. Our data includes information on amounts paid for claims, the date a claim was initiated, the type of claim, and demographic information of patients such as age and gender, as well as the average neighborhood income in their 6-digit postal code area.[8] Our main outcome variable is the total healthcare spending of an individual at the monthly level. To compute this measure, we aggregate payments for claims in each month, based on the date claims were initiated.

Our model and empirical approach are not well-suited to fit the data for individuals who suffer from severe health shocks. Therefore, we exclude the top 10% of spenders in the year 2013 from our estimation sample.[9] This reduces our sample from 86,497 to 77,848 individuals. We follow them over the course of the year 2013. When performing counterfactual simulations, we will assume that the top 10% of spenders do not change their behavior in response to changes in policy. This seems reasonable, as they are very likely to experience a severe health crisis and their yearly spending is far above any of the cost-sharing limits we consider in our counterfactual experiments.[10]

Table 1 reports summary statistics. The first two columns shows that average demographic and socioeconomic characteristics such as age, gender, and income at the 6-digit postcode level differ only slightly between the full sample (with the top 10% of spenders in 2013) and the estimation sample (without the top 10% of spenders in 2013). As expected, removing the highest spenders affects the metrics of care consumption substantially. For example, average spending after removing the top 10% of spenders in 2013 is reduced by more than 50%.

The last four columns of Table 1 report summary statistics for our estimation sample by risk score quartile. Risk scores are a measure of predicted healthcare spending. We compute

[8]This demographic information is very detailed. On average, there are 37 residents in a 6-digit postal code area. Neighborhood income was measured in the year 2008.

[9]In preliminary analyses, we found that for these top spenders, the time at which they hit the deductible is systematically related to subsequent spending patters. In principle, this could be accommodated when we structurally estimate our parameters, but in practice, this turned out to be difficult. The reason is that it is not feasible to allow for more unobserved heterogeneity and estimate the model for more than four groups for the risk score.

[10]The lowest yearly spending in the top decile of the spending distribution is €6370.45. Average spending per month is €1447.98. The actual policy in 2013 was a €350 deductible. The highest cost-sharing limit we consider in our counterfactual simulations is €500. Appendix Figure A.1 shows that all of the top 10% of spenders in 2013 have at least one month in which they spend more than €500. Table A.1 shows summary statistics for the top 10% of spenders.

Table 1: Summary statistics

	full sample	estimation sample				
	all	all	risk Q1	risk Q2	risk Q3	risk Q4
demographics (one observation per person)						
average age	50.59	49.78	39.92	47.47	55.98	57.67
share female	.52	.52	.49	.48	.52	.59
average income in neighborhood	2186	2205	2487	2194	2058	2045
care consumption (one observation per month and person)						
average spending	215.23	78.27	24.24	43.21	93.30	175.97
probability any spending	.419	.377	.137	.263	.505	.678
average spending if any spending	513.68	207.47	177.28	164.49	184.85	259.44
median spending if any spending	79.52	62.62	28.87	34.04	58.20	112.31
95th percentile spending if any spending	1682.49	856.00	792.24	742.43	774.47	979.70
standard deviation spending if any spending	2366.15	447.95	496.50	445.60	413.27	462.80
number individuals	86497	77848	20639	20925	20521	15763

Notes: This table reports summary statistics for the full sample and our estimation sample. Our estimation sample excludes the top 10% of spenders in 2013. The last 4 columns report statistics by risk score quartile. A detailed description of how the risk scores were calculated can be found in footnote 11. See footnote 8 for details on average income in the neighborhood. Table A.1 in the Online Appendix shows the last 5 columns for the 10% highest spenders.

risk scores for each individual in our sample using only information from the previous year and then divide our sample in four quartiles according to individuals' risk scores.[11] We see that average characteristics of individuals differ widely across risk score quartiles. Individuals in lower risk score quartiles are more likely to be younger, male, and live in neighborhoods with higher income. Our metrics of care consumption mostly increase with risk score quartile, in line with expectations.

4 Model

In this section, we describe our dynamic structural model of patient decision making. Section 4.1 provides an overview. In the subsequent sections, we provide details on the healthcare needs process (Section 4.2), the relationship between healthcare spending and out-of-pocket payments (Section 4.3), and preferences (Section 4.4). Section 4.5 describes optimal patient behavior. Finally, Section 4.6 provides a critical discussion of our modeling choices.

4.1 Overview

We model healthcare consumption at the monthly level. Patients have a finite horizon until the end of the calendar year. Each month, the sequence of events and decisions is as follows:

1. Patients enter the month with a given remaining deductible for the current calendar year.
2. They learn whether they have a healthcare need.
3. If they have a healthcare need, patients decide whether or not to visit a doctor.
4. If they decide to visit a doctor, they learn about the size of their healthcare need. They then decide how much care to consume.

[11] We use a linear regression to predict annual expenditures in year y based on age, gender, diagnosis for chronic conditions derived from pharmaceutical use, and medical spending in year $y-1$. The risk score of an individual is given by her predicted annual expenditures divided by average predicted annual expenditures. The larger the risk score, the more a person is predicted to spend, relative to the average. For our application, we use data from 2012 to construct risk scores for 2013. The same risk score measure is also used in Hayen et al. (2021) and Klein et al. (2022).

4.2 Healthcare needs

We use a two-part model for healthcare needs. The first part describes whether patients have a healthcare need. The second part describes the size of the need. This two-part model is able to capture the high frequency of months with zero care in our data and the persistence of healthcare costs over time.

Formally, h_{it} is a binary indicator for individual i having a healthcare need in month t. It is equal to 1 if an individual has a healthcare need, and equal to 0 if not. We write

$$p^0 \equiv \Pr\left(h_{it} = 1 | h_{it-1} = 0\right)$$

and

$$p^1 \equiv \Pr\left(h_{it} = 1 | h_{it-1} = 1\right).$$

The superscript denotes whether a patient had a healthcare need in the previous period. If $p^1 > p^0$, then healthcare needs are persistent.

λ_{it} is the size of the need for individual i in month t. If patients do not have a healthcare need, then $\lambda_{it} = 0$. If patients do have a healthcare need, then it is drawn from a log-normal distribution with parameters μ and σ,

$$\lambda_{it} \sim \exp\left(\mathcal{N}\left(\mu, \sigma\right)\right).$$

4.3 Out-of-pocket payments and non-medical costs

Patients face two types of costs, out-of-pocket payments and the non-medical costs of visiting a doctor.

In the Netherlands, insurance plans feature an annual deductible. For individuals in our sample, the deductible is €350. This means that individuals pay for the first €350 of annual care consumption out-of-pocket. At the beginning of month t, the remaining deductible is

$$R_{it} = \max\left\{350 - \sum_{s=1}^{t-1} c_{is}, 0\right\}, \tag{1}$$

where c_{is} is healthcare cost of patient i in month s. Out-of-pocket payments in t are then given by

$$C(R_{it}, c_{it}) = \begin{cases} c_{it} & \text{if } c_{it} \leq R_{it} \\ R_{it} & \text{if } c_{it} > R_{it}. \end{cases}$$

In addition to out-of-pocket payments for medical care, patients pay a non-medical cost to visit a doctor. This reflects e.g. traveling expenses and the opportunity cost of time. We denote this cost by κ_{it}. We assume that each period, patients draw a new value of the costs from a logistic distribution with location parameter $\bar{\kappa}$ and scale parameter s_κ.

4.4 Preferences

At any point in time, patients maximize the expected discounted sum of flow utilities until the end of the year. We use a finite horizon, because the deductible resets after the end of the year. We assume exponential discounting. Time preferences are summarized by the monthly discount factor δ.

The flow utility of individual i in month t is specified to be quadratic in the difference between healthcare consumption and healthcare needs, $c_{it} - \lambda_{it}$, and quasi-linear in out-of-pocket payments $C(R_{it}, c_{it})$,[12]

$$u(c_{it}, \lambda_{it}, R_{it}) = (c_{it} - \lambda_{it}) - \frac{1}{2\omega}(c_{it} - \lambda_{it})^2 - C(R_{it}, c_{it}). \tag{2}$$

The utility function has one parameter, ω. One advantage of this specification is that this parameter has a straightforward interpretation. If patients would maximize per period utility and ignore the dynamic part of the decision problem, then optimal consumption would be given by

$$c_{it}^* = \lambda_{it}$$

[12] Here, we follow Einav et al. (2013) who use this specification in a different context. They estimate a model for health insurance choice, in which individuals derive utility from medical expenditures covering their medical needs. Individuals choose a level of coverage, trading off higher insurance premiums against lower levels of coverage.

when patients have to pay the last euro of care consumption out-of-pocket and

$$c_{it}^{*} = \lambda_{it} + \omega$$

if they exceed the deductible limit in period t or earlier and therefore do not have to pay the last euro of care consumption out-of-pocket. This means that we can interpret ω as additional care consumption when care is free to the patient, relative to a situation in which the patient has to pay out-of-pocket for the care she consumes. This additional care consumption has been termed *ex post* moral hazard (Pauly, 1968; Cutler and Zeckhauser, 2000).

4.5 Patient decisions

Patients have a finite horizon until the end of the calendar year. Therefore, we can solve the model by backward recursion. At the beginning of each month t, a patient i knows her remaining deductible (R_{it}), whether she has a healthcare need (h_{it}) and the non-medical cost of visiting a doctor (κ_{it}).

We first discuss the situation in which a patient does not have a healthcare need ($h_{it} = 0$). We assume that then, the patient does not visit a doctor and care consumption is zero. The patient collects a flow utility equal to zero and moves to the next period.

Next, we turn to the case in which she has a need ($h_{it} = 1$). In this case, she decides whether or not to visit a doctor. She makes this decision under incomplete information because she learns the size of her need, λ_{it}, only when she visits a doctor. To determine the value of visiting a doctor, we first need to discuss the optimization problem a patient faces once she visits a doctor. In that case, she learns λ_{it} and optimally chooses c_{it} to solve the optimization problem

$$\max_{c_{it}} u(c_{it}, \lambda_{it}, R_{it}) + \delta \cdot \mathbb{E}_{h_{it+1}, \kappa_{it+1}} \left[V_{t+1}(R_{it+1}, h_{it+1}, \kappa_{it+1}) | h_{it} = 1 \right]. \tag{3}$$

$V_{t+1}(R_{it+1}, h_{it+1}, \kappa_{it+1})$ is the value function at the beginning of the next period.[13] The second term of (3) is the expected discounted sum of flow utilities in the future. The expectation is taken

[13] Recall that patients have a finite horizon and that we solve the model by backward recursion. From the perspective of period t, $V_{t+1}(R_{it+1}, h_{it+1}, \kappa_{it+1})$ is known. Moreover, when t is the last period, then $V_{t+1}(R_{it+1}, h_{it+1}, \kappa_{it+1}) = 0$.

over h_{it+1} and κ_{it+1}, as the evolution of R_{it+1} is given by (1) and known. This expectation is conditional on $h_{it} = 1$, as the probability of having a healthcare need in the next period depends on this.

Since a patient does not know λ_{it} when she decides on visiting a doctor, she has to compare the expected value of visiting a doctor with the expected value of not visiting a doctor. The value of visiting a doctor when having a healthcare need and after having paid the non-medical cost κ_{it} is given by

$$V_t^{go}(R_{it}) \equiv \mathbb{E}_{\lambda_{it}}\left[\max_{c_{it}} u(c_{it}, \lambda_{it}, R_{it}) + \delta \cdot \mathbb{E}_{h_{it+1},\kappa_{it+1}}\left[V_{t+1}(R_{it+1}, h_{it+1}, \kappa_{it+1})|h_{it} = 1\right]\right],$$

where we take the outer expectation of (3) over λ_{it}.

The value of not visiting a doctor when having a healthcare need is similar. The only difference is that we always have $c_{it} = 0$ when a patient does not visit a doctor. This gives

$$V_t^{ngo}(R_{it}) \equiv \mathbb{E}_{\lambda_{it}}\left[u(0, \lambda_{it}, R_{it}) + \delta \cdot \mathbb{E}_{h_{it+1},\kappa_{it+1}}\left[V_{t+1}(R_{it}, h_{it+1}, \kappa_{it+1})|h_{it} = 1\right]\right].$$

Note that here we use $R_{it+1} = R_{it}$, as $c_{it} = 0$ if patients decide not to visit a doctor.

Patients visit a doctor if $V_t^{go}(R_{it}) - \kappa_{it} \geq V_t^{ngo}(R_{it})$. Recall that we have assumed that κ_{it} is distributed logistic with location parameter $\bar{\kappa}$ and scale parameter s_κ. Therefore, the likelihood that a patient with a healthcare need visits a doctor is given by

$$\Pr\left(V_t^{go}(R_{it}) - \kappa_{it} \geq V_t^{ngo}(R_{it})\right) = \frac{1}{1 + \exp\left(-\frac{V_t^{go}(R_{it}) - \bar{\kappa} - V_t^{ngo}(R_{it})}{s_\kappa}\right)}. \tag{4}$$

In Appendix B.1, we provide further details on how we numerically solve the model.

4.6 Discussion

In designing our model, we have to make choices. For these, we are guided by two principles. On the one hand, our model should be well-suited for our purpose to simulate the effects of counterfactual cost-sharing policies. On the other hand, the model should be tractable and relatively simple, so that structural estimation is feasible.

For the policy options we consider, the out-of-pocket risk individuals face is small.[14] For that reason, we do not incorporate risk aversion in our model. A direct consequence is that our model cannot be used to answer the question what the optimal level of cost sharing would be.[15] However, it is well-suited for the counterfactual simulations we perform and for showing that high-risk individuals can benefit from higher levels of cost-sharing even if it leads to less redistribution from low to high risks.

Insurance contracts with a deductible—and of many other contracts with non-linear price schedules—give rise to dynamic incentives (Keeler et al., 1977). Dynamic incentives arise because deductible payments in the current period can reduce deductible payments in later periods, since patients do not have to make any further deductible payments for healthcare use in the remaining year after they have exceeded their annual deductible limit. To capture this, we estimate a dynamic structural model, in which patients make decisions throughout the year.

Our modeling of healthcare needs reflects three key features that are typical for the distribution of healthcare costs (French and Jones, 2004; Jones et al., 2013). The first key feature of this distribution is the high frequency of observations with zero healthcare spending. We account for this feature by modeling healthcare needs as a two-part process: the first part refers to the probability of having any healthcare needs; the second part refers to the distribution of healthcare needs given that there is a positive need. The second key feature is the heavy right tail that is typical for the distribution of healthcare spending. This motivates our choice of a log-normal distribution for the second part in our model of healthcare needs. The third key feature is that healthcare expenditures tend to be persistent over time. Therefore, we allow for persistence in the first part of our model for healthcare needs.

Previous studies frequently treat the patients' decision to seek any care and the decision on how much care to seek separately (e.g. Manning et al., 1987). This is also reflected in our model. Patients first decide on whether or not to visit a doctor, and in a second step they decide on the amount of healthcare spending. In reality, patients will often have some idea

[14] Section 7.1 below presents our choice of policy options.

