DISCUSSION PAPER SERIES

DP17057

A structural microsimulation model for demand-side cost-sharing in healthcare

Jan Boone and Minke Remmerswaal

PUBLIC ECONOMICS

ISSN 0265-8003

A structural microsimulation model for demand-side cost-sharing in healthcare

Jan Boone and Minke Remmerswaal

Discussion Paper DP17057
Published 21 February 2022
Submitted 21 February 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

A structural microsimulation model for demand-side cost-sharing in healthcare

Abstract

Demand-side cost-sharing schemes reduce moral hazard in healthcare at the expense of out-of-pocket risk. With a structural microsimulation model, we show that shifting the starting point of the deductible away from zero to 400 euros for all insured individuals, leads to an average 4% reduction in healthcare expenditure and 47% lower out-of-pocket payments for people without a chronic condition. We use administrative healthcare expenditure data for the Dutch population to analyze expenditure under different cost-sharing schemes. The model is estimated with a Bayesian mixture model to capture distributions of healthcare expenditure. This allows us to predict the effects of cost-sharing schemes that have not been used in the Netherlands.

JEL Classification: N/A

Keywords: moral hazard, risk, equity, out-of-pocket, shifted deductible, Co-Insurance, Bayesian mixture model, microsimulation model

Jan Boone - j.boone@tilburguniversity.edu
Tilburg University and CEPR

Minke Remmerswaal - M.Remmerswaal@cpb.nl
CPB Netherlands Bureau for Economic Policy Analysis

Acknowledgements
We gratefully acknowledge comments and suggestions by Bram Wouterse, Rob Aalbers, Pieter Boone, Rudy Douven, Richard van Kleef, Tobias Klein, Jana van Kuringen, Johan Polder, Gert-Anne van Pruissen, Martin Salm, Anouk van Schijndel, Wouter Vermeulen, Dani\"elle Willemse-Duijmelinck, Robin Zoutenbier, and seminar participants at CPB Netherlands Bureau for Economic Policy Analysis, the Centre for the Business and Economics of Health at the University of Queensland, the European Commission, 2019 ENTER Jamboree, the Dutch Ministry of Health, and Tilburg University.

A structural microsimulation model for demand-side cost-sharing in healthcare*

Jan Boone[1, 2, 3] and Minke Remmerswaal[1, 2]

[1]CPB Netherlands Bureau for Economic Policy Analysis, The Hague, the Netherlands
[2]Department of Economics, Tilec, Tilburg University, the Netherlands
[3]CEPR, London, United Kingdom

February 18, 2022

Demand-side cost-sharing schemes reduce moral hazard in healthcare at the expense of out-of-pocket risk. With a structural microsimulation model, we show that shifting the starting point of the deductible away from zero to 400 euros for all insured individuals, leads to an average 4% reduction in healthcare expenditure and 47% lower out-of-pocket payments for people without a chronic condition. We use administrative healthcare expenditure data for the Dutch population to analyze expenditure under different cost-sharing schemes. The model is estimated with a Bayesian mixture model to capture distributions of healthcare expenditure. This allows us to predict the effects of cost-sharing schemes that have not been used in the Netherlands.

Keywords: moral hazard, risk, equity, out-of-pocket, shifted deductible, co-insurance, Bayesian mixture model, microsimulation model

*We gratefully acknowledge comments and suggestions by Bram Wouterse, Rob Aalbers, Pieter Boone, Rudy Douven, Richard van Kleef, Tobias Klein, Jana van Kuringen, Johan Polder, Gert-Anne van Pruissen, Martin Salm, Anouk van Schijndel, Wouter Vermeulen, Daniëlle Willemse-Duijmelinck, Robin Zoutenbier, and seminar participants at CPB Netherlands Bureau for Economic Policy Analysis, the Centre for the Business and Economics of Health at the University of Queensland, the European Commission, 2019 ENTER Jamboree, the Dutch Ministry of Health, and Tilburg University.

1 Introduction

Most healthcare systems and health insurance plans have some form of cost-sharing in place. Cost-sharing is useful as it prevents moral hazard and curbs healthcare expenditure by shifting part of healthcare costs to users of care (Zweifel and Manning, 2000). As such, it can improve efficiency. However, cost-sharing is also the topic of political debates, as it introduces a risk of out-of-pocket expenditure and (chronically) ill people carry the largest financial burden because of high out-of-pocket payments. Cost-sharing as such can imply a trade-off between efficiency and equity (solidarity). In this paper, we develop a structural microsimulation model for demand-side cost-sharing in the Netherlands, which we estimate on a large administrative dataset of inhabitants in the Netherlands. With the model we can show how multiple demand-side cost-sharing schemes affect the trade-off between equity and efficiency. One of our main results is that shifting the starting point of the deductible in the Netherlands by 400 euros leads to an average 4% reduction in healthcare expenditure and 47% lower out-of-pocket payments for people without a chronic condition.

A cost-sharing scheme puts a price on healthcare for insured individuals. However, the effectiveness of the scheme in reducing moral hazard is determined by the *effective* price that individuals experience (Keeler et al., 1977). The effective price depends on the design of the cost-sharing scheme and individuals' health status (more specifically, individuals' expected healthcare costs), and whether they are 'at the margin'. To illustrate, a deductible of 100 euros in a healthcare plan for chronically ill individuals – who know they will have high healthcare expenditures – is unlikely to be very effective in reducing their healthcare expenditure: they are not at the margin. However, the same 100 euro deductible in a plan for students – who tend to be young and healthy – is likely to be more effective. In order to choose the optimal scheme, one must have detailed information about the distribution of healthcare expenditure of the targeted population. Moreover, the optimal design of a scheme will vary across populations and countries with varying expected healthcare expenditure for example due to differences in demographics, health status, and institutional setting.

We illustrate this in the four panels in Figure 1 for an increase in the deductible from 150 to 350 euros. The top-left panel of the figure shows mean healthcare expenditure for men and women for each age in the Netherlands for 2008.[1] Two solid, black horizontal lines represent the lowest deductible in our data of 150 euros (in place in the Netherlands in 2008, horizontal axis in the figure) and highest deductible of 350 euros (in 2013). The graph illustrates that the fraction of individuals affected by an increase in the deductible from 150 euros to 350 euros varies by age and gender: the average expenditure for men are around 350 euros between age

[1] The exact definition of healthcare expenditure that we use in this paper is given in Section 3.2. The figure is drawn using our training data set.

20 and 40, whereas average expenditures for women aged 30 to 40 years are well above the highest deductible level. Based on the average expenditures shown in the top-left panel in the figure for each gender-age category, it seems as if only a few persons would be affected (at the margin) when increasing the deductible from 150 to 350 euros. If we zoom in however, into the distribution for each gender-age category, we see that looking at mean values is not enough and that the distribution is important.

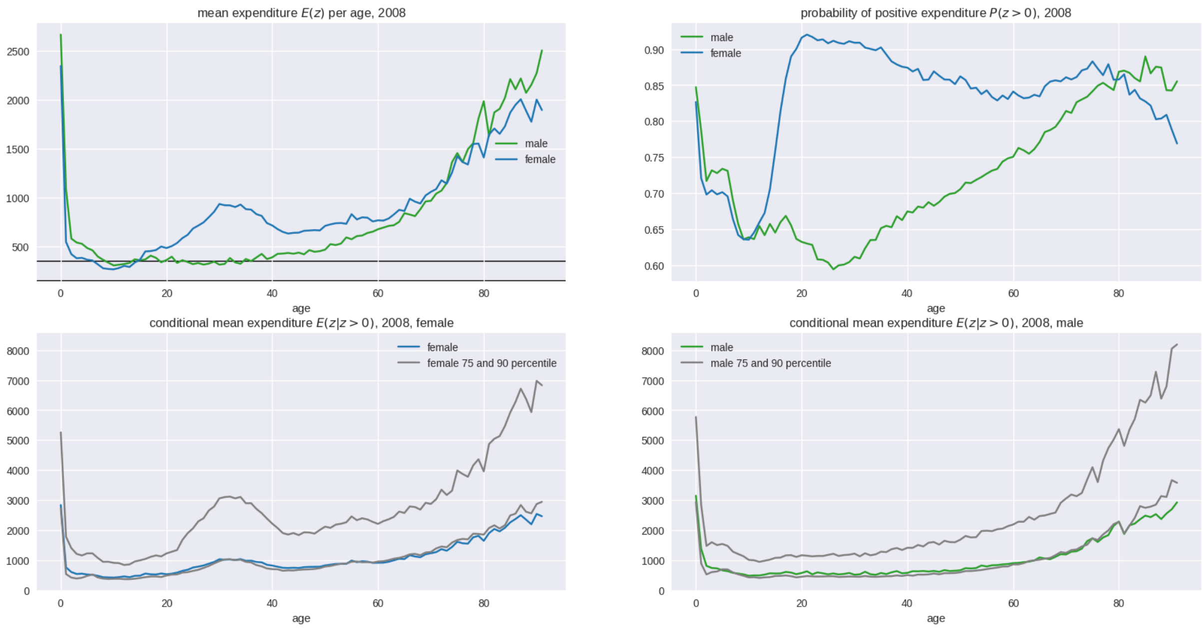

Figure 1: Healthcare expenditure across ages for men and women

The top-right panel shows for each gender-age category the probability that expenditure is positive. There is a substantial share of men and women who have no healthcare costs in a year. Also, the variation across gender and age is large. It is unlikely that everyone would react in the same way to an increase in the deductible. The bottom row in the figure shows the expected expenditure, conditional on positive expenditures. These figures also show the 75th and 90th percentile. The overall picture is clear: the distribution of healthcare expenditure per age group is highly skewed. Many people have zero expenditure: as shown in the top-right panel, this varies roughly between 40% and 5% per gender-age category. To be able to assess and quantify the effect of different cost-sharing schemes, modeling the distribution of healthcare expenditure across various groups of individuals is necessary. In this paper, we model these distributions explicitly and take into account that expenditure in a gender-age category is likely to be more elastic with respect to the deductible if people are more likely to be at the margin.

The paper begins with a theoretical model where a person can face two types of healthcare expenditures: exogenous and endogenous expenditures. Exogenous expenditures are of such high value that changes in the deductible do not affect the decision to accept such a treatment.[2] Think of breaking a leg and being plastered up in hospital. We identify such expenditures in the data as they do not vary with the size of the deductible over time. For the endogenous expenditures, people with a high deductible are more likely to forego the treatment. This decision is based on a person's expected out-of-pocket payment of a treatment ($EOOP$) where the expectation is taken over the distribution of his or her healthcare expenditures. This captures that a 70 year old – presumably – faces higher (exogenous) expenditure than a 20 year old. Hence, for a given treatment the $EOOP$ is lower for the former, as the 70 year old is likely to exhaust the deductible anyway.

This model is brought to the data by approximating the expenditure distribution for each gender-age-year combination with a mixture consisting of three distributions: zero expenditure, exogenous and endogenous expenditure distributions. An increase in $EOOP$ (say, due to an increase in the deductible) puts more weight on zero expenditure and less on the endogenous expenditure distribution. We assume Gaussian processes for both endogenous and exogenous healthcare expenditures across age. In contrast to age fixed effects, this implies that knowing the average healthcare expenditures of for example 20 year old males, helps to predict healthcare expenditures of 21 year old males. Figure 1 suggests that such correlation across age indeed exists. The estimations provide posterior distributions for the parameters in the model and with these parameters we can determine the expenditure distributions for each gender-age category. Once we have the expenditure distributions, we can simulate healthcare expenditure under different cost-sharing schemes. We use the posterior distributions throughout the paper and hence can indicate the uncertainty surrounding any results that we present.

Our approach, using the expected out-of-pocket payment $EOOP$ as the effective price that individuals face, is similar, yet different, to using the end-of-year price (Keeler et al., 1977; Ellis, 1986). For a deductible, the end-of-year price can be estimated as the probability that a person does not exceed the deductible until the end of the contract period, which is a year in the Netherlands. The end-of-year price captures how much one more euro of healthcare costs a person. Our data lack a within year time dimension, hence we cannot follow how individuals' healthcare expenditures evolve within the year. Instead, we integrate over the distributions of expenses that an individual faces over the coming period to calculate the price of treatment, $EOOP$. Whereas the end-of-year price is the (marginal) price of using *one* more euro of healthcare, $EOOP$ is the expected expenditure that corresponds to the price of accepting an *additional treatment*.[3]

The model is estimated on an administrative dataset which comprises healthcare expendi-

[2]This is given the relatively modest size of the deductible in Dutch healthcare.

[3]With only one type of expenditure (not distinguishing exogenous and endogenous expenditures) and considering an additional one euro expense, our $EOOP$ equals the end-of-year price.

tures of Dutch inhabitants between 2008 and 2013. The main results of the model are presented in Figure 2. The left panel presents the effect of different levels of deductibles, co-insurance schemes and shifted deductibles which start at 400 euros on healthcare expenditure. In this figure, we compare all results to a 300 euro deductible by normalizing expenditure on the expenditure with a 300 euro deductible. As expected, schemes with higher maximum out-of-pocket payments lead to lower healthcare expenditures. Having no demand-side cost-sharing, that is, the maximum out-of-pocket payment is zero, leads to approximately 8% higher costs compared to a 300 euro deductible. With a deductible equal to 600 euros, expenditures decrease by more than 10%. Increasing the deductible reduces expenditures per head, but at the cost of higher out-of-pocket risk for insured. This is illustrated in the right panel of the figure. With a 400 euro deductible, the average out-of-pocket payments in all gender-age categories are roughly 20% higher compared to the average out-of-pocket payments under a 300 euro deductible.

Co-insurance schemes and shifted deductibles with a maximum out-of-pocket payment of 300 euros reduce both healthcare expenditure and the average out-of-pocket payment per head. This is because with a co-insurance scheme, people face cost-sharing over a longer range of healthcare expenditures: not between zero and 300 euros, but between zero and 1,200 euros, and at a lower rate of 0.25 instead of 1.0. Note that we do not make any assumptions on how this trade-off between a longer range and lower rate affect expenditures. It follows endogenously from the model by using $EOOP$ and by integrating over the estimated distributions. In case of a shifted deductible which starts at 400 euros, people face no cost-sharing for expenditures in the interval $[0, 400]$. Then for expenditure $x \in [400, 700]$, they pay $x - 400$ euros. For $x > 700$, they pay the maximum out-of-pocket payment of 300 euros. The co-insurance scheme and the shifted deductible lead to a bigger reduction of healthcare expenditure compared to a 300 euro deductible because they increase the effective price of healthcare that individuals experience given their expected healthcare expenditures. With these schemes there are more persons at the margin and over a longer range of healthcare expenditures. The reduction in healthcare expenditures and out-of-pocket payments alleviates the trade-off between efficiency and equity.[4]

There is a large body of literature on demand-side cost-sharing and moral hazard. Many models in the cost-sharing literature, such as Einav et al. (2013), Cardon and Hendel (2001), and Bajari et al. (2014), add more structure to their models than we do because they also need to model the decision whether to buy insurance, and how generous the insurance should be. As we explain in Section 3, these decisions are not relevant in the context of mandatory basic insurance in the Netherlands and the simulations that we do. Our paper is most comparable to

[4]Note that for our analysis we do not need to model risk aversion because (basic) health insurance is mandatory in the Netherlands. However, in terms of the trade-off between efficiency and risk, it is clear that in all states of the world the out-of-pocket is (weakly) lower with a shifted deductible than with the same deductible level starting at 0.

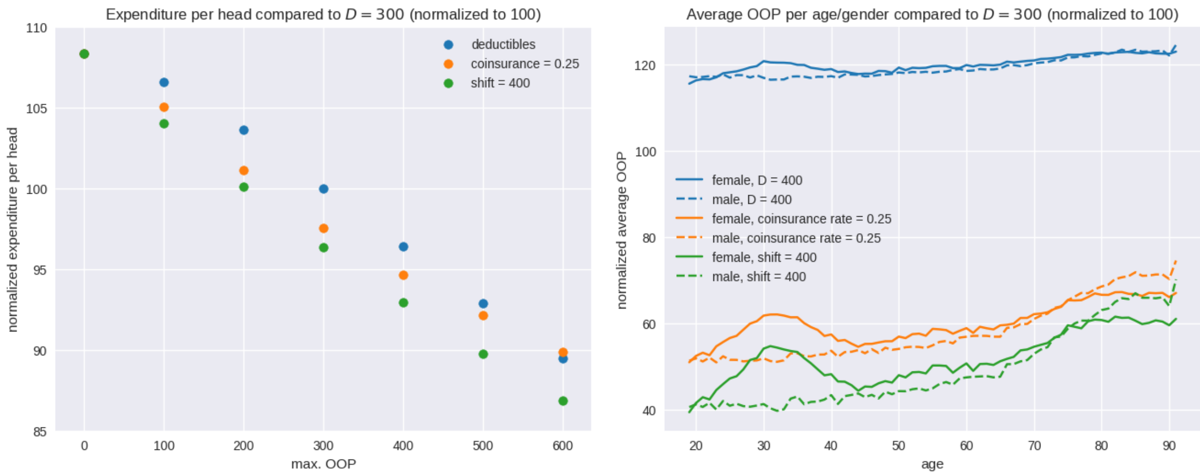

Figure 2: Effects of demand-side cost-sharing on expenditure per head and out-of-pocket payments

Einav et al. (2013) who develop a model to show evidence of selection on moral hazard. Like our paper, Einav et al. (2013) make a distinction between exogenous and endogenous healthcare expenditures and estimate the model using Bayesian methods with panel data. Their setting is, however, more specific: they study health insurance offered by one employer in the United States, whereas we have data for the entire population in the Netherlands. The results of our simulations can therefore be directly used by Dutch policy makers.[5] Furthermore, Einav et al. (2013) only compare a zero-deductible scheme to a high-deductible scheme, whereas we simulate and compare other types of cost-sharing than a deductible.

Other papers using Bayesian estimation techniques to estimate healthcare expenditure models are, for example, Deb et al. (2006), Jochmann and Leon-Gonzales (2004), and Hamilton (1999). Like the papers mentioned above, these papers also focus on modeling both the decision to purchase insurance as well as the level of healthcare consumption. Mukherji et al. (2016) estimate health demand for aging populations with a Bayesian model. They also adopt a mixture model, but with two components, to capture both the zeroes as well as the distribution of positive healthcare expenditures. Mukherji et al. (2016) stress the importance of explicitly modeling the nonlinear effects of age interacted with gender on healthcare expenditure. They use a flexible spline model, whereas we use Gaussian processes (GPs) to model average expenditures. Moreover, we have a four component mixture model to capture both endogenous and exogenous expenditures.

The Congressional Budget Office's Medicare Beneficiary Cost-Sharing Model has a similar purpose to ours: to estimate the budgetary effects of various cost-sharing schemes at a population level (Duchovny et al., 2019). Like this paper, they use a microsimulation model which

[5] In fact, a version of this model is used by the CPB Netherlands Bureau for Economic Policy Analysis to analyze and forecast the budgetary impact of cost-sharing schemes proposed by policy makers and political parties.

they estimate on administrative data, and they can also present distributional effects of cost-sharing schemes across beneficiairies. Unlike our paper, Duchovny et al. (2019) do not estimate the behavioral responses; they primarily apply the results from the RAND Health Insurance Experiment (Newhouse and the Insurance Experiment Group, 1993).

This paper builds on previous empirical work in which we study the effect of the deductible size for 18 year olds and estimate selection effects due to the voluntary deductible (Remmerswaal et al., 2019b). With a panel regression discontinuity design we exploit the introduction of the mandatory deductible at age 18 and the annual increase of the deductible size by the government. Remmerswaal et al. (2019b) find a deductible elasticity of -0.09. In Section 6 we show that the average deductible elasticity in our model here is the same.

Van Kleef et al. (2009) and Cattel et al. (2017) already show the shifted deductible's potential for reducing moral hazard in the Netherlands. In this paper, we build on this by quantifying *how much* a shifted deductible can reduce healthcare expenditures and, in fact, out-of-pocket payments.[6] Like our paper, Van Kleef et al. (2009) find that the optimal starting point is not zero (so not a traditional deductible), but positive for all individuals. According to Van Kleef et al. (2009), the optimal starting point of a 500 euro deductible is 879 euros. In this paper we do not perform a full grid search to find the optimal starting point, but our 'best' starting point is somewhat smaller. For a shifted deductible with a maximum out-of-pocket payment of 300 euros, a starting point of 400 euros gives the largest reduction of healthcare expenditure compared to a deductible of 300 euros which kicks in directly. Van Kleef et al. (2009) stress that different risk groups should face different starting points as by setting a uniform starting point, you either reduce incentives for the low risks (for whom the starting point is likely too high) or it has no effect on high risks (for whom the starting point is likely too low). In this paper, we do not use the (risk adjustment) indicator which labels individuals as chronically ill or a chronic user of medication to determine the optimal starting point for the following reason. In our data, the maximum deductible equals 350 euros which is far below the relevant margin for people with a chronic condition. Hence, we cannot estimate their elasticity in such a way that we can confidently do a counterfactual with a higher shift of, say, 1,000 euros. We do provide an additional analysis in which we simulate the effects of a starting point which is risk adjusted in the sense that it depends on age.