[15] Studies on the optimal level of patient cost-sharing focus on the trade-off between moral hazard and fiancial risk for patients (Cutler and Zeckhauser, 2000; Ho and Lee, ming) In our study, allowing for risk aversion would involve additional assumptions and lead to additional challenges. For instance, our data are not directly informative about risk aversion, because the deductible was low at the time of our study (see also footnote 2).

about how high the costs of treatment will be if they visit a doctor. In our model, we make the simplifying assumption that patients know only that they have a healthcare need. The exact size of the healthcare need is only revealed after a patient sees a doctor. We chose this model specification for three reasons: First, we wanted to allow for some uncertainty about the cost of treatment from the patient's perspective. This captures both the uncertainty about the price the doctor charges for the treatment and some uncertainty about the diagnosis. Second, evidence from previous studies shows that patients' responses to cost-sharing incentives for (expensive) inpatient care and (less expensive) outpatient care are about equal in relative terms (Manning et al., 1987; Shigeoka, 2014). This stylized fact is reflected in our model: since patients learn about the size of the healthcare need only after they visit a doctor, they will respond to cost-sharing incentives by reducing visits for expensive and less expensive needs in equal proportion. Third, it makes the model tractable and aids identification. A model in which patients select into visiting a doctor based on partial knowledge about their healthcare needs is harder to solve, as we would need to model the formation of their beliefs about their healthcare needs and then integrate over those beliefs. It is also harder to identify, as beliefs and healthcare needs are not observed in our data.

Finally, our model features a price response. The parameter ω can be interpreted as additional care consumption by patients who have a healthcare need and visit a doctor when care is free from the patient's perspective, relative to a situation in which the patient has to pay out-of-pocket for the care she consumes. This means that in our model, moral hazard is additive. We choose this specification for three reasons. First, it is in line with findings from previous studies, which have shown that patients tend to respond to cost-sharing incentives by reducing all types of care, e.g. both high-value care and low-value care (Manning et al., 1987; Einav and Finkelstein, 2018). Second, we see it as an advantage that our parameter has a straightforward interpretation. Third, it makes our model tractable and aids identification. If ω would instead be a random variable that is allowed to be correlated with λ_i, then it would take considerably longer to solve the model. Also, it would be very hard to credibly estimate the parameters of the joint distribution of healthcare needs and moral hazard given that we do not observe healthcare needs. Note, however, that we estimate the model separately for each risk score quartile.

Thereby, we allow for some heterogeneity of ω in the population.

5 Identification and estimation

5.1 Identification

We are able to estimate the model parameters if the model is identified. The model is identified if there is only one unique combination of parameter values for which the model can generate the observed patterns in the data. In the following, we informally discuss how the data are informative about the parameters of our model. We first discuss this for the parameters that affect healthcare costs at the intensive margin and then for parameters that affect healthcare costs at the extensive margin.

5.1.1 Intensive margin

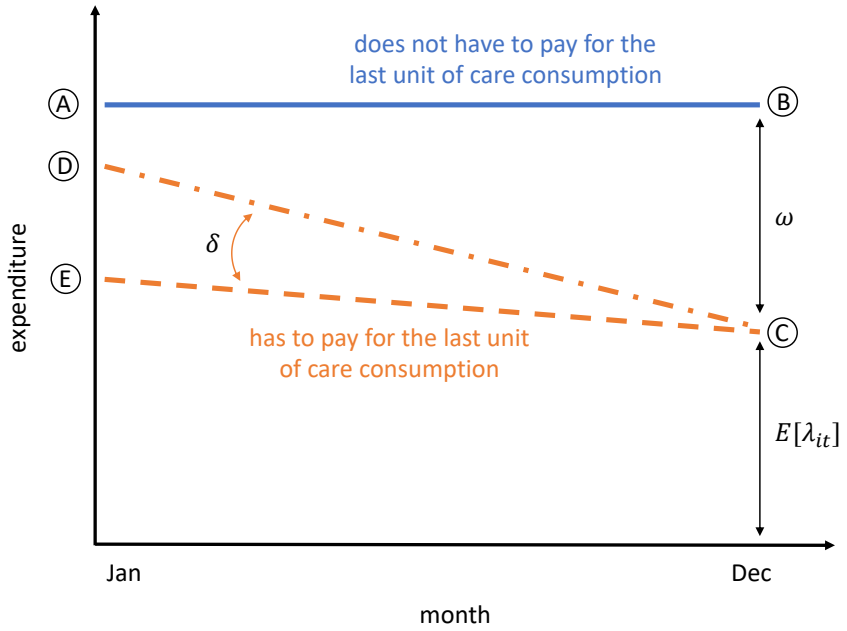
In this section, we discuss identification of the parameters μ and σ of the distribution of healthcare needs, the moral hazard parameter ω, and the discount factor δ.

Recall that patients do not know the size of the healthcare need, λ_{it}, when they decide whether they visit a doctor. They only know whether or not they have a healthcare need. This implies that patients do not select into visiting a doctor based on the size of their healthcare needs λ_{it}. Therefore, the distribution of healthcare costs conditional on them being positive is informative about the distribution of needs, λ_{it}.

Recall also that our utility function in (2) is specified so that healthcare spending is equal to $\lambda_{it} + \omega$ when patients do not have to pay for the last unit of care they receive. This means that the distribution of healthcare spending conditional on any spending, for patients who have no remaining deductible at the beginning of the period, is the distribution of $\lambda_{it} + \omega$.

To separately identify the mean of the healthcare needs λ_{it} and the scalar parameter ω, we turn to the last month of the year. This is useful because in December patients solve a static decision problem. Recall that then, healthcare spending is equal to λ_{it} when patients have to pay for the last unit of care they receive. This means that in December, the distribution of spending when patients visit a doctor and do have to pay out-of-pocket is the distribution of λ_{it}.

Figure 1: Identification



Notes: This figure illustrates how the mean of λ_{it}, ω and δ are identified. It shows expenditure at the intensive margin, separately over time and for patients who do and do not have to pay for the last unit of care they consume in that month. Point A is care consumption at the beginning of the year when patients do not have to pay. Point B is care consumption at the end of the year when patients do not have to pay. Point C is care consumption at the end of the year when patients have to pay. Point D and E are care consumption at the beginning of the year when patients have to pay. Point D is for a higher value of δ than point E. See Section 5.1.1 for details on the identification argument.

Figure 1 illustrates this. It shows average expenditure at the intensive margin, separately for each month and for patients who do not have to pay for the last unit of care in that month and for patients who do have to pay. When patients do not have to pay for the last unit of care consumption, then spending is given by point A at the beginning of the year and point B at the end of the year. C is care consumption at the intensive margin in December when patients do have to pay for the last unit of care they consume.

Next, we turn to identification of the discount factor δ.[16] The discount factor captures how much patients take the effects of current choices on future payments into account. To see which variation in the data is informative about this, think of a patient who has a healthcare need early in the year and has to pay for care consumption out-of-pocket. If $\delta = 0$, she will simply solve a static decision problem and care consumption will be equal to λ_{it}, as it would be in December. However, if $\delta > 0$, she will take into account that any euro that is spent on care out-of-pocket

[16]The intuition we provide here is based on theoretical results developed in the Online Appendix of Klein et al. (2022). These generalize results by Keeler et al. (1977) and Ellis (1986).

today will make it more likely that she will not have to pay for care consumed later in the year, since care above the deductible limit is free from the patient's perspective. This can be seen as a bonus (a reduction of expected out-of-pocket payments in the future) that is paid out with a delay and is therefore discounted. The less patients discount the future, i.e. the higher δ, the higher is the value of the bonus, which leads to higher care consumption in the current period.

Figure 1 illustrates this. The dash-dotted line is intensive margin care consumption for a high value of δ, when patients pay for the last unit out-of-pocket. The dashed line is intensive margin care consumption for a low value of δ, when patients pay for the last unit out-of-pocket. In the last period, patients solve a static decision problem, and therefore, care consumption does not depend on δ. It is given by point C in either case. In earlier periods, a higher δ implies higher care consumption. Point D is care consumption at the beginning of the year when patients discount the future less and have a high discount factor. Point E is care consumption when patients discount the future more and have a low discount factor.

5.1.2 Extensive margin

In this section, we discuss identification of the parameters $\bar{\kappa}$ and s_κ of the distribution of κ_{it} and the parameters p^0 and p^1 of the healthcare needs process.

The point of departure is identification at the intensive margin. Above we have argued that the data are informative about the parameters that affect choice at the intensive margin, μ, σ, ω, and δ. For the following discussion, we therefore treat these parameters as known. This means that we know $V_T^{go}(R_{it})$ and $V_T^{ngo}(R_{it})$. These are the values to visiting and not visiting a doctor in the last period, for any remaining deductible R_{it}, respectively. Consider the case in which the patient visited a doctor in the previous period. This means that the patient had a healthcare need in the previous period, which in turn means that the probability that she has a healthcare need in the current period is given by p^1. Hence, the observed probability that the patient will visit the doctor is given by p^1 times (4),

$$p^1 \cdot \frac{1}{1+\exp\left(-\frac{V_T^{go}(R_{iT})-\bar{\kappa}-V_T^{ngo}(R_{iT})}{s_\kappa}\right)}.$$

This probability is a nonlinear parametric function of p^1, $\bar{\kappa}$, s_κ, and of R_{iT}. We observe this probability for different values of R_{iT}. Therefore, we can solve for p^1, $\bar{\kappa}$, and s_κ.

This leaves us with the question how p^0 is identified. This is a challenge since some patients who have a healthcare need still don't visit a doctor, and thus the probability of visiting a doctor given that patients had no healthcare need in the previous period is unobserved. We assume that the healthcare needs process is ergodic. This allows us to express p^0 as a function of p^1 and the unconditional probability of visiting a doctor, which we denote by π. p^1 is identified by the argument spelled out above, which means that p^0 is identified when π is identified. For this, we can make an argument that parallels the one we made above. In particular, the unconditional probability that a patient visits a doctor, which is observed, is given by

$$\pi \cdot \frac{1}{1+\exp\left(-\frac{V_T^{go}(R_{iT})-\bar{\kappa}-V_T^{ngo}(R_{iT})}{s_\kappa}\right)}.$$

As before, π is identified from variation in R_{iT}.

Once we know π, we can recover p^0, as it holds for the stationary distribution that

$$\begin{bmatrix} 1-\pi & \pi \end{bmatrix} \begin{bmatrix} 1-p^0 & p^0 \\ 1-p^1 & p^1 \end{bmatrix} = \begin{bmatrix} 1-\pi & \pi \end{bmatrix},$$

which implies

$$p^0 = \frac{\pi(1-p^1)}{1-\pi}.$$

5.2 Estimation

For given parameters, we can numerically solve the model described in Section 4. This is done by backward recursion, starting with the last month. Appendix B.1 provides details.

Once we have solved the model for given parameters, we can simulate care consumption. This in turn enables us to estimate the parameters of the utility function, the healthcare needs process, and the non-medical costs of visiting a doctor using the generalized method of moments (GMM). We estimate all model parameters separately for individuals in each risk score quartile.

We use 4 sets of moment conditions. These are related to the probability of any spending,

the probability of any spending given spending in the previous period, mean spending, and the variance of log spending. We calculate the first three sets of moments separately for 4 intervals for the remaining deductible and for each month except January.[17] For the variance of log spending, we use the same months and use only observations that have crossed the deductible. This means that we use $3 \cdot 11 \cdot 4 + 11 = 143$ moments. Appendix B.2 provides details.

6 Results

6.1 Parameter estimates

Table 2 reports our estimates of the structural parameters for each risk score quartile. Our estimates of ω are large for the first three risk score quartiles and smaller for the fourth quartile

The 95% confidence intervals for δ do all not contain 0, which suggests that patients are forward-looking. δ is higher for higher risk score quartiles. The lowest risk score quartile has a monthly discount factor of 0.74, while the highest risk score quartile has a monthly discount factor of 0.94.

Our estimates for the mean of the net non-medical costs of visiting a doctor, $\tilde{\kappa}$, which consist of the sum of the non-medical costs of visiting a doctor and the expected flow utility from not visiting a doctor (see Appendix B.1.3), are somewhat higher (less negative) for higher risk score quartiles. This could either be the case because higher risk scores have a higher non-medical cost of visiting a doctor or a lower expected flow dis-utility from not visiting a doctor. The distribution of the net non-medical costs of visiting a doctor is rather wide. The scale parameter s_κ varies from 6.46 to around 25.72, meaning that the standard deviation of the distribution of κ_{it} varies from around 11.72 to around 46.65.[18]

There is a large difference in the probabilities of having a healthcare need across risk score quartiles: conditional on not having a healthcare need in the previous period, the chance of having a healthcare need in the current period, p^0, is more than twice as high for individuals

[17] The reason for excluding January is that January expenditures could partly be driven by healthcare needs from the previous year. For instance, some patients may not have been able to see a doctor in December because of the holiday season.

[18] Recall that κ_{it} follows the logistic distribution. The standard deviation of the distribution of κ_{it} is equal to s_κ times $\pi/\sqrt{3}$.

Table 2: Parameter estimates

	risk score quartile							
parameter	Q1		Q2		Q3		Q4	
moral hazard (ω)	87.13	(8.36)	71.05	(4.92)	73.33	(4.51)	17.37	(9.52)
monthly discount factor (δ)	0.74	(0.02)	0.79	(0.01)	0.91	(0.00)	0.94	(0.01)
mean net non-medical cost of visiting doctor[a] ($\tilde{\kappa}$)	-59.62	(5.01)	-57.88	(2.63)	-50.62	(1.76)	-46.16	(3.26)
standard deviation net non-medical cost of visiting doctor (s_κ)	6.46	(0.79)	6.93	(0.46)	23.06	(0.72)	25.72	(1.77)
likelihood need if no need previous month (p^0)	0.23	(0.00)	0.35	(0.00)	0.52	(0.00)	0.47	(0.05)
likelihood need if need previous month (p^1)	0.41	(0.01)	0.51	(0.00)	0.70	(0.00)	0.92	(0.02)
location parameter healthcare need (μ)	3.29	(0.17)	3.32	(0.09)	3.63	(0.05)	3.75	(0.06)
scale parameter healthcare need (σ)	1.74	(0.06)	1.68	(0.04)	1.51	(0.02)	1.69	(0.02)

Notes: This table reports parameter estimates for our model. Standard errors are in parentheses. See Section 5.2 and Appendix B.2 for details.
[a] $\tilde{\kappa}$ is the sum of the mean of the fixed cost of visiting a doctor, $\bar{\kappa}$, and the flow utility from not visiting a doctor. See Section B.1.3 in the appendix for details.

in the 4^{th} risk score quartile than for individuals in the 1^{st} risk score quartile (0.47 vs. 0.23). Healthcare needs are persistent, as p^1 is estimated to be substantially higher than p^0 for each risk score quartile. This means that it is more likely that a patient has a healthcare need when she has a healthcare need in the previous month.

The location parameter μ of the healthcare needs distribution is increasing in the risk score quartile, while the scale parameter σ does not increase or decrease in the risk score quartile. Expected healthcare needs are given by $\exp(\mu + \sigma^2/2)$. For the four risk score groups, they are €121.12, €113.22, €118.10, and €178.80, respectively.