The theoretical model is presented in Section 2. In Section 3, we describe the institutional setting of healthcare in the Netherlands and the data. We explain how we parameterize and identify the model in Section 4, and discuss the estimation methodology and fit of the model

[6]Note that Cattel et al. (2017) and Van Kleef et al. (2009) use a different terminology than we do for the differences between cost-sharing schemes. We use the term 'shifted deductibles' for all deductibles with a starting point strictly above zero. Cattel et al. (2017) and Van Kleef et al. (2009) refer to shifted deductibles with a uniform starting point as 'doughnut deductibles' and deductibles with a differentiated or individual starting point as 'shifted deductibles'.

in Section 5. Section 6 presents the results of simulating deductibles, co-insurance rates, and shifted deductibles with the model. We present robustness tests in Section 7 and conclude in Section 8.

The Python code for our paper is available at `https://gitlab.uvt.nl/janboone/micro_simulation_model`.

2 The model

The main aim of this paper is to simulate healthcare expenditures per individual for different demand-side cost-sharing schemes. We do this by modeling the distribution of expenditures per gender-age category. The effect of a change in cost-sharing comes through a change in the price of care: the expected out-of-pocket payment ($EOOP$). $EOOP$ depends on the level and type of cost-sharing and the healthcare expenditures an individual expects to make.

To illustrate, consider a person who is offered a treatment. An increase in the deductible level will on average imply that he or she will pay more for the same treatment out-of-pocket. However, if this person has high expected costs, such that he or she is more likely to exhaust the deductible with other treatments anyway, then the additional out-of-pocket payment of the offered treatment is very low or zero and the treatment is more likely accepted. We assume that the treatment is accepted if the treatment's value exceeds its $EOOP$.

We allow for two types of treatment: high value treatments which are exogenous with regard to cost-sharing, and lower value treatments which are endogenous with respect to cost-sharing. The 'exogenous treatments' lead to a distribution of expenditure against which the $EOOP$ is calculated for an 'endogenous treatment'. As either treatment type can be offered or not to an individual, we use a four component mixture model to capture healthcare expenditures.

Note that the model is estimated on Dutch administrative data for 2008 to 2013 (see Section 3.2). In those years, only a deductible was in place. Some of the choices for the model below are driven by the fact that we want to simulate various types of cost-sharing, not only the deductible for which we have data.

2.1 Expenditures

In our model, the distribution of total healthcare expenditure per capita per year is generated by two types of treatment. The first type, denoted by x, is exogenous to cost-sharing, and consists of high value procedures that are always carried out, regardless of the size of cost-sharing (over the relevant Dutch policy range).[7] One can think of plastering up a broken leg or being taken to hospital after a stroke or cardiac arrest. The second treatment type, y, is endogenous with

[7] The model is based on the Dutch healthcare system, in which the level of cost-sharing is low.

regard to the deductible and consists of treatments where persons *do* take the deductible into account. For example, if a physician offers a person extra imaging services, he or she may or may not accept this.[8]

We assume that healthcare expenditures, conditional on being positive, are lognormally distributed. Below we show that this assumption has two major computational benefits when estimating the model, and in Section 3.4 we show that this lognormal assumption is a reasonable representation of the data. Hence, we transform our observed healthcare expenditures Z into logs: $z = \ln(1+Z)$, where 1 (euro) is added to avoid taking the logarithm of zero.[9] As a result, $Z = 0$ if and only if $z = 0$; otherwise $z > 0$. As we assume that $Z|Z > 0$ (i.e. Z conditional on Z being positive) has a lognormal distribution, $z|z > 0$ is normally distributed.

As x and y denote *positive* healthcare expenditures, we also model the probability that someone gets offered an x treatment, that is, has a positive draw of x, as ψ_x. This treatment, if offered to a patient, is always accepted since it does not depend on the deductible by definition. The probability that someone is offered a y treatment is denoted by ψ_y. The probability that this treatment is rejected is denoted by F, to be determined below. We assume that the Bernoulli draws with probabilities ψ_x and ψ_y are independent.

2.2 Mixture model

In the way we view our data, observed (log) total healthcare expenditure is essentially the sum of x and y; $z = x + y$. This implies that the model consists of four components as illustrated in Table 1.

There are four possible outcomes for log expenditure: $z = 0$ if there is neither an x nor an y treatment. The probability that this happens is the multiplication of the probability that $x = 0$, $1 - \psi_x$, and the probability that $y = 0$. The latter can happen in two ways: either no y treatment was offered $(1 - \psi_y)$, or it was offered, but rejected $(\psi_y F)$. Further, the probability that $x > 0$ is given by ψ_x and the probability that $y > 0$ by $\psi_y(1-F)$: y is offered and accepted. Hence, the probability that both $x > 0$ and $y > 0$ is given by $\psi_x\psi_y(1 - F)$.

Table 1: The distribution of total log expenditure

Component	Probability
$x = y = 0$	$(1 - \psi_x)(1 - \psi_y + \psi_y F)$
$x > 0 = y$	$\psi_x(1 - \psi_y + \psi_y F)$
$y > 0 = x$	$(1 - \psi_x)\psi_y(1 - F)$
$x, y > 0$	$\psi_x\psi_y(1 - F)$

[8]Note that we do not label some (types of) healthcare expenditures as x or y ourselves; the distinction between x and y comes from the observation that some expenditures vary with the deductible level and others do not.

[9]If a person has positive healthcare expenditure, these expenditures tend to be of the order of 100 euros and higher; adding one euro is immaterial.

In line with the normality assumption on $z|z>0$, we also assume that $x|x>0$ and $y|y>0$ are normally distributed. This implies that with the exception of $x=y=0$, each component in Table 1 is normally distributed. Let the parameters for the $x|x>0$ distribution be given by μ_x, σ_x and similarly μ_y, σ_y for $y|y>0$. Then the distribution of the last component, $x, y>0$ in Table 1 is normal with parameters $\mu_x+\mu_y$ and $\sqrt{\sigma_x^2+\sigma_y^2}$, since we assume that the x and y processes are independent (conditional on age and gender). This is the first computational gain of assuming a lognormal distribution for healthcare expenditure in estimating our model: we have analytical expressions for each of the four components.

2.3 Out-of-pocket payments

The second computational advantage of assuming a lognormal distribution becomes clear when calculating $EOOP$. We think of $EOOP$ as follows: at the start of the period, before exogenous expenditure x has been realized, a person is offered an endogenous treatment y. When the treatment is offered, he or she does not know exactly the price of this treatment; e.g. the doctor advises to see a dermatologist, but the patient does not know which treatment will be needed exactly. Nor are the exogenous x treatments known that may be needed later in the year. Therefore, we model $EOOP$ as an integral over both x and y for each gender-age category.

Such a definition of $EOOP$ captures that a woman at age 30 has a lower $EOOP$ than a man at the same age (see Figure 1). When being offered a treatment, the woman knows that she is likely to exceed the deductible anyway and hence the y treatment is basically free. For the 30 year old male, this is not the case and accepting the y treatment will turn out to cost him money at the end of the year. For a given value distribution of the treatment, the probability that the man accepts the treatment is lower than for the woman.

Although there is evidence that people respond to spot prices and not to end-of-year prices because they are myopic, discount future costs, or lack information (Brot-Goldberg et al., 2017; Remmerswaal et al., 2019a), we assume here they have a least some idea of their health status, the healthcare they need, and the cost of those treatments, and we incorporate that in their $EOOP$.[10] We model $EOOP$ as the expectation across the distributions of x and y.[11] As we will see below, the average y expenditures are rather modest. This is to be expected as some of these treatments are rejected because their value does not exceed their cost, which can maximally be 350 euros, the maximum deductible size in our data. As y expenditures therefore tend to be relatively low, it is not necessary to allow for different distributions of the y treatments.[12]

[10] For example, Einav et al. (2015) also assume individuals are forwardlooking in their model. Furtermore, Klein et al. (2020) find that persons in Dutch health insurance are forwardlooking.

[11] In Section 7 we also test the validity of this assumption by changing the specification of $EOOP$.

[12] Technically speaking this is feasible. We can allow for n different y distributions. If a person is offered a y treatment, we first draw the y distribution from which he or she will draw. Then, with this distribution, we calculate $EOOP$. We can choose a value for n or estimate it as well. Note that having a sequence of y treatments

Moreover, as shown below, the fit of our model is fairly good. In this sense, there is no need to extend it with a number of distributions from which y can be drawn.

First drawing a value of y and have a person decide whether to accept the treatment based on this expenditure would complicate our estimation. If, say, more expensive y treatments are more likely to be rejected then the distribution of accepted y treatments is no longer normal (as the distribution of offered y treatments is). This implies that in the estimation we would have to simulate the accepted y distribution which increases the computational complexity considerably.

We use the following well known result:

Lemma 1 *Consider a variable χ which is lognormally distributed with parameters μ, σ. Let N denote the cumulative distribution function of a standard normal distribution with a mean of zero and a standard deviation of 1. Then the probability that $\chi < D$ is given by*

$$P(\chi < D) = N\left(\frac{\ln(D) - \mu}{\sigma}\right) \tag{1}$$

Further, let f_χ denote the density function of χ, then

$$\int^D x f_\chi(x)dx = P(\chi < D)E(\chi|\chi < D) = e^{\mu+\frac{\sigma^2}{2}} N\left(\frac{\ln(D) - \mu - \sigma^2}{\sigma}\right) \tag{2}$$

With a deductible level of D, the out-of-pocket payment (OOP) as a function of μ, σ is given by

$$OOP(\mu, \sigma) = \int \min\{x, D\} f_\chi(x)dx = e^{\mu+\frac{\sigma^2}{2}} N\left(\frac{\ln(D) - \mu - \sigma^2}{\sigma}\right) + \left(1 - N\left(\frac{\ln(D) - \mu}{\sigma}\right)\right) D \tag{3}$$

Consider a person who is offered a y treatment and who considers the expected cost of accepting this treatment. This expected cost is given by:

$$EOOP = (1 - \psi_x)OOP(\mu_y, \sigma_y) + \psi_x(OOP(\mu_x + \mu_y, \sqrt{\sigma_x^2 + \sigma_y^2}) - OOP(\mu_x, \sigma_x)) \tag{4}$$

With a probability $1 - \psi_x$, this person has no other (exogenous) costs during the year. $EOOP$ is then given by the expected y cost. With a probability ψ_x, there will be an x as well as a y cost. $EOOP$ is then determined by the difference between the out-of-pocket cost of both x and y and the out-of-pocket cost of only x. Because costs are lognormally distributed and y costs are only known *after* accepting the treatment, there is an analytic expression for this out-of-pocket expenditure $OOP(\mu_x + \mu_y, \sqrt{\sigma_x^2 + \sigma_y^2})$. This is the second computational gain of assuming a

being offered is also feasible but more complicated. In that case, we need to solve the dynamic optimization problem where a person takes into account that accepting the first y treatment makes the following y treatments cheaper. Each expenditure then has the additional benefit of lowering the effective or end-of-year price (Keeler et al., 1977; Ellis, 1986). Both extensions are left for future research.

lognormal distribution of healthcare expenditures.

In our simulations, we consider cost-sharing schemes of the form:

$$\max\{0, \min\{\delta(Z - \Delta), D\}\} \tag{5}$$

where Z are an individual's total healthcare expenditure (in levels) and $\Delta \geq 0$ denotes the shift or the starting point of the cost-sharing scheme. Δ is zero for a traditional deductible, but positive for a shifted deductible. The co-insurance rate $\delta \in \langle 0, 1]$ is the percentage of healthcare costs an individual has to pay out-of-pocket. D denotes the maximum out-of-pocket expenditure.

Equation (5) encompasses many types of cost-sharing schemes. For example, a traditional deductible can be summarized as: no shift ($\Delta = 0$), persons pay the full price until the maximum is reached ($\delta = 1$), and a maximum out-of-pocket ($D > 0$). A typical co-insurance scheme also has no shift ($\Delta = 0$) and a maximum out-of-pocket payment ($D > 0$), but individuals pay a percentage of total healthcare costs out-of-pocket ($\delta \in \langle 0, 1\rangle$). A shifted deductible can be described as: cost-sharing does not kick in directly ($\Delta > 0$), persons pay the full price until the maximum is reached ($\delta = 1$), and a maximum out-of-pocket ($D > 0$). But combinations are also possible. For example, Figure 3 illustrates a payment scheme with $\Delta = 200$ euros, so the first 200 euros of healthcare costs are paid by the insurance company. After that, $\delta = 0.5$, so the insured person pays 50% of her expenditures above Δ out-of-pocket. The maximum out-of-pocket is 400 euros ($D = 400$). This maximum is reached when an individual's healthcare expenditures are $Z = \Delta + D/\delta = 1,000$ euros or more. Expenditure above this is paid by the insurer. As the next corollary shows, we can use Lemma 1 for analytical expressions for the $EOOP$ for all cost-sharing schemes in the family in Equation (5). For cost-sharing schemes outside this family, one can simulate the distributions for the out-of-pocket payments using stochastic integration.

The generalized expression for the OOP is:

$$OOP = \int_{\Delta}^{+\infty} \delta(Z - \Delta) f_{\chi}(Z) dZ \tag{6}$$

Corollary 1 *The generalized version of $OOP(\mu, \sigma)$ can be written as*[13]

$$\begin{aligned} OOP(\mu,\sigma) = \delta \Bigg[& e^{\mu+\frac{\sigma^2}{2}} \left(N\left(\frac{\ln(\Delta + D/\delta) - \mu - \sigma^2}{\sigma}\right) - N\left(\frac{\ln(\Delta) - \mu - \sigma^2}{\sigma}\right)\right) \\ & -\Delta\left(N\left(\frac{\ln(\Delta + D/\delta) - \mu}{\sigma}\right) - N\left(\frac{\ln(\Delta) - \mu}{\sigma}\right)\right)\Bigg] \\ & + \left(1 - N\left(\frac{\ln(\Delta + D/\delta) - \mu}{\sigma}\right)\right) D \end{aligned}$$

where D denotes the maximum OOP, Δ the shift in the starting point, and δ the co-insurance

[13] See Appendix 9.A for proof of Corollary 1.

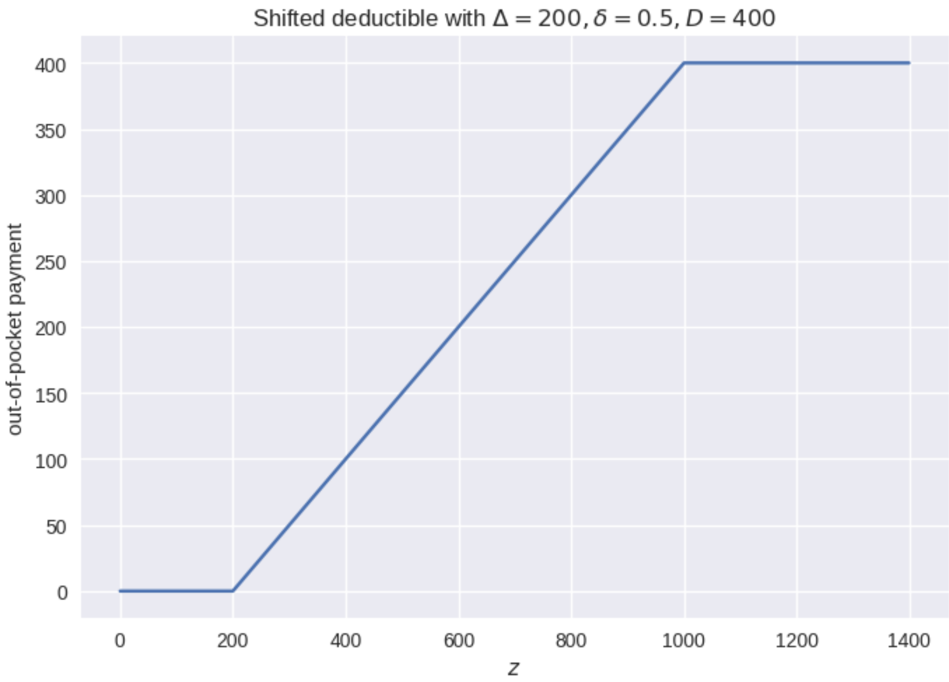

Figure 3: Out-of-pocket payments with a shifted deductible as a function of expenditure z.

rate.

2.4 Accepting a treatment

In our model, a change in the type or level of cost-sharing affects healthcare expenditure through a change in the price, $EOOP$, of an offered treatment and thus the probability that this treatment is rejected.[14] To model this we assume that the value of y treatments is distributed with a cumulative distribution function F. For given $EOOP$, an agent rejects treatments with utility below $EOOP$. The probability that this happens is given by $F(EOOP)$. This function captures our utility structure: $F(x)$ denotes the probability that treatment utility is less than x.

In the model that we estimate below, the probability that an offered y treatment is rejected is given by:

$$F(EOOP) = 1 - \zeta e^{-\nu EOOP} \tag{7}$$

where $\nu > 0$ and $\zeta \in \langle 0, 1]$. The parameters capture the price responsiveness of healthcare to changes in cost-sharing. The parameters in this specification of F can be interpreted as follows: we assume that the hazard rate is constant, $\nu = f(x)/(1 - F(x))$, and ζ denotes the probability

[14]Modeling the relationship between $EOOP$ and the probability that a treatment is rejected (F), and not, for example, the relationship between the deductible size and healthcare spending, allows us to model the effect of many types and levels of cost-sharing schemes not observed in the data. For a co-insurance rate of 25%, for example, we simply compute $EOOP$ given the co-insurance rate and the distributions of x and y, and relate $EOOP$ to F.

that a free treatment is accepted: $1 - F(0) = \zeta$. That is, $\zeta < 1$ indicates that there is disutility associated with treatment which can exceed treatment value. This captures travel to and waiting time at a provider or side effects of a treatment. As $EOOP$ goes to plus infinity, the treatment is rejected with probability 1: no treatment generates infinite utility.

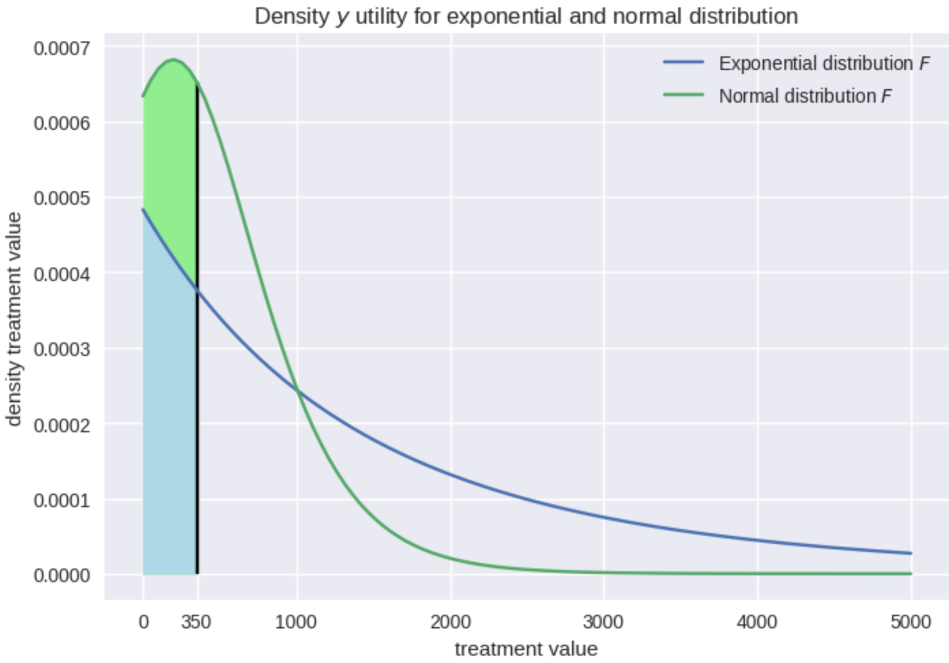

Figure 4: Two distributions of y treatment value

This is illustrated for both the exponential distribution and a normal distribution (which we use in a robustness analysis) in Figure 4.[15] The shaded blue area denotes the additional probability that a y treatment is rejected when $EOOP = 350$ instead of $EOOP = 0$. In other words, $F(350)$ denotes the mass of y treatments with value between 0 and 350. The sum of the green and blue area below the green curve denotes the same mass for a normal distribution of treatment value.

The function F captures ex post moral hazard: the probability of accepting a treatment. We do not model ex ante moral hazard: if healthcare is expensive, people are more careful with their health to prevent falling ill. This would lead to an effect of $EOOP$ on the probability of needing treatment ψ_x, ψ_y.