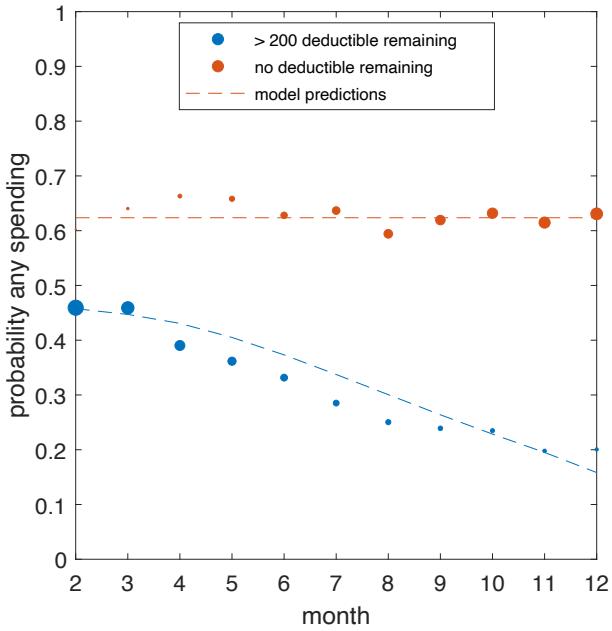
6.2 Model fit

To assess the fit, we compare patterns in the estimations sample with predictions from the model. Figure 2 shows results for the 3^{rd} risk score quartile and for 2 remaining deductible groups: those above the deductible at a given point in time and those with more than €200 remaining deductible at a given point in time. Figure 2a shows results for the extensive margin. Figure 2b shows results for the intensive margin. Appendix Figures A.3 and A.5 show results for other risk score quartiles. In addition, Appendix Figure A.4 shows results for the extensive margin when we condition on any spending in the respective previous month. In each figure, the size of dots is proportional to the respective number of observations.

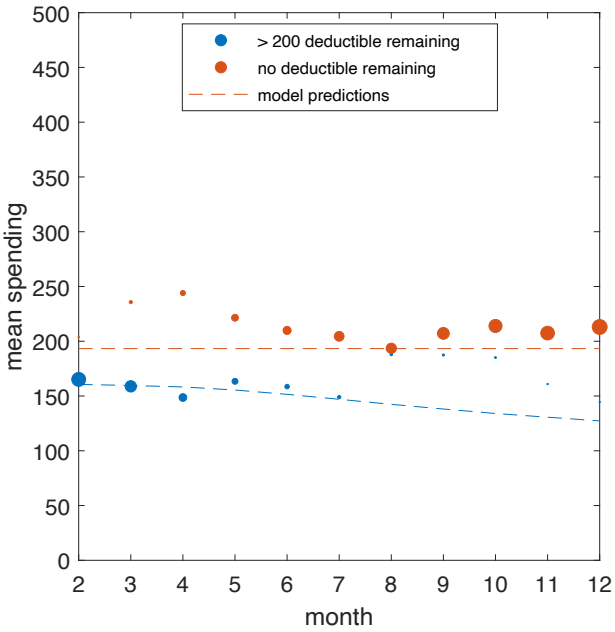
Overall, our model performs quite well in matching the data. As predicted by our model, individuals who do not have any deductible remaining (and thus face no cost-sharing) have a rather stable probability of having any spending across all months in the year. For individuals with a remaining deductible of more than €200, the probability of any spending decreases over the course of the year. The model rationalizes this via forward-looking behavior: in earlier months, there is a higher chance that individuals exceed the deductible before the end of the year, which lowers their effective price for healthcare needs. The model also performs quite well in matching intensive margin moments.

Figure 2: Model fit

(a) Probability of any spending



(b) Spending conditional on any spending



Notes: The figures plot the probability of any spending (top) and mean spending conditional on any spending (bottom) against months for the 3^{rd} risk score quartile. We distinguish between individuals with more than €200 remaining deductible at the beginning of the respective month and those who have no deductible remaining at the beginning of the respective month. The dots are the means in the data. The size of dots is proportional to the respective numbers of observations. The dashed lines are the predictions generated by our model. Figure A.3 shows the fit at the extensive margin for all 4 risk score quartiles. Figure A.4 does the same but conditions on spending in the respective previous period. Figure A.5 shows the fit at the intensive margin for all 4 risk score groups.

7 Patient cost-sharing, redistribution, and welfare

7.1 Policy options

We now use our model to study whether more cost-sharing leads to less redistribution from low-risk to high-risk individuals and lower welfare of high-risk individuals. We perform simulations for a range of counterfactual policies. The alternative patient cost-sharing policies are: 1) no patient cost-sharing, 2) a deductible of €150, 3) a deductible of €350 (the policy in place in the year 2013), 4) a deductible of €500, 5) a deductible of €700 over a 2-year period, 6) a donut hole contract for which the first €350 of annual care is cost-free from the patient's perspective, followed by a deductible of €350 and full insurance coverage afterwards, 7) a co-insurance rate of 75% (with an out-of-pocket maximum of €350, 8) co-payments of €30 for each month with medical treatment, 9) co-payments of €50 for each month with medical treatment, and 10) a deductible of €350 with a monthly cap of €150.

This choice of policy options is motivated both by the Dutch policy debate and by previous academic studies. In the Netherlands, patient-cost sharing is an important and controversial topic in the political debate. In general, more left-wing parties favor lower or no patient cost-sharing whereas more right-wing parties favor keeping the current patient cost-sharing scheme. Many of the policy options we examine were proposed by political parties in their party programs for the national elections in the year 2021. While some parties proposed to keep the mandatory deductible at the current level, others suggested halving it or abolishing it. One party suggested replacing the deductible with a co-insurance rate up to a maximum amount, and some parties suggested replacing the deductible for specialist care by fixed co-payments.[19] No party suggested increasing the size of the deductible. We nonetheless study the effects of a €500 deductible. A current proposal by the Dutch Ministry of Health, Welfare, and Sports suggests a deductible with a maximum payment of €150 per treatment, which motivates our analysis for a deductible with a monthly cap. The importance of the length of the deductible period for health insurance contracts with a deductible is pointed out already by Keeler et al.

[19] VVD and SGP proposed to keep the mandatory deductible at the current level. PvdA and GL suggested to halve the mandatory deductible. SP and DENK want to abolish it. CDA suggested to replace the deductible by a co-insurance rate up to a maximum amount. D66 and CU suggested to replace the deductible for specialist care by fixed co-payments (CPB, 2021).

(1977). This motivates our evaluation of a 2-year deductible period. van Kleef et al. (2009) suggest that donut hole contracts, where the deductible starts after a certain amount of free consumption, are better at reducing moral hazard than traditional deductibles. This motivates our evaluation of a donut hole contract. Donut hole contracts were also used for Medicare Part D plans in the United States.[20]

Our model allows us to evaluate the partial equilibrium effects of changes to contracts. This means that we do not take into account possible changes in prices or provider behavior. We can obtain predictions for alternative contract designs by changing the cost function in (2), $C(.,c_{it})$. For counterfactual contracts with different deductible amounts we adjust the range of possible deductible amounts individuals are allowed to have in the model. Counterfactuals that change the unit cost of utilizing healthcare, such as coinsurance contracts, can be implemented by multiplying the cost function, $C(.,c_{it})$, by the new unit cost. Similarly, we can also adjust the cost function, $C(.,c_{it})$, for evaluating counterfactual contracts with a donut hole design and with fixed co-payment amounts. Contracts that apply for different lengths of time (as opposed to the ones that apply for a single year) require changing the final time period, T, for the model.[21] For the yearly €350 deductible with a monthly cap, we use the cost function for the yearly deductible, but apply it to costs that are capped at €150 per month.

The policy options differ in multiple dimensions. Therefore, we cannot order all of them with respect to the strength of financial incentives or the "amount" of cost-sharing.[22] It is however possible to order deductible contracts according to their cost-sharing limits. Higher deductibles always provide incentives that are at least as strong as lower deductibles. Therefore, in the remainder, we will sometimes focus in our discussion on comparing the €350 deductible contract that was in place to deductible contracts with other cost-sharing limits and the contract with no cost-sharing.

[20]This was phased out in 2020. Since then, there is a so-called coverage gap. Insurees pay 25% of drug costs out-of-pocket in the coverage gap.

[21]For annual contracts ($T = 12$), we use the exact same health shocks across policies to ensure changes in outcomes stem entirely from changes in consumption choices due to differences in incentives across contracts. For contracts longer than a year ($T > 12$), we use the same health shocks for the first 12 months and generate new ones for later months.

[22]To see this, compare the €350 deductible to the donut hole contract with a donut hole from €350 to €700. For some individuals, the donut hole contract provides stronger incentives than the deductible. These are individuals who are likely to have yearly healthcare spending between €350 and €700. For others, it provides less strong incentives. This is the case for individuals who are likely to spend less than €350.

7.2 Overall effects of counterfactual policies

Table 3 reports average monthly outcomes for our estimation sample. Turning first to different deductible amounts, we see that care consumption decreases in the size of the deductible (column 1). Both changes at the intensive margin (column 2) and at the extensive margin (column 3) contribute to the reduction in spending.[23] When we move from no cost-sharing to a €350 deductible, intensive margin spending decreases by 15.6% and the probability of any spending decreases by 24.3%. For higher deductibles, fewer patients exceed the deductible limit by the end of the year (column 4). Columns 5 and 6 show the mean and the standard deviation of out-of-pocket payments, respectively. Out-of-pocket payments rise with increases in the deductible amount, but only up to a deductible of €350. When the deductible increases from €350 to €500, out-of-pocket payments decrease. The reason for this is that patients consume less care. The standard deviation of out-of-pocket payments always increases for higher deductible amounts, indicating an increase in the financial risk borne by patients in contracts with higher deductibles. We also calculate the semi-arc elasticity of healthcare spending with respect to the cost-sharing rate. This is defined as the percentage change in healthcare consumption divided by change in price (see e.g. Aron-Dine et al., 2013). It can be interpreted as a measure for the strength of the response to the financial incentives the respective policy provides. To compute this measure, we first calculate the change in average healthcare spending relative to no cost sharing. To turn this into a percentage change, we divide it by the average spending between the two policies. Then, we divide this percentage change by the average out-of-pocket payments under cost sharing divided by the average healthcare spending. Column 8 of Table 3 shows that increases in the size of the deductible lead to a change in the semi-arc elasticity from -0.35 for a €150 deductible to -2.79 for a €500 deductible.[24]

[23] Table A.2 in the Online Appendix reports the probability that individuals with a healthcare need choose to not visit a doctor. This probability is generally higher for persons with higher risk scores. Contracts that lead to lower spending also tend to lead to a higher probability of not visiting a doctor despite having a healthcare need. In Section 7.5 below, we discuss whether this is a cause of concern in the present context.

[24] Although these values seem rather large relative to the frequently cited elasticity estimate of -0.2 found in the RAND health insurance experiment (Manning et al., 1987; Keeler and Rolph, 1988; Aron-Dine et al., 2013), these two quantities are not comparable. Brot-Goldberg et al. (2017) calculate semi-arc elasticities for the RAND health insurance experiment for a specific change in contracts to be -2.11 (see page 19 of the NBER working paper version of their paper). We present semi-arc elasticities instead of elasticities, because semi-arc elasticities take differences in co-payment rates between contracts into account, whereas elasticities are just percentages changes in quantity if one price is 0, which is the case for the comparisons in our study.

Table 3: Effects of counterfactual policies in estimation sample

policy option	spending	any spending	spending if any	fraction hitting[d]	out-of-pocket	standard deviation OOP[e]	semi-arc elasticity[f]
no cost sharing	97.04	0.51	192.72	1.00	0.00	0.00	
€150 deductible	92.79	0.50	182.46	0.91	11.75	1.85	-0.35
€350 deductible (status quo)	66.35	0.38	162.62	0.52	18.16	9.88	-1.37
€500 deductible	31.26	0.19	146.57	0.19	11.51	12.62	-2.79
Two-year €700 deductible[a]	37.36	0.21	147.88	0.29	10.20	6.98	-3.25
Donut hole from €350 to €700[b]	71.39	0.41	173.87	0.64	12.31	12.42	-1.77
75% coinsurance with €350 maximum[c]	76.98	0.45	167.63	0.57	21.75	8.86	-0.82
Co-payment €30	89.67	0.47	192.92		14.02	4.83	-0.51
Co-payment €50	80.78	0.42	193.03		21.08	7.96	-0.70
€350 deductible with monthly €150 cap	74.95	0.45	160.95	0.62	20.65	8.87	-0.93

Notes: The table reports outcomes produced by our model (columns) for different insurance contracts (rows). We used estimates from Table 2 to simulate these outcomes. The table shows averages across individuals and months. To obtain those, we first calculate averages for each risk score quartile and then take a weighted average across these risk score quartiles, with weights equal to the share of each risk score quartile in our data (proportional to the numbers of observations in the last four columns of Table 1).

[a] The deductible resets after 2 years. The cost-sharing limit for this 2-year period is twice the status quo for the one-year deductible.

[b] The donut hole contract has the following structure: the first €350 of healthcare are free, the next €350 of healthcare have to be paid out-of-pocket. After that, care is free again.

[c] A coinsurance of 75% is accompanied by an out-of-pocket maximum of 350 euros.

[d] Fraction hitting refers to the share of individuals with a marginal price of 0 at the end of the contract duration. This is not well-defined for co-payments. For the last policy option, this is the fraction of individuals who exhaust the deductible.

[e] Standard deviation OOP denotes the standard deviation of the yearly out-of-pocket payment divided by 12.

[f] The semi-arc elasticity is calculated as the ratio of two quantities. The first quantity is the difference in healthcare consumption between the respective policy and no cost sharing, divided by the average of those two quantities. The second quantity is the average out-of-pocket payment (column 5) divided by average spending (column 1).

Table 4: Overall effects of counterfactual policies

policy option–add–	spending	out-of-pocket	premium
no cost sharing	232.13	0.00	232.13
€150 deductible	228.29	11.82	216.47
€350 deductible (status quo)	204.50	19.26	185.25
€500 deductible	172.92	14.52	158.40
Two-year €700 deductible	178.41	12.10	166.32
Donut hole from €350 to €700	209.04	14.00	195.04
75% coinsurance with €350 maximum	214.07	22.49	191.58
Co-payment €30	225.49	15.62	209.87
Co-payment €50	217.49	23.97	193.52
€350 deductible with monthly €150 cap	212.24	21.50	190.74

Notes: The table reports average outcomes (columns) for a number of counterfactual policies (rows). The reported averages are for the full sample. To obtain them, we simulate outcomes for our estimation sample (90%). We then combine them with averages for the top spenders (10%). See notes to Table 3 for details on the policies and on the way we simulate averages for our estimation sample. For the top spenders (10%), we assume that their monthly spending is not affected by the policy changes. See Section 3 for a related discussion. We assume that out-of-pocket spending by top spenders is the yearly out-of-pocket maximum divided by 12. The premium is defined in (5). It is given by the difference between the first column and the second.

The remaining rows show results for the other counterfactual policies. Overall, we find that spending is lowest for a €500 deductible contract, closely followed by the 2-year €700 deductible. These are also the contracts with the lowest fraction of patients exceeding the cost sharing limit and the highest (in terms of magnitude) semi-arc elasticity.

Our research question is whether more cost-sharing leads to less redistribution between risk groups and lower welfare of high-risk individuals. This does not only concern the 90% of the individuals in our estimation sample, but also the 10% highest spenders that we have excluded from it. From now on, we show results for the full sample. For this, we combine the results from our counterfactual simulations, which make use of our model, and data on spending by the excluded individuals. We assume throughout that spending of the 10% highest spenders does not vary across counterfactual policies. But of course, cost sharing limits are different and therefore, out-of-pocket payments vary also for the top 10% spenders. We assume that their out-of-pocket payments are always equal to the maximum under the respective policy we consider.

Table 4 shows that predicted spending for the full sample is always substantially higher than for the estimation sample (Table 3). This is in line with the summary statistics presented in the first two columns in Table 1. For the €350 deductible, predicted spending for the full sample

is €204.50. For a €500 deductible, it is €172.92. This suggests that increasing the deductible to €500 leads to a reduction in spending by 15.4%. The second column of Table 4 shows that average out-of-pocket payments decrease by 24.6%, from €19.26 to €14.52 per month.

In the Netherlands, insurance plans are community rated and most health insurance companies (including the one that provided our data) are not-for-profit cooperatives. For the purpose of our simulations, we assume that insurance premiums are equal to average expenditures minus average out-of-pocket payments.[25] That is, the premium under policy τ is

$$premium(\tau) = \bar{c}(\tau) - \overline{oop}(\tau), \tag{5}$$

where $\bar{c}(\tau)$ is average monthly spending in the full sample and $\overline{oop}(\tau)$ are average monthly out-of-pocket expenditures in the full sample.

The third column of Table 4 shows that increasing the annual deductible from €350 to €500 leads to a reduction in the monthly premium by 14.5%, from €185.25 to €158.40.

7.3 Redistribution between risk groups

The Dutch health insurance system redistributes resources from individuals with low health risks to individuals with high health risks. We now explore how such redistribution differs across alternative insurance plans.

We measure redistribution with a simple metric: the euro amount of care received by individuals in a risk score quartile minus the amount paid by these same individuals, which includes both premiums and out-of-pocket payments. Specifically, our measure of redistribution for risk score quartile r under policy τ is

$$redistribution_r(\tau) = \bar{c}_r(\tau) - \overline{oop}_r(\tau) - premium(\tau), \tag{6}$$

where $\bar{c}_r(\tau)$ is average monthly healthcare spending of risk score quartile r under policy τ and

[25] For three reasons, the levels of premiums computed in our study differ from actual premiums patients pay when they buy insurance. First, our model only covers care that counts towards the deductible (Section 3). Second, premiums in our simulations do not take employer contributions into account. Third, they also don't take administrative costs into accounts. For our purpose, the level of premiums is not relevant. Rather, it is relevant how premiums are affected by changes to patient cost-sharing schemes.

Table 5: Spending by risk score in full sample

	risk score quartile			
policy option	Q1	Q2	Q3	Q4
no cost sharing	79.37	120.19	201.24	523.83
€150 deductible	72.40	115.21	199.38	522.19
€350 deductible (status quo)	43.44	82.57	178.90	508.90
€500 deductible	26.66	54.56	135.66	471.25
Two-year €700 deductible	23.93	49.74	140.76	495.44
Donut hole from €350 to €700	67.23	102.86	167.46	495.26
75% coinsurance with €350 maximum	62.63	102.53	180.24	507.11
Co-payment €30	79.37	120.18	195.11	503.60
Co-payment €50	79.36	120.09	184.93	482.17
€350 deductible with monthly €150 cap	60.92	98.85	178.33	507.10

Notes: The table reports spending under a number of policies (rows). See notes to Table 3 for details on the policies. The table reports monthly averages for the full sample. We obtain those averages by combining results for counterfactual simulations for our estimation sample with results for the 10% top spenders. Table A.4 in the Online Appendix reports results for our estimation sample.

$\overline{oop}_r(\tau)$ are the average monthly out-of-pocket payments by risk score quartile r.[26]

Table 5 reports spending by risk score quartile under the policies we consider. It shows that for each contract, mean spending is higher for higher risk scores. There are remarkable differences in the way contracts affect spending across risk score quartiles. For instance, a €350 deductible reduces spending relative to lower deductibles or no cost-sharing especially strongly for the first risk score quartile, and less so for the fourth risk score quartile; increasing the deductible from €350 to €500 reduces care more for the fourth risk score quartile than for the first risk score quartile.[27]

In Table 6, we report our measure of redistribution for each risk score quartile and each

[26]Note that the premium, as defined in (5), does not depend on the risk score. If it would be risk score-specific, then redistribution between risk groups would be 0 by definition. Also by definition, our measure of redistribution is 0 when we take the (weighted) average across groups. The weights are proportional to the numbers of individuals in the last row of Table 1.

[27]The table also shows that it is possible to shift financial incentives to higher risk scores using a donut hold contract. We see that the donut hole contract leads to higher spending for low-risk patients and lower spending for high-risk patients. The reason is that under a donut hole contract, low-risk individuals are less likely to pay out-of-pocket for the last unit of care consumed in the year, while high-risk individuals are more likely to pay out-of-pocket for the last unit of care consumed in the year. One can also design donut hole contracts in which the beginning of the donut hole differs across risk score. In Figure A.6 in the Online Appendix, we quantify the effects of varying the beginning of the donut hole, while keeping the length constant at €350. We find that a donut hole contract that starts at 0, which is the standard deductible contract, leads to the lowest spending for the first two risk score quartiles. For the third risk score quartile, a donut hole contract that starts at about €310 leads to the lowest spending. For the fourth risk score quartile, a donut hole contract that starts at about €270 leads to the lowest spending. The associated reductions in spending relative to a deductible with the same cost-sharing limit are about €13 and €18, respectively.

Table 6: Redistribution between risk groups in full sample

	risk score quartile			
policy option	Q1	Q2	Q3	Q4
no cost sharing	-152.75	-111.94	-30.89	291.71
€150 deductible	-154.36	-113.26	-29.58	293.26
€350 deductible (status quo)	-150.73	-117.50	-32.28	296.72
€500 deductible	-134.51	-109.06	-44.07	284.54
Two-year €700 deductible	-143.46	-118.81	-44.35	303.27
Donut hole from €350 to €700	-135.05	-103.33	-44.52	279.80
75% coinsurance with €350 maximum	-144.86	-110.23	-37.55	289.13
Co-payment €30	-139.23	-102.91	-33.35	272.06
Co-payment €50	-128.72	-95.45	-36.93	258.03
€350 deductible with monthly €150 cap	-143.72	-111.34	-38.14	289.72

Notes: The table reports our measure of redistribution under a number of policies (rows). Our measure is defined in 6. It is equal to care consumption minus out-of-pocket payments minus premiums. See notes to Table 3 for details on the policies. This table reports monthly averages for the full sample. We obtain those averages by combining results for counterfactual simulations for our estimation sample with results for the 10% top spenders. Table A.4 in the Online Appendix reports results for our estimation sample. See Section 7.3 for further details.

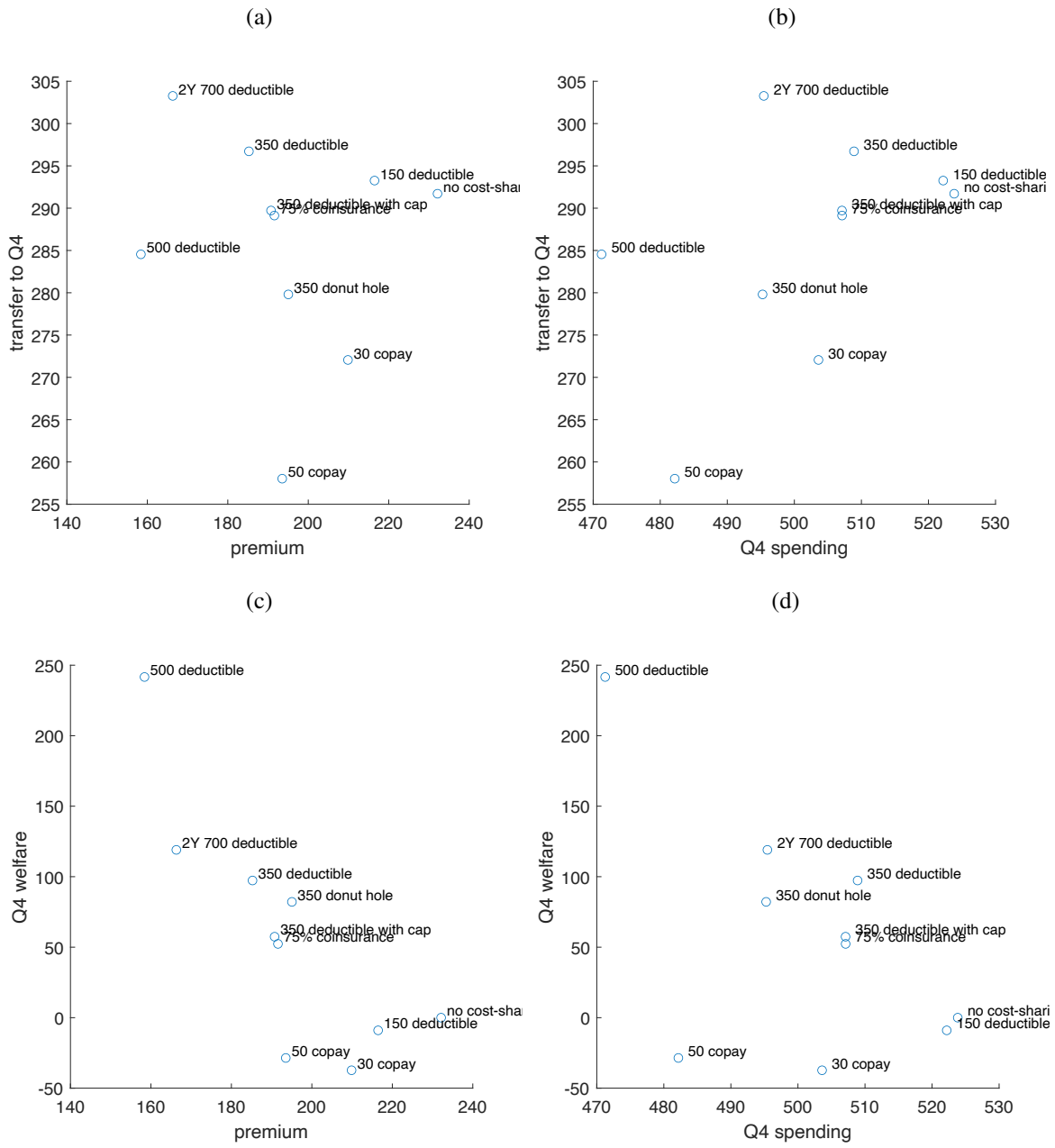
policy option. In line with expectations, health insurance redistributes money from low-risk to high-risk individuals. This holds regardless of the specifics of the contract. Among the one-year deductible contracts, redistribution from low-risk to high-risk individuals is largest for the €350 deductible contract that was in place in 2013. The healthiest individuals (the first risk score quartile) pay €150.73 per month in excess of what they use. The sickest individuals (the fourth risk score quartile) use €296.72 of care per month in excess of what they pay. This redistribution is large compared to average income: a loss of 7.0% of average disposable income for the first risk score quartile and a gain of 13.7% for the fourth risk score quartile.[28]

Interestingly, reducing the deductible to €150 (or a move to no cost sharing) would lead to a similar amount of redistribution from low-risk to high-risk individuals. However, increasing the deductible to €500 leads to less redistribution between the lowest and highest risk score quartile.

Figure 3 shows general patterns. Figure 3 (a) shows that policies that lead to lower premiums do not generally lead to less redistribution towards individuals in the fourth risk score quartile. Figure 3 (b) shows that policies that lead to lower spending by individuals in the fourth risk

[28] Average available income per month adjusted for household size in the Netherlands in the year 2013 was €2158. Source: Statistics Netherlands https://www.cbs.nl/nl-nl/cijfers/detail/83739NED?q=besteedbare%20inmomsten%202013, accessed June 20, 2022.

Figure 3: Policy effects



Notes: These figures visualize the relationship between transfers to Q4 (last column of Table 6) and Q4 welfare (last column of Table 7) on the one hand and premiums (last column in Table 4) and Q4 spending (last column of 5) on the other hand. Reported for full sample.

score quartile do also not generally lead to less redistribution towards individuals in the fourth risk score quartile. In the next section, we discuss Figures 3 (c) and (d).

7.4 Welfare

Our model-based measure of welfare that is associated with policy τ is given by

$$\begin{aligned} W(\tau) =& \{(1-\Pr(h_{i1}=1))\cdot\delta\cdot\mathbb{E}_{h_{i2},\kappa_{i2}}[V_2(R_{i1}(\tau),h_{i2},\kappa_{i2})|h_{i1}=0] \\ &+\Pr(h_{i1}=1)\cdot\mathbb{E}_{\kappa_{i1}}[V_1(R_{i1}(\tau),1,\kappa_{i1})]\} \\ &-\sum_{t=1}^{T(\tau)}\delta^{t-1}\cdot premium(\tau). \end{aligned} \tag{7}$$

$R_{i1}(\tau)$ is the cost-sharing limit associated with contract τ. $T(\tau)$ denotes the length of the contract in months. With probability $(1-\Pr(h_{i1}=1))$ a patient has no healthcare need. Then, she transitions directly to the next period so that $\delta\cdot\mathbb{E}_{h_{i2},\kappa_{i2}}[V_2(R_{i1}(\tau),h_{i2},\kappa_{i2})|h_{i1}=0]$ is the *ex ante* value function under policy τ when patients do not have a healthcare need. With probability $\Pr(h_{i1}=1)$, she has a healthcare need. Then, $\mathbb{E}_{\kappa_{i1}}[V_1(R_{i1}(\tau),1,\kappa_{i1})]$ is the *ex ante* value function under policy τ.[29] In our model, utility is quasi-linear in money. Therefore, $W(\tau)$ is expressed in terms of euros and takes into account the discounted sum of premiums the patient pays.

We use welfare when care is free as a baseline, which we denote by $\tau=0$. For each other policy option with a contract length of 12 months (all but the 2 year deductible), Table 7 reports the difference between $W(\tau)$ and $W(0)$. For the 2 year deductible, we first calculate the difference between $W(\tau)$,which is the discounted sum over 24 months, and $W(0)+\delta^{12}\cdot W(0)$, which is the value of two one year contracts without cost sharing. Then, we annuitize this difference by dividing by $1+\delta^{12}$. In that way, all welfare measures are comparable.

The first column of Table 7 reports overall effects. We find that cost-sharing generally increases welfare.[30] The reason is that individuals value the premium reductions more than the care they would have otherwise consumed. We find that welfare increases in the size of the

[29] See (9) in Online Appendix B.1.3. As before, we keep the dependence on risk score quartiles implicit.

[30] The exception is the €150 deductible. The reason is that out-of-pocket payments under a €150 deductible happen relatively early in the year, while the premium reduction is uniform over time. Patients discount the future and therefore prefer no cost sharing over a €150 deductible. This timing effect is also present for most of the other contracts and works against our general finding that welfare is higher under cost-sharing.

Table 7: Welfare in full sample

		risk score quartile			
policy option	all	Q1	Q2	Q3	Q4
€150 deductible	-45.94	-53.47	-63.91	-57.39	-8.99
€350 deductible (status quo)	33.57	20.00	6.75	10.27	97.31
€500 deductible	154.43	113.58	111.69	149.71	241.63
Two-year €700 deductible	78.64	84.47	75.26	36.76	119.02
Donut hole from €350 to €700	87.27	93.04	92.92	81.26	82.14
75% coinsurance with €350 maximum	11.08	9.47	-3.07	-14.00	52.34
Co-payment €30	6.09	31.75	22.00	8.62	-37.30
Co-payment €50	26.12	58.56	43.31	32.02	-28.57
€350 deductible with monthly €150 cap	17.18	20.32	4.98	-13.35	57.42

Notes: The table reports the difference in welfare under a number of policies (rows) and welfare under no cost sharing. See notes to Table 3 for details on the policies. The table reports averages for the full sample. We obtain those averages by combining results for counterfactual simulations for our estimation sample and results for the 10% top spenders. For the top spenders we calculate the welfare difference as the negative of the sum of the difference in out-of-pocket payments and the difference in the premium. Table A.5 in the Online Appendix reports results for our estimation sample. See Section 7.4 for further details.

deductible. It is highest for the €500 deductible contract.