Summarizing, we model the distribution of total healthcare expenditures per capita per year as the sum of exogenous and endogenous treatments. A change in cost-sharing changes the effective price ($EOOP$) of an offered endogenous treatment. As such, it changes the probability that this treatment is accepted and the distribution of healthcare expenditure, z.

[15] In fact, the distributions drawn are the average densities for the F distributions (averaged over the posterior distributions of their parameters) for 30 year old females. We come back to this below.

3 Data and setting

We first discuss the institutional setting of healthcare in the Netherlands in the period 2008 to 2013. Then we provide details on our data.

3.1 Institutional Setting

This paper focuses on demand-side cost-sharing in curative healthcare in the Netherlands. Dutch curative healthcare comprises hospital care, general practitioner care, physiotherapy, mental healthcare, et cetera.[16] Long-term care and social care in the Netherlands are organized differently from curative healthcare, and therefore outside the scope of this paper. In this entire paper when we write healthcare, we refer to curative healthcare.

Curative healthcare is organized at the national level by regulated competition (Van de Ven and Schut, 2008). Health insurers negotiate and contract with healthcare providers, on behalf of their clients. Health insurance is mandatory and each individual aged 18 or over must purchase health insurance from one of the health insurers. Everyone below 18 years old is automatically insured and they do not pay a health insurance premium nor do they face cost-sharing. The government regulates the market to ensure access to healthcare and protect risk solidarity. For example, the government has set up regulation that makes sure that health insurers cannot refuse an individual from buying their health insurance. There is an elaborate risk adjustment scheme to compensate health insurers for healthcare costs of high risk individuals. The government also sets the coverage of the mandatory basic benefit package and the (minimum) level of cost-sharing, which is the same for everyone. There have been changes in the basic benefit package in the period of this study. These changes, and some other policy changes, are summarised by Remmerswaal et al. (2019a).

Healthcare costs are financed from three sources. The first source is the premium that each inhabitant above the age of 18 pays for his or her health insurance. Annual premiums are between 1,000 to 1,250 euros. There is also an income dependent subsidy for persons with a low income. Dutch inhabitants also pay for healthcare through taxes. The size of the payment can vary per person and depends on income. The last source is demand-side cost-sharing. Every person aged 18 or over faces a deductible which kicks in directly (the starting point is zero). This deductible was introduced in 2008 and since then, the government has increased its size annually (see Table 2).

Table 2: Deductibles in the Netherlands for 2008 to 2013

Year	2008	2009	2010	2011	2012	2013
Mandatory deductible (€)	150	155	165	170	220	350

[16]For an exhaustive list, see Appendix 9.C.

The deductibles in Table 2 are *mandatory* deductibles. It is possible to increase the size of this mandatory deductible by choosing a so-called voluntary deductible. Each year, Dutch inhabitants (aged 18 or over) can choose a voluntary deductible of 100, 200, 300, 400, or 500 euros, in return for a discount on their premium. For example, if someone chose the maximum voluntary deductible in 2008, then he faced a total deductible of 650 euros. A voluntary deductible is chosen by only (roughly) 10% of the insured population. The size of the premium discount is set by health insurers and the average discount for a 100 euro voluntary deductible in 2013 was 45 euros and 230 euros for a 500 euro voluntary deductible.

Some health services in the basic benefit package are exempted from cost-sharing. These services are primary care, general practitioner care, maternal care and obstetric care. These health services cover a small fraction, roughly 8%, of total costs in curative healthcare. For health services such as hospital care, physiotherapy, and pharmaceutical care, cost-sharing does apply.

Health insurers offer supplementary insurance on top of health insurance for care in the basic benefit package. Such supplementary insurance policies cover other health services than the basic benefit package, such as contact lenses and glasses, alternative medicine, extra dental checks, and cosmetic surgery. Supplementary insurance is therefore an addition to regular insurance, not a substitute. Supplementary insurance is optional and can be bought from a different insurer than insurance for the basic benefit package.[17] Importantly and unlike other countries, it is not possible to cover mandatory deductibles in the Netherlands through supplementary insurance nor to get faster access to treatment covered in basic insurance. In this sense, the basic and supplementary markets are orthogonal. In this paper, we only study the basic insurance market and not the supplementary insurance market.

3.2 Data

Proprietary healthcare claim data are used to estimate this model. The data include all claims of all, approximately 17 million, Dutch inhabitants for years 2006 to 2013. The data have been collected by Dutch health insurers and assembled by Vektis. The data have been pseudonymized, are not publicly available, and do not suffer from underreporting of healthcare claims.[18] The same data and a similar cleaning procedure were used in Remmerswaal et al. (2019a) and Remmerswaal et al. (2019b). In this paper, we therefore repeat the cleaning procedure (see Appendix 9.B) and data description. We exclude data for years 2006 and 2007, because another cost-sharing scheme, a no-claims rebate, was in place. In those years, persons with low

[17] Over 85% of the Dutch population bought supplementary health insurance in our data period (Dutch Healthcare Authority, 2014).

[18] Healthcare providers are only compensated for treatments which have been reported to the patient's health insurer. Healthcare providers send their bills to the insurer electronically, who subsequently will bill the patient (if the deductible is not exhausted).

healthcare consumption would get a bonus at the end of a year, instead of paying a deductible. Remmerswaal et al. (2019a) show the rebate has a smaller effect on healthcare expenditure than a deductible. We omit 2006 and 2007 and only use the years a deductible was in place, to simplify the model. After cleaning and excluding the years 2006 and 2007, the train data comprise over 58 million observations.[19]

Total healthcare expenditures of each Dutch inhabitant is one of the main variables in the data. This expenditure can be separated into 21 healthcare categories, such as general practitioner care, maternity care, hospital care, and mental care. As we are interested in the effect of cost-sharing, our expenditure variable only includes cost categories that fall under the deductible (see Appendix 9.C). All variables in the data are only available at a year level. As a result, we do not know how healthcare expenditures evolve within the year.[20]

Several person characteristics are available, such as gender, age, indicators of chronic use of care, and chronic use of medication, and a person's annual choice of a voluntary deductible. Age is available in years and registered for December 31st in every year.[21] To make sure we have enough observations per gender-age category, we pool everyone older than 90 years in age category 91. We use DCG to abbreviate diagnosis cost group ('diagnosekostengroep'); this is an indicator for chronic illness and high healthcare costs in previous years. Similarly, PCG is an abbreviation of pharmaceutical cost group ('farmaciekostengroep'), which indicates chronic use of medication.[22] Lastly, the annual choice of a voluntary deductible ranges between 0 (no voluntary deductible was chosen) to 500 euros (the maximum voluntary deductible).

3.3 Sample and selection

The model is estimated on a baseline sample. This baseline sample is similar to Remmerswaal et al. (2019a) and omits specific observations and cost categories from the data. First, all persons with any mental healthcare expenditures are excluded, because additional co-payments for mental healthcare were introduced in 2012. As this interacts with the deductible (before and after 2012), it complicates the derivation of the expected out-of-pocket expenditure. Furthermore, dental costs are excluded because the coverage of dental care changed between 2008 and 2010 complicating comparisons across years. As dental costs are relatively low, omitting these costs has little effect. Persons who chose a voluntary deductible in the data are also excluded. If they chose the voluntary deductible at least once, we exclude them in all years, to control for potential selection effects of the voluntary deductible.

[19] See Section 3.5 below to read more about the train data.

[20] Klein et al. (2018) use (other) Dutch data to analyze the dynamics of expenditures within a year.

[21] An individual born on December 1st in 1963 is classified in our data as 50 years old in 2013, even if he or she was 49 years old for 11 months that year.

[22] DCG and PCG are variables from the Dutch risk adjustment system, which aim to identify chronic disorders that are correlated with high healthcare expenditures.

This last selection criterium may raise concerns about selection bias in our simulations. One could argue that a change in demand-side cost-sharing, could change people's choice of voluntary deductible and invalidate our conclusions. We believe this is not the case however, for the following reasons.

First, the goal of our model is to predict healthcare expenditure for the people who satisfy our selection criteria; this is the majority of the Dutch population. To predict expenditure for people with, say, a DCG or with a voluntary deductible – and how the chosen voluntary deductible changes with, say, a deductible shift – is left for future research.

Second, we omit citizens from our data who chose a voluntary deductible in *any* year in our data.[23] Hence, people in our sample did not choose a voluntary deductible over the deductible range in Table 2. As long as our simulations lead to out-of-pocket payments consistent with this range, we do not expect these people to now choose a voluntary deductible.

Moreover, our main finding is that a shifted deductible reduces both total expenditure and out-of-pocket expenditure. If this would induce people to take more risk (as out-of-pocket risk is reduced by the shift) in the form of a voluntary deductible, this would strengthen our result that a shift reduces total expenditure. Indeed, one would expect the voluntary deductible to reduce moral hazard further. Next, revealed preference suggests that people are better off by taking up this voluntary deductible if they do so. Hence, although out-of-pocket expenditure may increase due to the voluntary deductible, this must be utility enhancing as otherwise people would not do this. The same reasoning implies that a shifted deductible would not reduce the voluntary deductible uptake for people with a voluntary deductible. That could potentially be problematic as it would increase moral hazard and invalidate our result that a shifted deductible reduces total healthcare expenditure.

Since basic health insurance is mandatory in the Netherlands, we do not (need to) model the decision to buy insurance. Coverage of basic insurance is set by the government and does not vary (much) between insurers. In addition, we focus on the majority of people who do not opt for a voluntary deductible over the relevant range. This allows us to focus on the effect of cost-sharing on the decision to accept a treatment or not. That is, we do not need to model risk aversion. Furthermore, our main policy implication is that a deductible shift reduces both total expenditure and out-of-pocket risk. This is desirable from a utility point of view, irrespective of the degree of risk aversion. This allows our model to be simpler than some of the moral hazard models discussed in the introduction of this paper.

The only difference between the baseline sample in this paper compared to Remmerswaal et al. (2019a) is that we also exclude persons who are chronically ill or chronic users of medication (and who have been labeled with a DCG of PCG in the data). The share of persons with

[23]See Remmerswaal et al. (2019b) for an analysis of selection effects caused by the voluntary deductible.