More generally, Figures 3 (c) and (d) show that policies that lead to lower premiums and lower spending by individuals in the fourth risk score quartile tend to lead to higher welfare of individuals in the highest risk score group.

7.5 Robustness

Our main finding is that policies that lower spending do not necessarily led to less redistribution from low-risk to high-risk individuals; and even if they do, policies that lead to lower spending generally lead to higher welfare for all risk score groups. In this last subsection, we discuss to what extent this finding may be altered when we broaden our evaluation criteria and also take into account that out-of-pocket risk differs across contracts, that patients may face liquidity constraints, and that unmet needs may be more harmful to patients than they believe when they decide not to visit a doctor.

7.5.1 Out-of-pocket risk

Recall that in our model, utility is quasi-linear in money. This implies that patients are risk-neutral. If they are in fact risk-averse, then our measure of welfare could be too high for policy options that lead to higher out-of-pocket risk.

Table 3 shows the standard deviation of average monthly out-of-pocket expenditures for each policy. It is 0 when care is free, €9.88 for a €350 deductible (the status quo), and biggest at €12.62 for a €500 deductible. We follow Handel et al. (2024), who also assess the effect of out-of-pocket risk on patient welfare in the Netherlands, to translate this into a risk premium.[31] They assume constant absolute risk aversion (CARA) preferences. A standard value of absolute risk aversion is 10^{-5}. For this value and for individuals in our estimation sample, the risk premium for the €350 deductible contract is equal to $0.5 \cdot 10^{-5} \cdot (12 \cdot 9.58)^2 = 0.07$ euros per year.[32] For the €500 deductible contract, the risk premium increases to $0.5 \cdot 10^{-5} \cdot (12 \cdot 13.16)^2 =$

[31] We do not introduce risk aversion into our model, but directly calculate the risk premium associated with the risk patients face.

[32] We multiply by 12 because the table reports the standard deviation of yearly out-of-pocket expenditures divided by 12. Therefore, the risk premium we calculate here is yearly.

0.11 euros. For extreme risk aversion of 10^{-3} it increases from €7.03 per year to €11.47 per year. Table 4 implies that yearly premiums decrease from €2223.00 to €1900.80 when the deductible is increased from €350 to €500. Table 7 shows that yearly welfare increases from €33.57 to €154.43. This suggests that the effects on out-of-pocket risk are unlikely to affect our conclusions.

7.5.2 Liquidity constraints

One may also be concerned that patients may face liquidity constraints when they face higher levels of cost-sharing. For instance, when the yearly deductible is increased by €150, then it could be that a patient is asked to pay €150 more at some point in time during the year. However, in the Netherlands, such a payment would usually come with a delay of several months, and patients have the possibility to pay it in installments (see Hayen et al., 2021, for details). In addition, an increase in the deductible by €150 (from €350 to €500) would lead to premium reductions of $12 \cdot (185.25 - 158.40) = 322.16$ euros per year (based on Table 4). This means that liquidity constraints are actually relaxed. Therefore, overall, we believe that liquidity constraints are unlikely to change our conclusions.

7.5.3 Unmet needs and behavioral hazard

A final cause of concern could be that as a response to cost-sharing, patients harm themselves by deciding to not visit a doctor when they have a healthcare need. For instance, Chandra et al. (2021) show that not taking certain drugs as a response to cost-sharing can increase mortality.

One advantage of estimating a structural model with a latent healthcare needs process is that it leads to estimates of the probability of not visiting a doctor when patients have a healthcare need, both for the policy that is in place and for counterfactual policies. Table A.2 shows that the likelihood of not visiting a doctor when patients have a healthcare need increases from 0.14 to 0.34 when the deductible increases from €350 to €500.

Our model attributes this effect to two sources: a non-medical cost of visiting a doctor and moral hazard. Table A.2 shows that even when care is free, the likelihood of not visiting a doctor with a healthcare need is 0.02. This is driven by the non-medical cost of visiting a

doctor. Table A.2 also shows that the likelihood to visit a doctor varies across policies. This variation is driven by moral hazard. Patients find it more attractive to visit a doctor when they do not have to pay for care themselves, because this allows them to exert moral hazard, which they value. Therefore, they are more likely to visit a doctor when there is a lower level of cost sharing.

In our model, individuals take the expected negative utility from unmet needs into account when they decide whether or not to visit a doctor. However, it is possible that individuals underestimate the value of certain medical treatments. Baicker et al. (2015) refer to such a divergence between the private and the social valuation of care as behavioral hazard. In our paper, we do not explicitly measure the negative health consequences of unmet needs. Therefore, we cannot directly answer the question how costly unmet needs are to society and whether accounting for behavioral hazard would overturn our conclusions.

However, it is interesting to observe that increasing the deductible from €350 to €500 would lead to a premium reduction of about €322.16 per year (see Section 7.5.2) and an increase in welfare of about €120.86 per year (see Section 7.4). This suggests that an increase in the size of the deductible may be desirable, as long as the associated additional costs related to unmet needs that the individual has not taken into account are smaller than about €120 per year and patient. It is useful to compare these €120 to the total healthcare costs under the current policy, which are €2454.10 per year (12 times €204.50 in Table 4). This means that our conclusions would only be overturned if the negative consequences of behavioral hazard would be more than about 5% of current healthcare costs when the deductible is increased from €350 to €500. While in principle possible, this seems to be unlikely.

8 Conclusion

This paper studies whether more cost-sharing leads to less redistribution from individuals with low health risks to individuals with high health risks and lower welfare of high-risk individuals. We develop and estimate a model of healthcare utilization using data from a large Dutch insurer. Individuals make decisions along both the extensive and intensive margins, and the model ex-

plicitly accounts for within-year spending dynamics that arise from non-linear prices common in health insurance contracts.

Our estimates allow us to study the impact of various counterfactual insurance contracts on a variety of outcomes, such as spending and insurance premiums. Our general finding is that among the policies we study there is no strong dependence between redistribution and the amount of cost-sharing. At the same time, more cost sharing generally leads to higher welfare.

We find that for all risk score groups, welfare is highest for the €500 deductible contract (the highest deductible amount we consider). Thus, even high-risk individuals would benefit from somewhat higher levels of cost-sharing. This is the case despite the fact that this particular policy leads to less redistribution from low-risk to high-risk individuals relative to the €350 deductible that was actually in place. The reason for this is a version of the tragedy of the commons. Patients consume more care when they do not have to pay for it individually by means of out-of-pocket payments. However, if everyone does this, then everyone pays in the end because the premium will increase. The additional care that everyone consumes because of this has low value to individuals.

References

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

Aron-Dine, A., L. Einav, and A. Finkelstein (2013). The RAND health insurance experiment, three decades later. *Journal of Economic Perspectives 27*(1), 197–222.

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

Baicker, K., S. Mullainathan, and J. Schwartzstein (2015). Behavioral hazard in health insurance. *Quarterly Journal of Economics 130*(4), 1623–1667.

Brot-Goldberg, Z. C., A. Chandra, B. R. Handel, and J. T. Kolstad (2017). What does a deductible do? The impact of cost-sharing on health care prices, quantities, and spending dynamics. *Quarterly Journal of Economics 132*(3), 1261–1318.

Cai, Y. and K. L. Judd (2012). Dynamic programming with shape-preserving rational spline hermite interpolation. *Economics Letters 117*(1), 161–164.

Chandra, A., E. Flack, and Z. Obermeyer (2021). The health costs of cost-sharing. NBER Working Paper No. 28439.

Chandra, A., J. Gruber, and R. McKnight (2010). Patient cost-sharing and hospitalization offsets in the elderly. *American Economic Review 100*(1), 193–213.

CPB (2021). Keuzes in kaart 2022-2025: Economische analyse van verkiezingsprogrammaâs. CPB Netherlands Bureau for Economic Policy Analysis, The Hague, Netherlands.

Cutler, D. and R. Zeckhauser (2000). The anatomy of health insurance. In J. Newhouse and A. Culyer (Eds.), *Handbook of Health Economics*, Volume 1, Chapter 11, pp. 563 – 643. Elsevier Science.

Dalton, C. M., G. Gowrisankaran, and R. J. Town (2020). Salience, Myopia, and Complex Dynamic Incentives: Evidence from Medicare Part D. *Review of Economic Studies 78*(2), 822–869.

Ecorys (2011). Evaluatie naar het verplicht eigen risico. Ecorys, Rotterdam, The Netherlands.

Einav, L. and A. Finkelstein (2018). Moral hazard in health insurance: What we know and how

we know it. *Journal of the European Economic Association 16*, 957 – 982.

Einav, L., A. Finkelstein, S. P. Ryan, P. Schrimpf, and M. R. Cullen (2013). Selection on moral hazard in health insurance. *American Economic Review 103*(1), 178–219.

Einav, L., A. Finkelstein, and P. Schrimpf (2015). The response of drug expenditure to nonlinear contract design: Evidence from Medicare Part D. *Quarterly Journal of Economics 130*(2), 841–899.

Ellis, R. P. (1986). Rational behavior in the presence of coverage ceilings and deductibles. *RAND Journal of Economics 17*(2), 158–175.

French, E. and J. B. Jones (2004). On the distribution and dynamics of health care costs. *Journal of Applied Econometrics 19*(6), 705–721.

Guo, A. and J. Zhang (2019). What to expect when you are expecting: Are health care consumers forward-looking? *Journal of Health Economics 67*, 102216.

Handel, B., J. Kolstad, T. Minten, and J. Spinnewijn (2024). The socio-economic distribution of choice quality: Evidence from health insurance in the netherlands. *American Economic Review: Insights*.

Hayen, A. P., T. J. Klein, and M. Salm (2021). Does the framing of patient cost-sharing incentives matter? the effects of deductibles vs. no-claim refunds. *Journal of Health Economics 80*, 102520.

Hayen, A. P., M. J. van den Berg, B. R. Meijboom, J. N. Struijs, and G. P. Westert (2015). Incorporating shared savings programs into primary care: From theory to practice. *BMC Health Services Research 15*(580), 1–15.

Ho, K. and R. S. Lee (forthcoming). Health insurance menu design for large employers. *RAND Journal of Economics*.

Jones, A. M., N. Rice, T. B. d'Uva, and S. Balia (2013). *Applied Health Economics*. Routledge.

Keeler, E. B., J. P. Newhouse, and C. E. Phelps (1977). Deductibles and the demand for medical care services: The theory of a consumer facing a variable price schedule under uncertainty. *Econometrica 45*(3), 641–655.

Keeler, E. B. and J. E. Rolph (1988). The demand for episodes of treatment in the health

insurance experiment. *Journal of Health Economics 7*(4), 337–367.

Klein, T. J., M. Salm, and S. Upadhyay (2022). Patients' response to dynamic incentives: Evidence from a differences-in-regression discontinuities design. *Journal of Public Economics 210*, 104660.

Kowalski, A. 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.

Manning, W., J. Newhouse, N. Duan, E. Keeler, A. Leibowitz, and M. Marquis (1987). Health insurance and the demand for medical care: Evidence from a randomized experiment. *American Economic Review 77*(3), 251–277.

McGuire, T. (2011). Demand for health insurance. In T. M. M.V. Pauly and P. Barros (Eds.), *Handbook of Health Economics*, Volume 2, Chapter 5, pp. 317 – 396. Amsterdam: Elsevier Science.

Pauly, M. V. (1968). The economics of moral hazard: Comment. *American Economic Review 58*(3), 531–537.

Shigeoka, H. (2014). The effect of patient cost sharing on utilization, health, and risk protection. *American Economic Review 104*(7), 2152–2184.

van Kleef, R., W. van de Ven, and R. van Vliet (2009). Shifted deductibles for high risks: More effective in reducing moral hazard than traditional deductibles. *Journal of Health Economics 28*(1), 198–209.

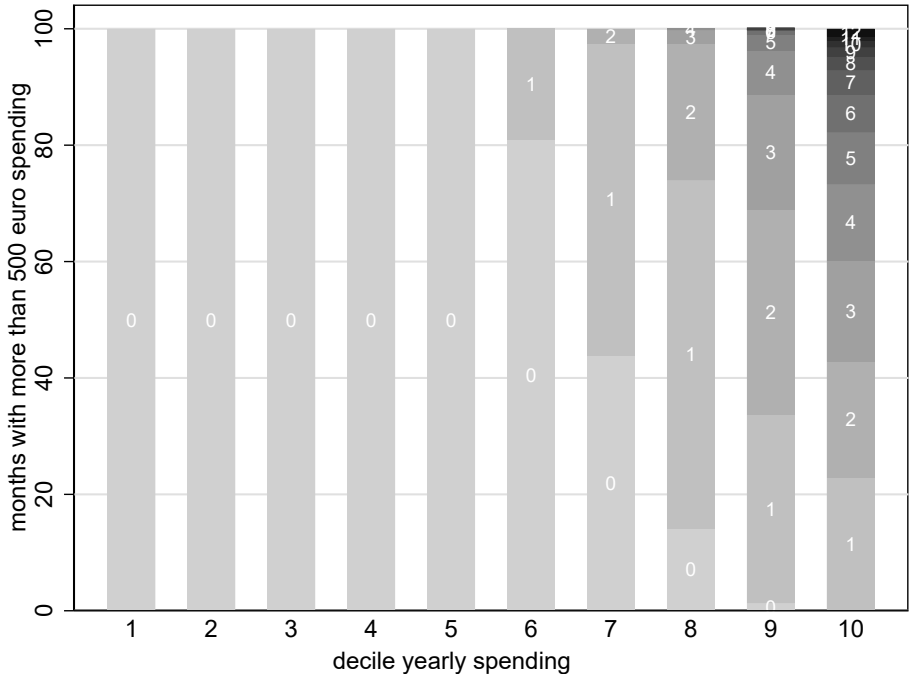
Zweifel, P. and W. G. Manning (2000). Moral hazard and consumer incentives in health care. *Handbook of Health Economics 1*, 409–459.

Online Appendix

This Online Appendix contains additional tables and figures, as well as technical details related to numerically solving the model and estimating the parameters.