Table 3: Summary of the sample selection in the train data

	Women	Men
Age (mean)	34.63	33.72
Number of observations	16,367,197	16,117,158
Fraction of positive expenditures	0.81	0.69
Expenditure (mean)	750.53	571.48
Expenditure (std. dev.)	2,667.00	2,963.61
Log expenditure (mean)	4.42	3.58
Log expenditure (std. dev.)	2.62	2.80

label DCG and/or PCG is relatively small and their distribution of healthcare expenditure is very different from people without it. This would complicate identification because we use the distributions of costs per gender-age category as an expectation for the people in the category. Furthermore, people labeled with DCG and/or PCG are unlikely to be affected at the margin by the deductible as their healthcare expenditures are well above the deductible range in our data. If we would use separate gender-age categories for people labeled with DCG and/or PCG then the number of observations in those categories would be too low to confidently identify the model parameters.

All in all, by using the baseline sample, we lose about 40% of the data with the selection steps described above. However, the observations that we lose, are not directly relevant for our research question, because the groups are rather inelastic to the changes in the mandatory deductible. For the DCG and PCG categories, their expected expenditures are (far) above the highest deductible in our data and that we consider in the simulations. People with a voluntary deductible can undo the changes in the mandatory deductible in our data. An increase in the mandatory deductible of, say, a 100 euros can be compensated by a reduction of 100 euros in the voluntary deductible. In the robustness analyses we come back to our sample selection.

To summarize, our main dependent variable of healthcare expenditure covers costs of health services for which the deductible applies except dental costs. We focus on the behavioral response of individuals without a voluntary deductible who are not chronically ill and do not use mental health services.

3.4 Descriptive statistics

Table 3 summarizes the data of the baseline sample for women and men separately. The average age is around 34 for both men and women. There are slightly more women in the sample and they have higher healthcare expenditures: the mean healthcare expenditures are 750 euros for women compared to 571 euros for men. The average healthcare expenditures are low, which indicates that the subsample is relatively healthy. About 80% of the women in our baseline sample has some healthcare expenditure, compared to 70% for men.

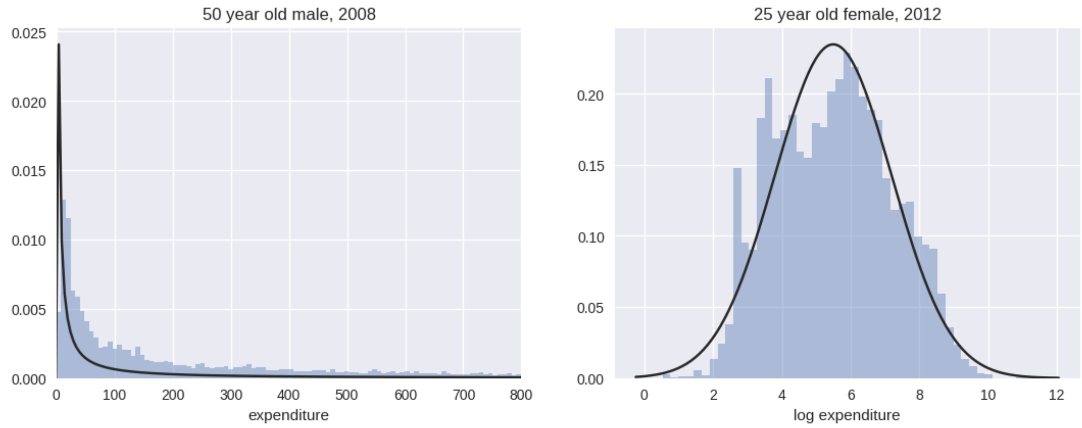

Figure 5: Two illustrative distributions for positive healthcare costs

Table 3 already indicates the skewness of healthcare expenditure in the data. Figure 5 illustrates this further. The figure shows how well a lognormal distribution fits healthcare expenditure (conditional on being positive) in the data. The histogram denotes the raw data (in levels on the left and in logs on the right). The plotted line denotes an estimated lognormal distribution (for just this distribution; not our estimated model). The fit is not unreasonable but not perfect either. It seems to capture the skewness of the distribution of healthcare expenditure well. In fact, the right panel suggests that it is better approximated by a mixture of normal distributions; which is what we estimate.

3.5 Train, validation and test data

In this study, we follow the machine learning custom to divide the data into a train, validation and test dataset (McElreath, 2020). The train dataset is used to estimate the parameters of the model. With the model and the estimated coefficients of these parameters, outcomes are predicted, which are compared to the validation data. If the fit is not good yet, we can go back to the model, make some adjustments, and repeat this procedure. To assess the final fit of the model, the predictions are compared to the test data. This last step we will do after the paper is finalized for publication. The current version of the paper only uses the train and validation sets.

By splitting the data into these datasets we prevent overfitting of the model. For this paper, we applied stratified sampling to make each dataset representative in terms of age, gender and years. The test and validation datasets each comprise 20% of the total data, and the train dataset the remaining 60%.

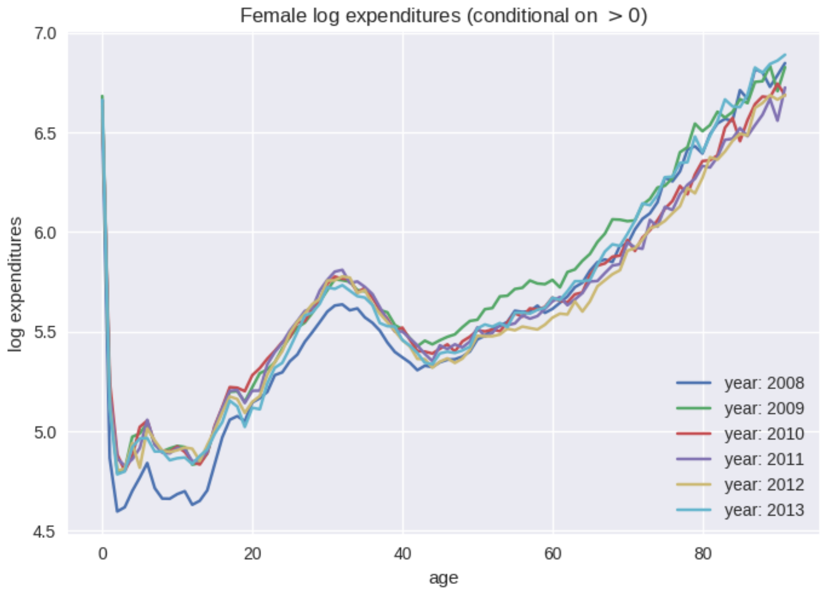

Figure 6: Log healthcare expenditure for women conditional on being positive

4 Econometric specification

This section explains how we parameterize and identify the model. Appendices 9.D and 9.E provide a more detailed overview of the specification of the model and the choice of priors.

4.1 Parameterization

As mentioned, the distribution of (log) total healthcare expenditure per capita per year, z, is modeled as a mixture distribution with four components. The distribution of expenditure by age is very different for men and women (see Figure 1), which is why we estimate the model separately for men and women. Our unit of observation is this expenditure distribution per gender-age-year and not, for example, healthcare expenditure of one individual.

We expect age to have an effect on components x and y, and therefore we model paramaters $\mu_x, \mu_y, \psi_x, \psi_y$ as Gaussian processes (GPs) with age. To illustrate this choice, consider Figure 6 which shows female log healthcare expenditures (conditional on being positive) across ages and for different years. The graph reveals a clear and stable age pattern across the years. The GP captures this pattern by assuming that expenditures are more similar for, say, 20, 21, and 22 year olds, than for 20 and 50 year olds. We assume that the covariance decreases with the square of the age difference.[24]

[24] In particular, the covariance between, say, $\mu_{x,a}, \mu_{x,a'}$ for two different ages a, a' is – up to a constant – given by $e^{-0.5(a-a')^2}$ (Rasmussen and Williams, 2005).

To capture the relatively small annual changes in basic insurance coverage we allow for year fixed effects in μ_x and μ_y. We model these changes as year fixed effects because they can be different from one year to the next and do not necessarily follow a coherent pattern (as they would when modeled as GP).

For ψ_x and ψ_y we assume that the probability of being offered a treatment varies with age, but not by year.[25] That is, we assume that the probability of being offered a treatment is driven by the probability of falling ill, which –we assume– does not vary (much) over time. The probability that you accept a y treatment varies across years with the deductible and the distributions of x and y expenditures.

The standard deviation of expenditure z does not show a clear pattern across age.[26] Hence, we model the standard deviations σ_x, σ_y as age fixed effects, not as a GP.

As mentioned in Section 3.1, cost-sharing in the Netherlands kicks in when a person turns 18. In our data, age is given in full years, and therefore we cannot distinguish a person who turns 18 in January from a person who turns 18 in December of the same year. Both are denoted as 18 in our data. However, the former will face cost-sharing the entire year, whereas the latter will not face any cost-sharing at all. We include a parameter $\alpha \in [0, 1]$ (which varies with gender) which weighs the effect of cost-sharing for these 18 year olds. We see α as the probability that a person faces a deductible when deciding on the y treatment. If birthdays are uniformly distributed over a year, we expect α to be around 0.5.

Finally, the parameters of the function F (see Equation (7)), ζ and ν, feature age fixed effects. That is, the distribution of treatment values is allowed to vary by age and gender and thus the price responsiveness can also vary by age and gender.

4.2 Identification

To identify the parameters described above, we use data on healthcare expenditures at the individual level for years 2008 to 2013 (see Section 3.2). We start from the idea that for a given gender-age-year combination the distribution of (log) expenditures can be approximated by a mixture of four distributions: expenditure equals 0, an x distribution, a y distribution and an $(x + y)$ distribution. Such mixture models are well known (see, for instance, McElreath, 2020; Gelman et al., 2013).

What we add to this is that the weight on the y distribution varies with $EOOP$. It is this

[25] There is one exception to this: ψ_x for women older than 21 years changed substantially in 2011. Before 2011, contraceptives were covered for all women aged 18 and over. However, since 2011 contraceptives are no longer covered by the basic insurance package for women older than 21 years. As a result, all (expenditures on) contraceptives for this group dropped out of our data. To be clear, this is not a substitution effect, but these purchases are no longer recorded in our data. We create a dummy which equals 1 for women above 21 years old from 2011 onward (and 0 otherwise). We allow the coefficient of this dummy to decay with age as older women are less likely to use contraceptives.

[26] As illustrated in Figure 12, the standard deviation for men randomly fluctuates around approximately 2.7.

variation in $EOOP$ that allows us to identify the parameters ζ, ν of the probability F that a y treatment is rejected. We have three main sources of variation for this identification.

The first source of exogenous (for a gender-age category) variation is the change in the size of the mandatory deductible over time. As presented in Table 2, this deductible was 150 euros in 2008, and it increased annually up to 350 euros in 2013. This change affects the $EOOP$ and thus the probability that y treatments are accepted. Thereby, the change in the deductible size affects the probability of having positive expenditures and the distribution of expenditures z.