A Additional tables and figures

Figure A.1: Months with more than €500 spending by decile yearly spending



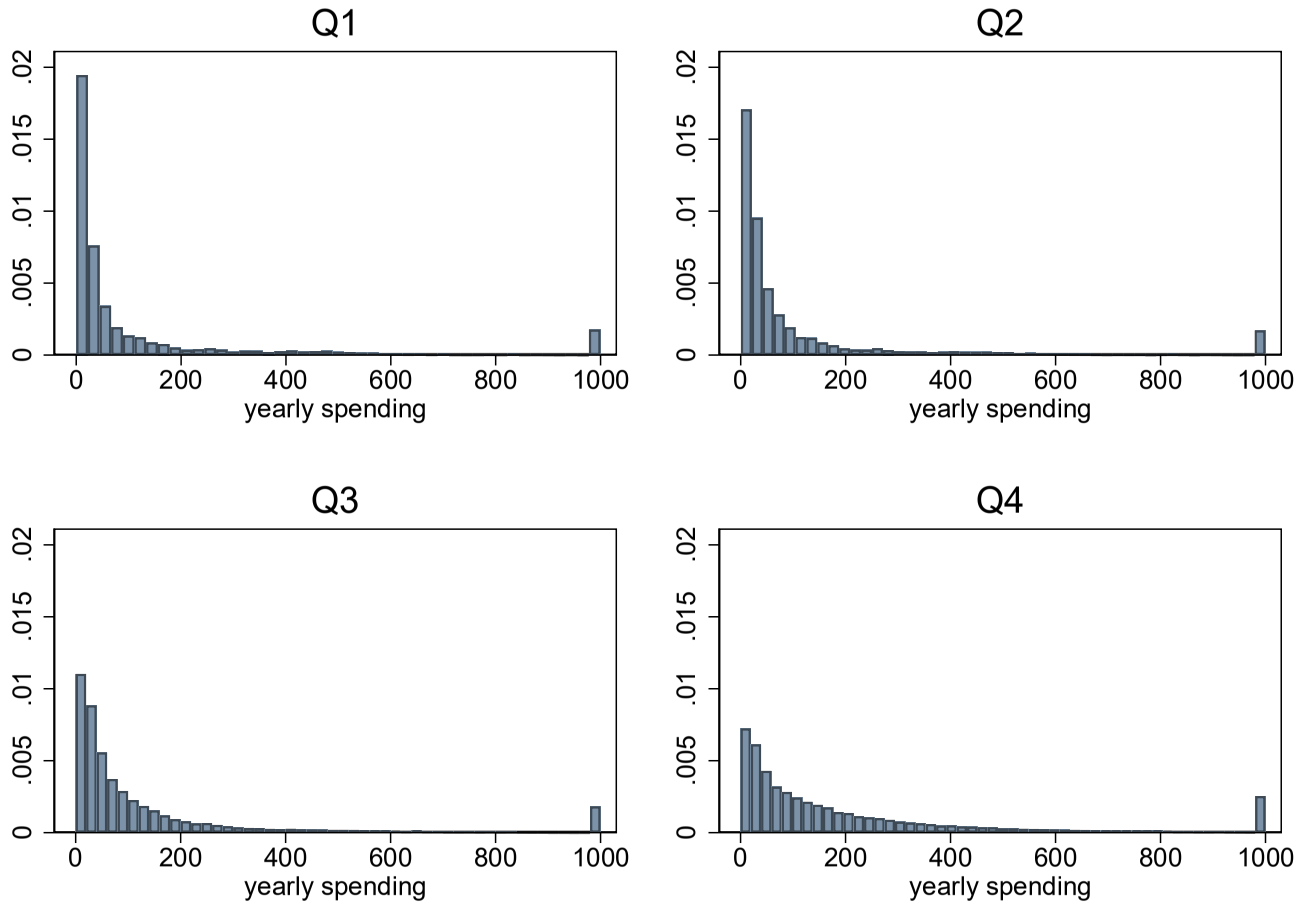
This figure shows the distribution of the number of months with spending above €500 by risk score decile. One observation per person. Reported for the full sample. The estimation sample excludes persons in the 10th decile for yearly spending.

Table A.1: Summary statistics for 10% highest spenders

	excluded from estimation sample				
	all	risk Q1	risk Q2	risk Q3	risk Q4
demographics (one observation per person)					
average age	57.86	41.20	50.61	60.23	59.32
share female	.53	.52	.52	.49	.55
average income in neighborhood	2009	2439	2159	2048	1950
care consumption (one observation per month and person)					
average spending	1447.98	1202.91	1339.76	1290.27	1520.25
probability any spending	.795	.441	.552	.732	.867
average spending if any spending	1821.92	2726.37	2424.94	1763.42	1753.49
median spending if any spending	257.66	193.54	156.86	163.04	286.62
95th percentile spending if any spending	9092.61	11931.24	11858.91	9471.72	8397.42
standard deviation spending if any spending	5152.59	6380.55	6422.43	4872.77	5039.72
number individuals	8649	404	770	1556	5919

Notes: This table reports summary statistics for the 10% highest spenders in 2013. See Table 1 for variable definitions and statistics for the full sample and the estimation sample. The 10% highest spender are excluded from the estimation sample. See Section 3 for details.

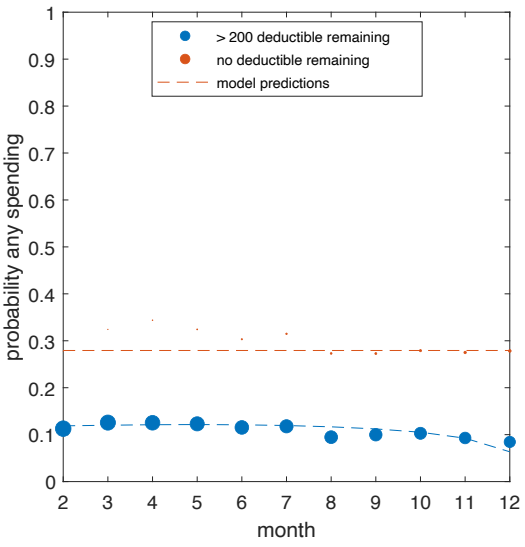
Figure A.2: Spending if positive



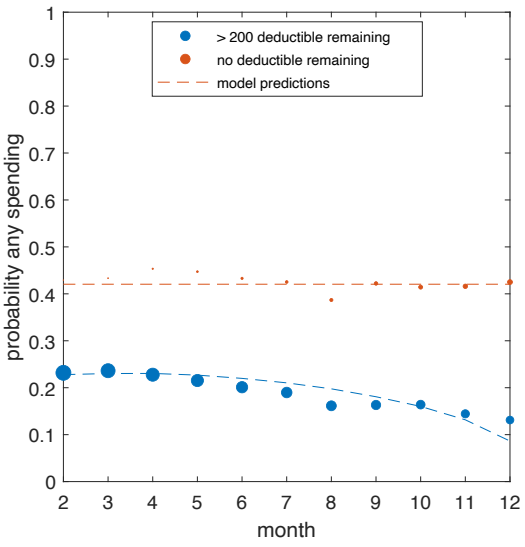
This figure shows the distribution of monthly spending by risk score quartile. One observation is a person-month when spending is positive. To produce this figure, we re-coded spending above €10000 as €10000. Reported for the estimation sample.

Figure A.3: Probability of any spending

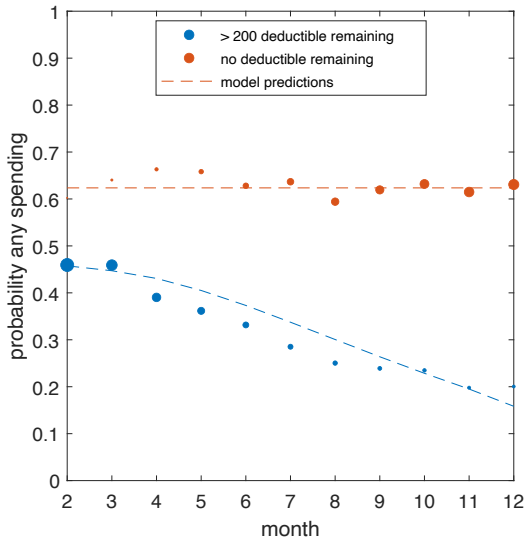
(a) Risk score quartile 1



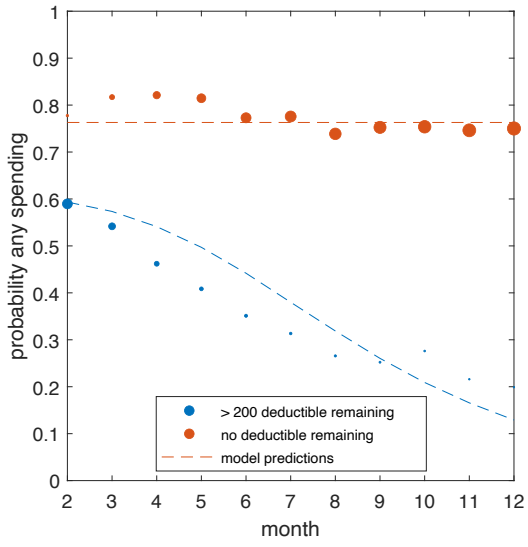
(b) Risk score quartile 2



(c) Risk score quartile 3



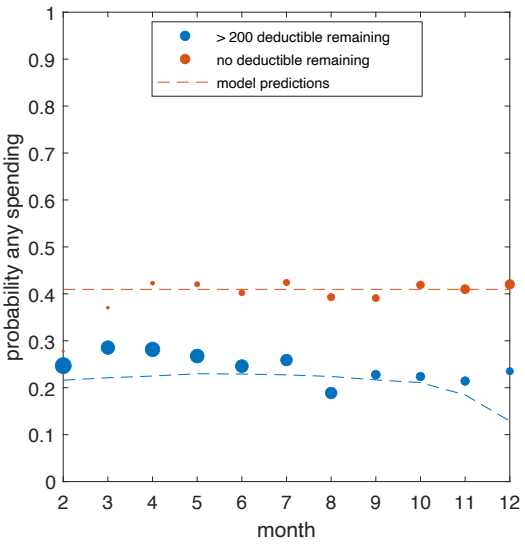
(d) Risk score quartile 4



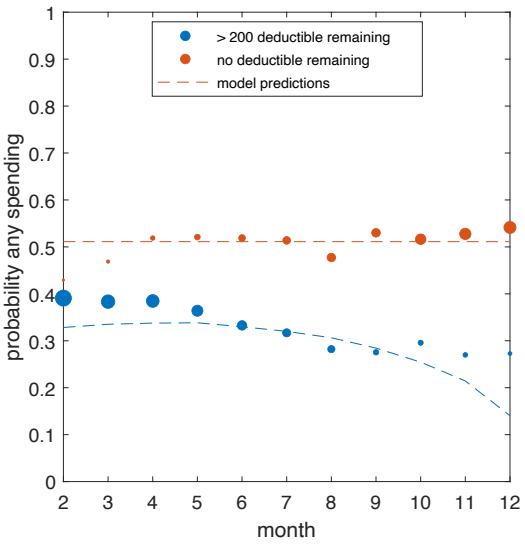
Notes: The figures plot mean probability of any spending on months across all risk score quartiles and remaining deductible groups. Each solid dot is the probability of spending at time t computed from the data, while the dashed-line denotes the predictions generated by our model.

Figure A.4: Probability of any spending if spending in $t-1$

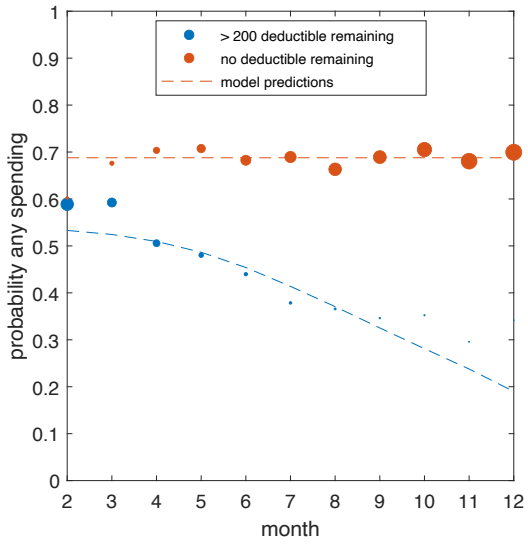
(a) Risk score quartile 1



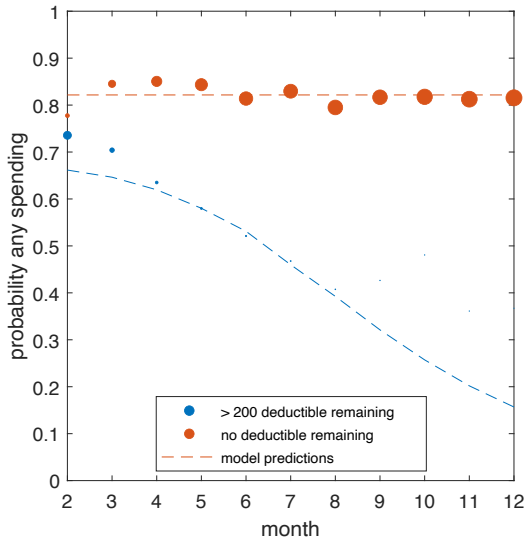
(b) Risk score quartile 2



(c) Risk score quartile 3



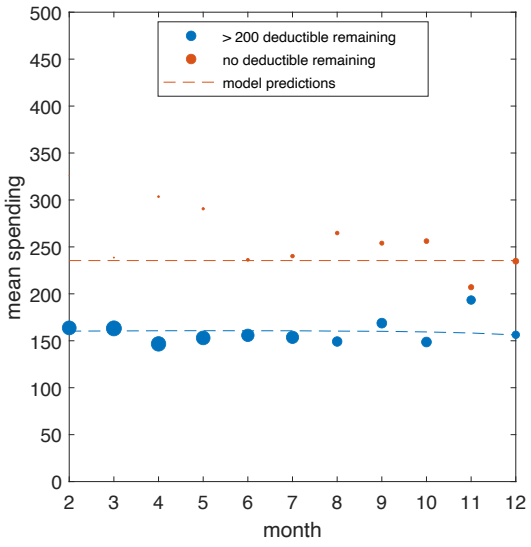
(d) Risk score quartile 4



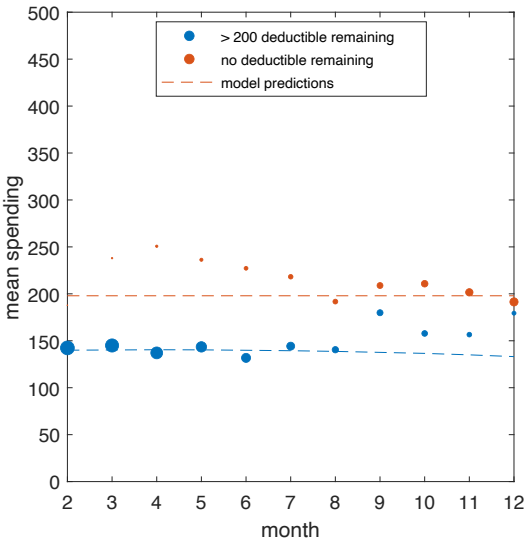
Notes: The figures plot mean probability of spending conditional on spending in the previous period against months across all risk score quartiles and remaining deductible groups. Each solid dot is the probability of any spending conditional on spending at time $t-1$ computed from the data, while the dashed-line denotes the predictions generated by our model.