Second, there is (crosssectional) variation in the healthcare distributions among different gender-age categories even when they face the same (mandatory) deductible per year. For example, the healthcare expenditure distribution of an 80 year old man differs greatly from the healthcare expenditure distribution of a 25 year old man. The former has higher expected expenditures than the latter (see Figure 1) and we expect him to have higher x expenditures as well. This is indeed the case as illustrated in Figure 7: the (posterior) probability that ψ_x for 25 year olds exceeds ψ_x for 80 year olds is basically zero. This makes x treatments for 80 year olds more likely and hence $EOOP$ lower.

The third source of exogenous variation for identification is the age threshold for cost-sharing: only individuals aged 18 and over face cost-sharing, whereas persons below 18 years old do not face any cost-sharing. An advantage of this age threshold is that it provides more variation in the deductible size, namely a deductible of zero. The last two sources of crosssectional variation allow us to separate year fixed effects from yearly changes in the deductible.

In particular, consider an age $a > 18$,[27] gender g and year y combination. Under our model, the expenditure distribution for this combination allows us to identify ψ_x and $\psi_y(1 - F)$ which determine the weights on the zero, the x, the y and $x + y$ distributions and the mean and standard deviations of the x and y distributions. The variation in the deductible over time allows us to separately identify ψ_y and the parameters ζ, ν of F.

4.3 Prior distributions

Bayesian estimation methods start with prior distributions for the parameters which are then updated –when confronted with the data– into posterior distributions. Specifying priors is not completely routine. The priors should not exclude plausible parameter values. On the other hand, they should not put (much) weight on implausible outcomes either (McElreath, 2020). Since we have data on the whole Dutch population there are enough observations to determine the posterior distributions in a fairly robust way, even in our train data. The exact choice of priors can be found in Appendix 9.E. In the appendix we also argue that there are some bounds

[27] Note that for people below 18, we have $EOOP = 0$. Hence for this group we cannot separately identify ψ_y, ζ, ν. But we do not need to identify these separately as this group is not used in our simulations below.

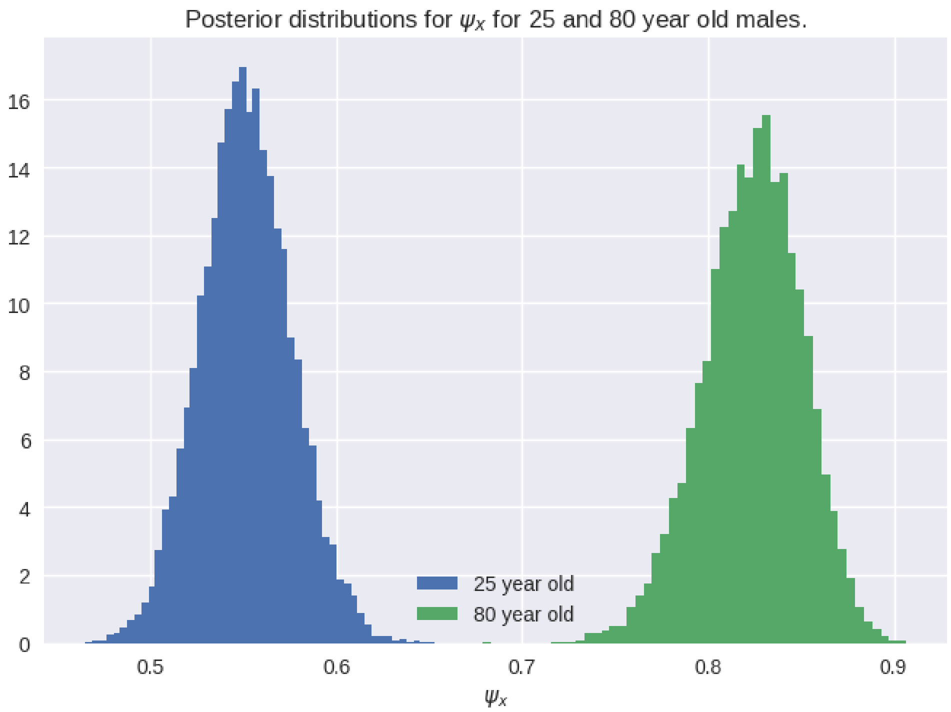

Figure 7: Posterior distributions of ψ_x for 25 and 80 year old men

on what healthcare expenditure per head can be. Then we simulate expenditures directly from the priors to see whether the priors satisfy these bounds. We also illustrate that the main results are robust to different prior choices.

5 Estimation

In this section we explain our estimation methodology and present a summary of estimated parameters. Since the main goal of the paper is to simulate outcomes for a number of cost-sharing schemes, we present these outcomes together with the uncertainty that surrounds them. Therefore, we put more emphasis on the posterior distributions than papers like Einav et al. (2013) and Geweke et al. (2003). For each parameter we draw 10,000 samples from the posterior. With these samples we calculate average effects and their uncertainty. This is in contrast to maximum likelihood estimators where (only) the "most likely" parameters (mode of the distribution) are used. Using the posterior, we also examine the fit of our predictions with the validation data.

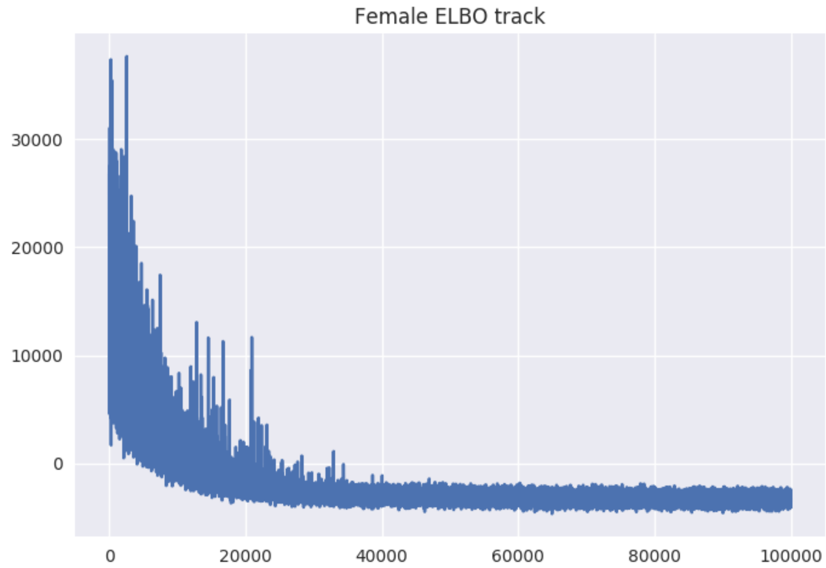

Figure 8: ELBO model estimation for women

5.1 Estimation methodology

The model is estimated using Bayesian methods with PyMC3 in Python (Salvatier et al., 2016). Since standard Bayesian Markov Chain Monte Carlo (MCMC) methods like Metropolis and NUTS do not scale well with our large dataset, we use a variational inference approach (ADVI, or automatic differentiation variational inference) with minibatches. ADVI is especially suitable for more complex models, such as our mixture model, which are estimated on large datasets (Kucukelbir et al., 2017). Contrary to Metropolis and NUTS estimators, the ADVI estimator approximates the posterior with well known distributions which speeds up estimation considerably. MCMC methods are computationally intensive but can provide (asymptotically) exact samples from the target density; for ADVI we do not have such an asymptotic result. Blei et al. (2017) argue that although variational approaches tend to underestimate the variance of the posterior densities, for mixture models the results of variational inference may be better than MCMC.

Figure 8 shows the ELBO (evidence lower bound) plot for estimating the model for women.[28] The plot is skewed and flat on the right hand size, which implies convergence of the ADVI algorithm.

[28]To keep this paper short, we show in most cases graphs for one gender only. For the other gender, graphs are presented in Appendix 9.F.

5.2 Parameter estimates

Here we summarize the posterior distributions of the directly relevant parameters for women and men: $\mu_{x,y}$ and $\sigma_{x,y}$, ψ_x and ψ_y, α, ν, and ζ.

Table 4 gives a first summary with mean values and standard deviations of the posterior distributions for the main parameters. Note that this is a broad brush description of the posteriors because we have aggregated across samples, ages, and years. We find that μ_x is on average approximately 5.0 for both men and women, while μ_y is lower, below 3.0. The standard deviations of the posterior distributions for $\mu_{x,y}$ are between 0.15 and 0.21, respectively for men and women. The standard deviation of x treatments is around 1.2 for both sexes; for y it is approximately 0.4. The probability of being offered a y treatment is close to 0.5 for both sexes. Women are more likely to be offered x treatments than men: 0.8 vs 0.7 on average.

Recall that α indicates the probability that an 18 year old faces a deductible. We find that on average α equals 0.5 which is consistent with birthdays being approximately uniformly distributed across the year.

The value of ν is small but positive. Recall that ν is multiplied by the $EOOP$, which is maximally 350 euros (the largest deductible size) in our data, to compute F, the probability that a treatment is rejected (see Equation (7)). A small but positive ν is therefore in line with our expectations. The average value for ζ can be interpreted as women and men would accept 60% of y treatments offered to them if they were free ($EOOP = 0$).

Although Table 4 gives an indication of the size of the parameters, the beauty of Bayesian analysis is to work with the (posterior) *distributions* of parameters. For example, Figure 9 gives the distribution of ν for women. The expectation of this distribution is around 0.001. When determining the fit of the estimations and the effects of various simulated cost-sharing schemes below, we draw values for ν from this distribution.[29] The left figure shows that 20 year old

[29] We do not present plots of each posterior distribution here, as that would make the paper too long. Our code repository contains the posterior samples and hence the interested reader can plot the distribution for each

Table 4: Summary of posterior distributions

Variable	Women		Men	
	Mean	Std. dev.	Mean	Std. dev.
μ_x	4.830	0.210	4.827	0.143
μ_y	2.625	0.186	2.813	0.211
σ_x	1.111	0.133	1.240	0.157
σ_y	0.411	0.099	0.425	0.104
ψ_y	0.537	0.070	0.474	0.151
ψ_x	0.785	0.072	0.675	0.097
α	0.502	0.157	0.509	0.157
ν	0.001	0.001	0.001	0.001
ζ	0.653	0.132	0.591	0.146

women tend to be more elastic (higher ν) with respect to the $EOOP$ than 70 year old women. As the figure on the right shows, this translates into a higher probability that a treatment is rejected by 20 year old females than 70 year olds for a given value of $EOOP$. Put differently, 70 year old women tend to be offered more valuable y treatments. The effect that 70 and 20 year old women face different cost distributions is captured by $EOOP$ and does not affect the parameters ν and ζ of the utility distribution F. Furthermore, to give an idea of the uncertainty surrounding F for both age categories we also plot the 75th percentile of F. To illustrate the interpretation of this 75th percentile, consider an $EOOP = 500$. On average, 20 year old women will reject a treatment with $EOOP = 500$ with a probability of 0.6. We are 75% sure that this rejection probability is less than 0.65, as is illustrated with the dashed blue line at this $EOOP$.