Figure A.5: Intensive margin fit

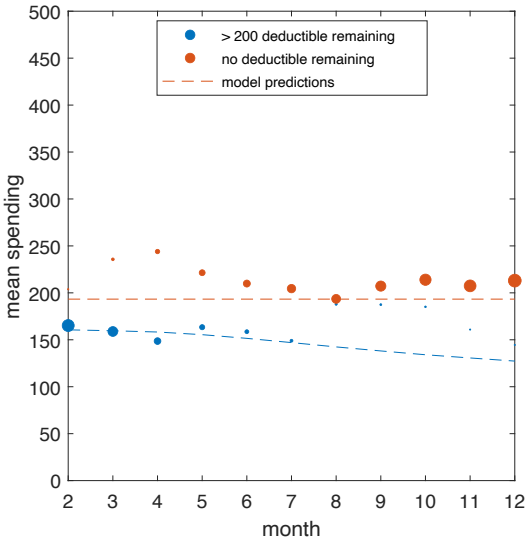
(a) Risk score quartile 1



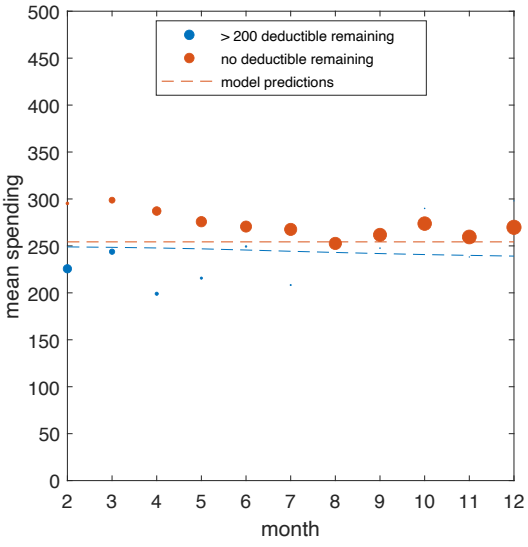
(b) Risk score quartile 2



(c) Risk score quartile 3



(d) Risk score quartile 4



Notes: The figures plot mean spending conditional on any spending on months across all risk score quartiles and remaining deductible groups. Each solid dot is mean spending conditional on any spending at time t computed from the data, while the dashed-line denotes the predictions generated by our model.

Table A.2: Probability of not going with a healthcare need in estimation sample

		risk score quartile			
policy option	all	1	2	3	4
no cost sharing	0.02	0.00	0.00	0.01	0.09
€150 deductible	0.03	0.00	0.00	0.02	0.10
€350 deductible (status quo)	0.14	0.14	0.16	0.10	0.19
€500 deductible	0.34	0.25	0.36	0.34	0.45
Two-year €700 deductible	0.32	0.27	0.41	0.32	0.28
Donut hole from €350 to €700	0.12	0.03	0.05	0.15	0.28
75% coinsurance with €350 maximum	0.08	0.03	0.03	0.09	0.20
Co-payment €30	0.06	0.00	0.00	0.05	0.24
Co-payment €50	0.11	0.00	0.00	0.11	0.39
€350 deductible with monthly €150 cap	0.08	0.02	0.03	0.10	0.20

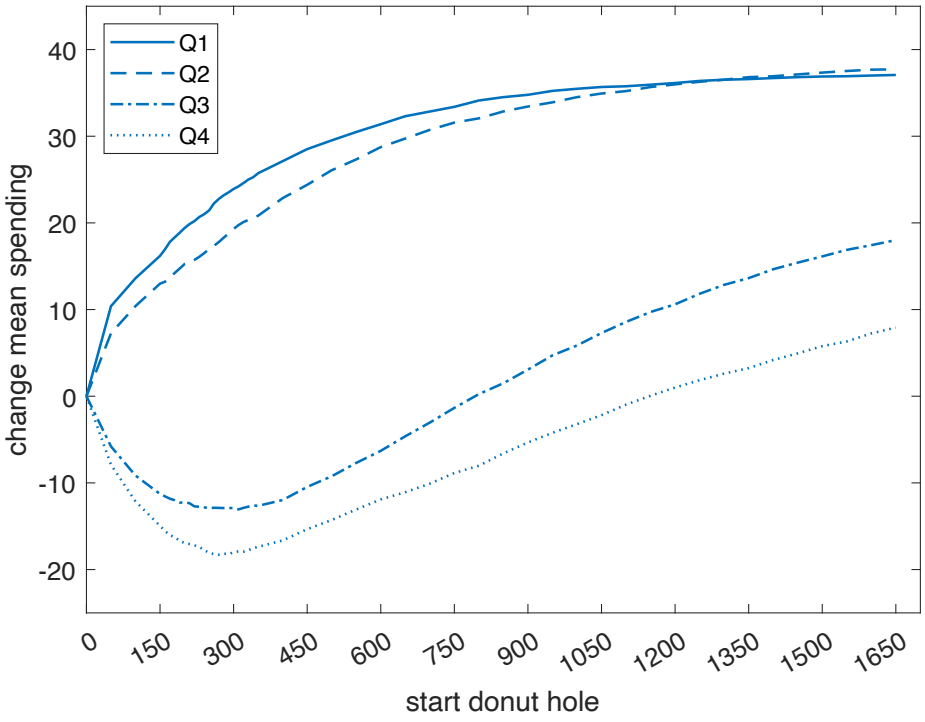
Notes: The table reports the probability to not go to the doctor despite having a healthcare need under a number of policies (rows). Reported for the estimation sample.

Table A.3: Spending by risk score in estimation sample

	risk score quartile			
policy option	Q1	Q2	Q3	Q4
no cost sharing	57.38	75.31	118.66	149.68
€150 deductible	50.28	70.15	116.66	147.42
€350 deductible (status quo)	20.74	36.31	94.64	129.13
€500 deductible	3.64	7.27	48.11	77.35
Two-year €700 deductible	0.85	2.27	53.59	110.63
Donut hole from €350 to €700	45.00	57.35	82.32	110.38
75% coinsurance with €350 maximum	40.31	57.00	96.07	126.68
Co-payment €30	57.38	75.30	112.06	121.85
Co-payment €50	57.37	75.21	101.12	92.37
€350 deductible with monthly €150 cap	38.57	53.19	94.02	126.66

Notes: The table reports spending under a number of policies (rows). Reported for the estimation sample. See notes to Table 5 for details.

Figure A.6: Spending changes across different donut hole contracts



Notes: The figure depicts changes in mean spending for different donut hole starting points. Reported by risk score quartile. The change is always relative to spending under a €350 deductible. If the starting point is 0, we have a deductible contract; if the starting point is 350, we have a contract where the first €350 is free, the next €350 has to be paid for and then care is free again. Reported for estimation sample.

Table A.4: Redistribution in estimation sample

	risk score quartile			
policy option	Q1	Q2	Q3	Q4
no cost sharing	-174.75	-156.82	-113.46	-82.45
€150 deductible	-176.45	-158.30	-112.30	-81.49
€350 deductible (status quo)	-173.03	-163.23	-116.30	-82.20
€500 deductible	-156.78	-155.01	-130.07	-104.34
Two-year €700 deductible	-165.99	-165.29	-130.73	-80.30
Donut hole from €350 to €700	-156.85	-148.19	-128.73	-101.79
75% coinsurance with €350 maximum	-166.92	-155.47	-121.49	-90.27
Co-payment €30	-160.81	-147.17	-115.52	-106.57
Co-payment €50	-150.02	-139.30	-119.09	-124.50
€350 deductible with monthly €150 cap	-165.78	-156.64	-122.19	-89.76

Notes: The table reports redistribution under a number of policies (rows). Reported for the estimation sample. See notes to Table 6 for details.

Table A.5: Welfare in estimation sample

		risk score quartile			
policy option	all	Q1	Q2	Q3	Q4
€150 deductible	-48.28	-52.72	-63.28	-59.13	-8.44
€350 deductible (status quo)	34.55	23.81	12.29	11.36	108.34
€500 deductible	161.65	120.19	122.26	157.74	273.31
Two-year €700 deductible	67.81	88.15	80.26	29.27	74.84
Donut hole from €350 to €700	102.91	99.00	103.22	93.22	120.26
75% coinsurance with €350 maximum	15.18	13.53	3.13	-11.20	67.68
Co-payment €30	26.99	37.79	32.45	24.12	9.36
Co-payment €50	61.39	68.62	60.73	58.35	56.77
€350 deductible with monthly €150 cap	21.21	24.53	11.34	-10.98	71.87

Notes: The table reports welfare under a number of policies (rows). Reported for the estimation sample. See notes to Table 7 for details.

Table A.6: OOP by risk score quartile in estimation sample

	risk score quartile			
policy option	Q1	Q2	Q3	Q4
no cost sharing	0.00	0.00	0.00	0.00
€150 deductible	10.25	11.97	12.49	12.44
€350 deductible (status quo)	8.52	14.29	25.69	26.09
€500 deductible	2.02	3.88	19.78	23.29
Two-year €700 deductible	0.52	1.23	18.00	24.61
Donut hole from €350 to €700	6.81	10.49	16.01	17.13
75% coinsurance with €350 maximum	15.65	20.89	25.99	25.37
Co-payment €30	8.32	12.61	17.72	18.55
Co-payment €50	13.86	20.98	26.69	23.34
€350 deductible with monthly €150 cap	13.60	19.09	25.46	25.68

Notes: The table reports the standard deviation of yearly out-of-pocket spending divided by 12 under a number of policies (rows). See notes to Table 3 for details on the policies. Reported for the estimation sample.

Table A.7: Std. OOP by risk score quartile in estimation sample

	risk score quartile			
policy option	Q1	Q2	Q3	Q4
no cost sharing	0.00	0.00	0.00	0.00
€150 deductible	4.06	2.01	0.34	0.69
€350 deductible (status quo)	11.17	12.26	8.02	7.46
€500 deductible	7.05	9.80	17.17	17.75
Two-year €700 deductible	2.95	4.52	11.85	9.17
Donut hole from €350 to €700	11.09	12.53	13.02	13.25
75% coinsurance with €350 maximum	10.78	9.54	7.06	7.79
Co-payment €30	4.56	4.91	4.94	4.92
Co-payment €50	7.60	8.19	8.15	7.89
€350 deductible with monthly €150 cap	10.09	9.86	7.77	7.38

Notes: The table reports average monthly out-of-pocket spending under a number of policies (rows). See notes to Table 3 for details on the policies. Reported for the estimation sample.

B Technical Details

This appendix provides technical details that are related to numerically solving the model and estimating the parameters.

B.1 Solving the model

B.1.1 Overview

Recall that at the start of each month $t = 1, \ldots, T$, patients know their remaining deductible, R_{it}, whether they have a healthcare need, h_{it}, and the non-medical cost to visiting a doctor, κ_{it}. Denote the associated value function by $V_t(R_{it}, h_{it}, \kappa_{it})$ and define the terminal condition $V_{T+1}(\cdot, \cdot, \cdot) = 0$. We solve the patients' decision problem by backward recursion, starting with the last month, T. It is useful to keep in mind that from the perspective of month t, we can treat $V_{t+1}(R_{it+1}, h_{it+1}, \kappa_{it+1})$ as known.

B.1.2 Solving model on a grid

Our model features choice and state variables that are continuous. We solve the model on a grid for combinations of R_{it} and λ_{it}. For this, we use a grid for candidate policy combined with interpolation to find the optimal policy. Below we provide details in steps 3 and 4 for the last period and steps 6 and 7 for earlier periods.

The grids we use are the following:

1. c_{it}: 0 to 1000 with increments of 5 followed by 5000 and 5,000,000.
2. R_{it}: 0 to 350 with increments of 5.
3. λ_{it}: the same grid as c_{it}.

B.1.3 Expected maximal utility

Recall that patients visit a doctor if $V_t^{go}(R_{it}) - \kappa_{it} \geq V_t^{ngo}(R_{it})$ and that κ_{it} is distributed logistic with location parameter $\bar{\kappa}$ and scale parameter s_κ.

To derive an expression for the value function when having a healthcare need, $\mathbb{E}_{\kappa}[V_t(R_{it},1,\kappa_{it})]$, one can re-interpret the cost as

$$\kappa_{it} = \bar{\kappa} + s_{\kappa} \cdot (\varepsilon_{i1t} - \varepsilon_{i0t}),$$

where ε_{i0t} and ε_{i1t} are taste shocks that are distributed according to re-centered type 1 extreme value distributions with location parameter equal to minus Euler's constant and scale parameter equal to 1. The taste shocks ε_{i0t} and ε_{i1t} are related to not visiting a doctor and visiting a doctor, respectively. Therefore, the expected (over κ_{it}) maximal utility conditional on having a healthcare need is[1]

$$\mathbb{E}_{\kappa}[V_t(R_{it},1,\kappa_{it})] = s_{\kappa} \ln\left(\exp\left(\frac{V_t^{go}(R_{it}) - \bar{\kappa}}{s_{\kappa}}\right) + \exp\left(\frac{V_t^{ngo}(R_{it})}{s_{\kappa}}\right)\right). \tag{8}$$

An expression for the value of not visiting a doctor, $V^{ngo}(R_{it})$, is given in Section 4.5. It can be re-written as

$$V_t^{ngo}(R_{it}) = \mathbb{E}_{\lambda_{it}}\left[u(0,\lambda_{it},R_{it})\right] + \delta \cdot \mathbb{E}_{h_{it+1},\kappa_{it+1}}\left[V_{t+1}(R_{it},h_{it+1},\kappa_{it+1})|h_{it}=1\right].$$

The first part is the expected flow utility of not visiting a doctor. The second part is the continuation value. It follows from(1) that this is a constant and enters the patient's decision in an additive way. Therefore, instead of estimating $\bar{\kappa}$, we estimate $\tilde{\kappa} \equiv \mathbb{E}_{\lambda_{it}}\left[u(0,\lambda_{it},R_{it})\right] + \bar{\kappa}$. In words, $\tilde{\kappa}$ captures the sum of the mean of the fixed costs of visiting a doctor and the flow utility from not visiting a doctor; a patient visits a doctor if $V_t^{go}(R_{it}) - \kappa_{it} \geq V_t^{ngo}(R_{it})$ or $V_t^{go}(R_{it}) \geq \tilde{\kappa} + \delta \cdot \mathbb{E}_{h_{it+1},\kappa_{it+1}}\left[V_{t+1}(R_{it},h_{it+1},\kappa_{it+1})|h_{it}=1\right]$. The expected maximal utility is

$$\begin{aligned} &\mathbb{E}_{\kappa_{it}}[V_t(R_{it},1,\kappa_{it})] \\ &= \bar{\kappa} + s_{\kappa} \ln\left(\exp\left(\frac{V_t^{go}(R_{it})}{s_{\kappa}}\right) + \exp\left(\frac{\tilde{\kappa} + \delta \cdot \mathbb{E}_{h_{it+1},\kappa_{it+1}}\left[V_{t+1}(R_{it},h_{it+1},\kappa_{it+1})|h_{it}=1\right]}{s_{\kappa}}\right)\right). \end{aligned} \tag{9}$$

In principle, we could estimate $\bar{\kappa}$, because it follows from (2) and the log-normality of λ_{it}

[1]The advantage of thinking of ε_{i0t} and ε_{i1t} as distributed according to the re-centered type 1 extreme value distribution is that they are mean 0, respectively.

that

$$\mathbb{E}_{\lambda_{it}}\left[u\left(0,\lambda_{it},R_{it}\right)\right]=\mathbb{E}\left[-\lambda_{it}-\frac{1}{2\omega}\cdot\lambda_{it}^{2}\right]=-\exp\left(\mu+\frac{\sigma^{2}}{2}\right)-\frac{1}{2\omega}\cdot\exp\left(2\mu+2\sigma^{2}\right).$$

However, this is not necessary. The reason is that patient decisions and welfare differences (between counterfactuals) only depend on $\tilde{\kappa}$.

B.1.4 Approximating expectations

We approximate all of the expected values through discretization. For example, to approximate the expectation of $\bar{V}_t^{go}(R_{it},\lambda_{it})$ over λ_{it} we use

$$\int_0^{\infty}\bar{V}_t^{go}\left(R_{it},\lambda\right)f\left(\lambda\right)d\lambda\approx\sum_{j=1}^{G}\bar{V}_t^{go}\left(R_{it},\lambda^{j}\right)p_j, \tag{10}$$

where G denotes the total number of intervals used, λ^j is the mid-point of interval j, and p_j denotes the probability mass for the specific interval, obtained by subtracting the log-normal cumulative distribution function at the upper end of interval j from the log-normal cumulative distribution function at the lower end of interval j. The intervals in G were determined by using the grid for λ_{it} described in Section B.1.2. The upper end of each grid point was determined by adding half of the distance between consecutive grid points, while for the lower end, we subtracted this amount.

B.1.5 Backward recursion

We first solve the model in the final period. This means doing the following:

1. For each $\{c_{iT},R_{iT},\lambda_{iT}\}$ combination produced by our grids (see Section B.1.2), compute the respective flow utility, $u(c_{iT},\lambda_{iT},R_{iT})$.

2. Select, for each $\{R_{iT},\lambda_{iT}\}$, the point on the grid of c_{iT} that maximizes the flow utility. Call this the grid point solution.

3. Using flow utility computed at each grid point for c_{iT}, interpolate flow utility values not

on the exact grid of c_{iT} by fitting a cubic spline for each $\{R_{iT}, \lambda_{iT}\}$.[2]

4. Select, for each $\{R_{iT}, \lambda_{iT}\}$, the c_{iT} that maximizes the fitted flow utility. For this, consider only values close to the grid point solution.

5. Store the resulting computed flow utility as $V_T^{go}(R_{iT}, \lambda_{iT})$, for each possible combination of $\{R_{iT}, \lambda_{iT}\}$.

6. Calculate $V_T^{go}(R_{iT})$ by integrating over λ through discretization (see (10)).

7. Set the expected flow utility of not going to 0.[3] Since this is the last period, this is also $V_T^{ngo}(R_{iT})$.

8. Obtain $\mathbb{E}_{\kappa_i}[V_T(R_{iT}, 0, \kappa_{iT})]$, by evaluating the flow utility at $c_{iT} = 0$ when $\lambda_{iT} = 0$.

9. Compute $\mathbb{E}_{\kappa_i}[V_T(R_{iT}, 1, \kappa_{iT})]$ as described in Section B.1.3 above.

Then, moving on to the second to last period, we do the following:

1. For each $\{c_{iT-1}, R_{iT-1}, \lambda_{iT-1}\}$, compute the flow utility.

2. For each combination of $\{c_{iT-1}, R_{iT-1}\}$, compute R_{iT} (the remaining deductible in the final period), given by $\max(R_{iT-1} - c_{iT-1}, 0)$.

3. Compute $\mathbb{E}_{h_{iT}}[V_T(R_{iT}, h_{iT}, \kappa_{iT}) | h_{iT-1}]$ by taking a weighted average of $\mathbb{E}_{\kappa_i}[V_T(R_{iT}, 0, \kappa_{iT})]$ and $\mathbb{E}_{\kappa_i}[V_T(R_{iT}, 1, \kappa_{iT})]$, where the weights are given by either $1 - p^0$ and p^0 or $1 - p^1$ and p^1 depending on whether $h_{iT-1} = 0$ or $h_{iT-1} = 1$, respectively.

4. When $h_{iT-1} = 1$, obtain $\mathbb{E}_{h_{iT}}[V_T(R_{iT}, h_{iT}, \kappa_{iT}) | h_{iT-1} = 1]$ and add δ times this expectation to the flow utility computed in step 1. This yields the value function of visiting a doctor $\tilde{V}_{T-1}^{go}(c_{iT-1}, R_{iT-1}, \lambda_{iT-1})$ for each combination of $\{c_{iT-1}, R_{iT-1}, \lambda_{iT-1}\}$.[4]

[2] In general, cubic spline interpolation can introduce artificial non-monotonicities in the value function, leading to accuracy and convergence problems. See, e.g., Cai and Judd (2012). Shape-preserving splines may lead to more accurate results. We are less concerned about this in the current context because we use a very fine grid (Section B.1.2).

[3] See Section (B.1.3).

[4] $\tilde{V}_{T-1}^{go}(c_{iT-1}, R_{iT-1}, \lambda_{iT-1})$ is the value of going to the doctor when knowing λ_{iT-1} and for a given value of c_{iT-1}. This will be used to compute the value of going to the doctor when having a healthcare need, $V_{T-1}^{go}(R_{iT-1})$.

5. Select, for each $\{R_{iT-1}, \lambda_{iT-1}\}$, the point on the grid of c_{iT-1} that gives the highest value of $\tilde{V}^{go}_{T-1}(c_{iT-1}, R_{iT-1}, \lambda_{iT-1})$.

6. Using $\tilde{V}^{go}_{T-1}(c_{iT-1}, R_{iT-1}, \lambda_{iT-1})$ computed at each grid point for c_{iT-1}, interpolate values of $\tilde{V}^{go}_{T-1}(c_{iT-1}, R_{iT-1}, \lambda_{iT-1})$ not on the exact grid of c_{iT-1} by fitting a cubic spline for each $\{R_{iT-1}, \lambda_{iT-1}\}$.

7. In an area around the c_{iT-1} obtained in step 5, search for the maximum of the fitted value function. The resulting c_{iT-1} is the optimal policy for the given state variables.

8. Store the resulting value function, evaluated at the optimal policy c_{iT-1} obtained in the previous step, as $\bar{V}^{go}_{T-1}(R_{iT-1}, \lambda_{iT-1})$, for each possible combination of $\{R_{iT-1}, \lambda_{iT-1}\}$.[5]

9. Compute $V^{go}_{T-1}(R_{iT-1})$ by integrating over λ_{it}.

10. $V^{ngo}_{T-1}(R_{iT-1})$ is given by the value obtained in step 3, where $h_{iT-1} = 1$, multiplied by δ since the flow utility of not visiting a doctor is normalized to 0.[6]

11. Obtain $\mathbb{E}_{\kappa_i}[V_{T-1}(R_{iT}, 0, \kappa_{iT})]$ by evaluating the quantity in step 3, with $h_{iT-1} = 0$, and multiply this with δ.

12. Compute $\mathbb{E}_{\kappa_i}[V_T(R_{iT}, 1, \kappa_{iT})]$ as described in Section B.1.3 above.

Notice that, if we replace T with $T-1$ and $T-1$ with $T-2$ in the procedure outlined above, we solve the model for $T-2$. This recursive solution admits, in the end, an optimal policy function that describes how much healthcare services individuals would consume for every $\{R_{it}, \lambda_{it}, t\}$ combination when they visit a doctor.

After solving the model, it follows from (4) and the discussion in Section B.1.3 that likelihood that a patient visits a doctor in t given h_{t-1} is given by

$$p^{h_{t-1}} \cdot \frac{1}{1 + \exp\left(-\frac{V^{go}_t(R_{it}) - \tilde{\kappa} - V^{ngo}_t(R_{it})}{s_\kappa}\right)}.$$

[5] $\bar{V}^{go}_{T-1}(R_{iT-1}, \lambda_{iT-1})$ is the value of visiting a doctor for a given combination of R_{iT-1} and λ_{iT-1}. This becomes known to the patient when she visits a doctor.

[6] This value will be different from the one used in $V^{go}_{T-1}(R_{iT-1})$ because R_{iT} will be different.

B.2 Estimation

B.2.1 Overview

Collect the parameters in θ and denote the parameter space by Θ. Then, our estimates are given by

$$\hat{\theta} = \arg\min_{\theta \in \Theta} m(\theta)' W m(\theta) \tag{11}$$

where $m(\theta)$ is a vector of moment conditions and W is a positive definite weighting matrix. This is the GMM estimator. We calculate standard errors using the usual formula. Next, we describe how we construct the moment conditions and the weighting matrix.

B.2.2 Moment conditions

$m(\theta)$ is a vector of moment conditions. To calculate these moment conditions, we solve the model and calculate the following four quantities for each t and each point on the grid for R_{it}:

1. expected spending when spending is positive
2. the likelihood that spending is positive
3. the likelihood that spending is positive given that the patient had a healthcare need in the previous period
4. the variance of log spending provided that spending is positive.

Then, for each patient i in each time period t, we use interpolation and the remaining deductible in our data to calculate values of these variables at the individual-month level. Denote the resulting predictions from our model by $\hat{y}^1_{it}(\theta)$, $\hat{y}^2_{it}(\theta)$, $\hat{y}^3_{it}(\theta)$, and $\hat{y}^4_{it}(\theta)$, respectively. From this, we calculate the following four variables:

1. $\hat{u}^1_{it}(\theta)$: an indicator for positive observed spending interacted with the difference between observed spending and $\hat{y}^1_{it}(\theta)$
2. $\hat{u}^2_{it}(\theta)$: the difference between an indicator for positive observed spending and $\hat{y}^2_{it}(\theta)$

3. $\hat{u}^3_{it}(\theta)$: an indicator for positive observed spending in the previous period interacted with the difference between an indicator for positive observed spending and $\hat{y}^3_{it}(\theta)$

4. $\hat{u}^4_{it}(\theta)$: an indicator for positive observed spending interacted with the difference between squared log spending minus average log spending squared minus $\hat{y}^4_{it}(\theta)$.

Observe that for any t and in expectation over i, each of these four variables is zero. We stack these four variables for all patients, i, and all 11 months from February to December, t, into vectors $\hat{u}^1(\theta)$, $\hat{u}^2(\theta)$, $\hat{u}^3(\theta)$, and $\hat{u}^4(\theta)$. The ordering within each of those vectors is first by patient, i, and then by month, t.

Next, we construct a matrix Z that includes indicators for combinations of months and four intervals for the remaining deductible. The rows of Z correspond to the rows of the three vectors $\hat{u}^1(\theta)$, $\hat{u}^2(\theta)$, and $\hat{u}^3(\theta)$, respectively. The columns contain indicators for interactions between all months from February onward and intervals for the remaining deductible. We order the columns first by month and then, for each month, by remaining deductible interval. The intervals we use for the remaining deductible are: above deductible ($R_{it} = 0$), up to 100 remaining deductible ($R_{it} \in (0, 100]$), between 100 and 200 remaining deductible ($R_{it} \in (100, 200]$), and more than 200 remaining deductible ($R_{it} \in (200, 350]$).

We also construct a matrix $\bar{Z}$. The rows of $\bar{Z}$ correspond to the rows of the vector $\hat{u}^4(\theta)$. The columns contain indicators for interactions between all months from February onward and being above the cost-sharing limit.

Denote the number of individuals in our data by N. Then, we have that each of the vectors $\hat{u}^1(\theta)$, $\hat{u}^2(\theta)$, $\hat{u}^3(\theta)$, and $\hat{u}^4(\theta)$ is of dimension $(N \cdot 11) \times 1$, because for each individual we use data for 11 months. Z is of dimension $(N \cdot 11) \times 44$, because Z contains indicators for combinations of 11 months and 4 intervals for the remaining deductible. $\bar{Z}$ is of dimension $(N \cdot 11) \times 11$, because there is an indicator for being above the deductible for each month.

Using this, we calculate:

$$m(\theta) = \begin{bmatrix} Z'\hat{u}^1(\theta)/N \\ Z'\hat{u}^2(\theta)/N \\ Z'\hat{u}^3(\theta)/N \\ \bar{Z}'\hat{u}^4(\theta)/N \end{bmatrix}.$$

$m(\theta)$ is a vector with $3 \cdot 44 + 11$ elements. The first element is the average of the product of $\hat{u}_{it}^1(\theta)$ and an indicator that is equal to 1 if the observation is for February and R_{it} is in the first deductible interval. The second element is the average of the product of $\hat{u}_{it}^1(\theta)$ and an indicator that is equal to 1 if the observation is for February and R_{it} is in the second deductible interval, and so on. The very last element of $m(\theta)$ is the average of the product of $\hat{u}_{it}^4(\theta)$ and an indicator that is equal to 1 if the observation is for December and the individual is above the cost-sharing limit.

B.2.3 Weighting matrix

We use a diagonal weighting matrix that puts more weight on elements of $m(\theta)$ that represent combination of calendar month and remaining deductible interval that are more commonly observed in our data. To obtain the numbers of observations, denote the diagonal of the matrix $Z'Z$ by $\vec{N}$. This is a vector with 44 elements. Its first element, $\vec{N}_1$, is the number of patients for whom the remaining deductible falls in the first interval in February. The second element, $\vec{N}_2$, is the number of patients for whom the remaining deductible falls in the second interval in February, and so on.

Likewise, denote the diagonal of the matrix $\bar{Z}'\bar{Z}$ by $\vec{\bar{N}}$. This vector has 11 elements, every fourth of $\vec{N}$ starting with the fourth, because we only use data for observations that fall in the last deductible interval where patients are above the cost-sharing limit.

The scaling differs across moments. Therefore, we calculate the variance of observed spending conditional on any spending, the variance of the indicator for any spending, the variance of an indicator for any spending given any spending in the previous period, and the variance of squared log spending conditional on any spending, for any month. Denote these estimates by $\hat{\sigma}_{1,t}^2$, $\hat{\sigma}_{2,t}^2$, $\hat{\sigma}_{3,t}^2$, and $\hat{\sigma}_{4,t}^2$, respectively.

The elements on the diagonal of the weighting matrix are given by $\vec{N}_1 / \hat{\sigma}_{1,1}^2$ up to $\vec{N}_{12} / \hat{\sigma}_{1,12}^2$ for moments based on $\hat{u}_{it}^1(\theta)$, followed by $\vec{N}_1 / \hat{\sigma}_{2,1}^2$ up to $\vec{N}_{12} / \hat{\sigma}_{2,12}^2$ for moments based on $\hat{u}_{it}^2(\theta)$, followed by $\vec{N}_1 / \hat{\sigma}_{3,1}^2$ up to $\vec{N}_{12} / \hat{\sigma}_{3,12}^2$ for moments based on $\hat{u}_{it}^3(\theta)$, and finally followed by $\vec{\bar{N}}_1 / \hat{\sigma}_{4,1}^2$ up to $\vec{\bar{N}}_{12} / \hat{\sigma}_{4,12}^2$ for moments based on $\hat{u}_{it}^4(\theta)$. This gives more weight to moments that are based on combination of calendar month and remaining deductible interval that are

more commonly observed and more weight to moments that are less noisy, as measured by $\hat{\sigma}^2_{1,t}$, $\hat{\sigma}^2_{2,t}$, $\hat{\sigma}^2_{3,t}$, and $\hat{\sigma}^2_{4,t}$.

B.3 Counterfactuals

To conduct our counterfactual simulations, we solve the model for a given policy and simulate paths of healthcare consumption for 10,000 simulated patients for each risk score quartile. For each simulated patient, healthcare consumption in the first period depends on whether patients had a need in the last period of the previous year. Therefore, for each simulated patient, we draw a healthcare need for the last period of the previous year from the ergodic distribution of the healthcare needs process (see Section 5.1.2).

One observation in the simulated data is a simulated patient in a given month for a given policy. From this, we calculate the numbers reported in the table by taking averages and standard deviations over simulated patients and months.