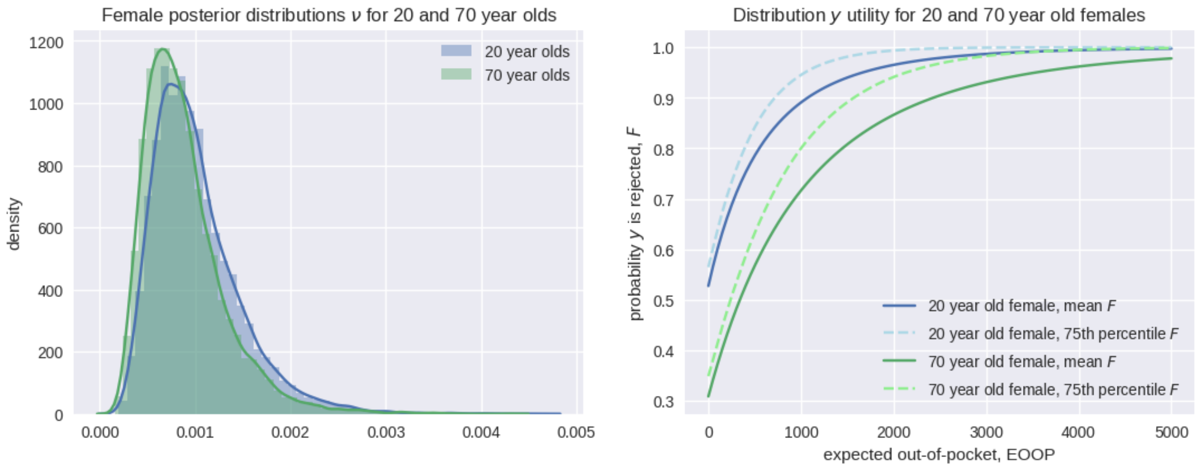

Figure 9: Posterior distributions of ν and F for women

As a final illustration of our estimates, and a first step towards evaluating fit, consider Figure 10. This figure plots the probability of positive expenditures, $1 - (1 - \psi_x)(1 - \psi_y + \psi_y F)$ (see Table 1), for men across age in 2008. Recall that the parameters ψ_x and ψ_y are modeled as GPs. That is, we do not draw values from a distribution for each age, but chains of values across all ages from the GP. The figure illustrates this by showing these draws for ψ_y and ψ_x as thin red lines. Darker colors red, indicate more draws at these values. On top of these draws, the realized fraction of positive expenditures are plotted for each category (age-male combinations) in 2008. Our predicted probability of positive expenditures is fairly close to the realized fractions of positive expenditures for each age category. But we tend to overestimate this probability for men around 80 years old.

parameter.

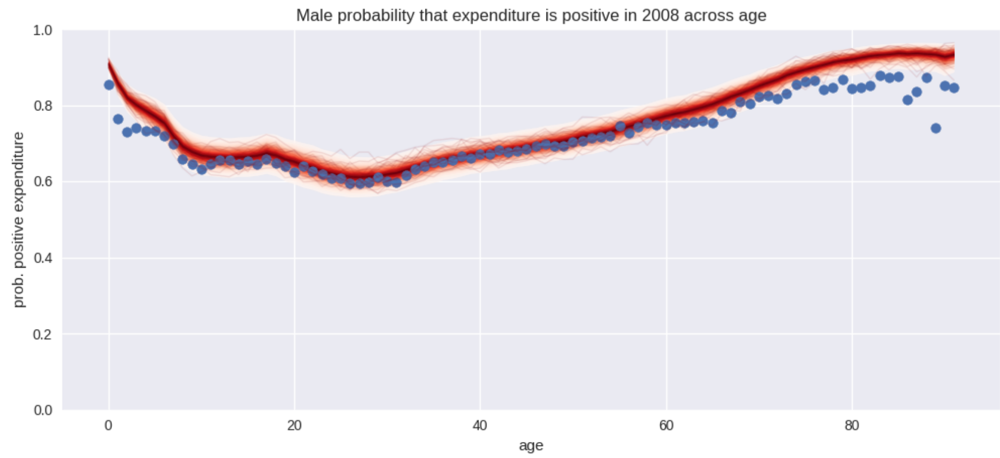

Figure 10: Predicted and realized probabilities of positive expenditures for men across age in 2008

5.3 Model fit

To examine the fit of the model, we first predict healthcare expenditures. For each parameter in the model we draw 10,000 samples from the underlying posterior distribution. For example, we have 10,000 μ_x's for a given year, gender, and age.[30] Next we generate a value with each sample. To illustrate, for x this implies generating a draw from the normal distribution with parameters μ_x and σ_x of that particular posterior sample. These 10,000 outcomes of x incorporate two forms of uncertainty: (i) for given μ_x, σ_x, we draw an x from this normal distribution, (ii) we have 10,000 values for μ_x, σ_x since we are uncertain about the parameters of the x distribution. With ψ_x and ψ_y we draw a value of 0 or 1 from the Bernoulli distribution, where 1 indicates a treatment offer is made (0 no offer is made).

With the samples of $\psi_x, \mu_{x,y}$, $\sigma_{x,y}$ and D in a given year, we generate a value of $EOOP$ (Equation (4)), which then combines with the ζ's and ν's to get values for F (Equation (7)). With these values for F we generate a value of 0 (y treatment is accepted) or 1 (treatment rejected) with a Bernoulli distribution.

In this way, we have 10,000 predicted healthcare expenditures for men and women, 92 age categories and 6 years.

Figure 11 shows a comparison of the predicted and observed mean healthcare expenditures for men in 2008 up to 2013. The solid line denotes average log healthcare expenditures per age category which are predicted with the model and the dots denote the mean log healthcare expenditures of the validation data.[31] The fit is quite good. Even the second moments, as shown

[30] In total, there are 10,000 x 6 years x 2 genders x 92 ages samples of μ_x (and similarly for other parameters of the model).

[31] To be clear, we first take the log of individual healthcare expenditures and then the mean of log expenditures per gender-age-year category.

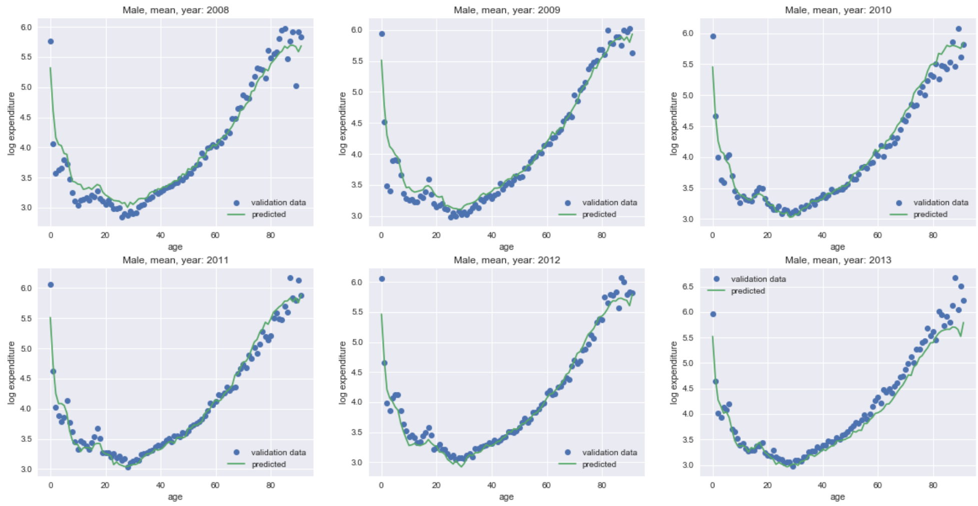

Figure 11: Average predicted vs average validation male log healthcare expenditures for 2008 to 2013

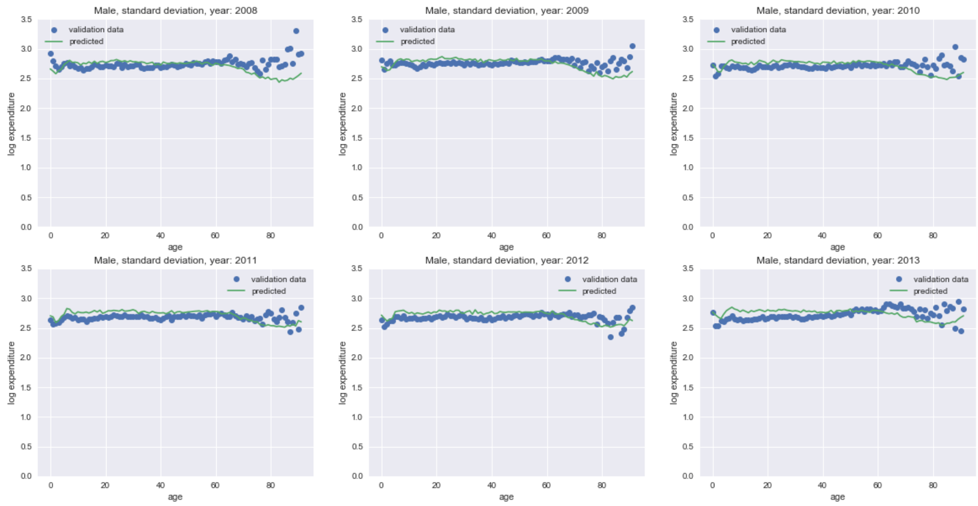

Figure 12: Standard deviation predicted vs validation male log healthcare expenditures for 2008 to 2013

in Figure 12, show a reasonable match.

An illustration of the fit on the distribution is to plot a histogram and a QQ plot of the validation and predicted data. The left panel in Figure 13 compares the distribution of healthcare expenditure for 50 year old women in 2011, conditional on the expenditures being below 3,000 euros.[32] The distributions of the predicted data and the validation data are close. The right

[32]The upper bound on expenditure is needed to keep the figure readable. To illustrate, the maximum expenditures for this group in 2011 is almost 100,000 euros, and more than 95% of the observations are below

panel in the figure shows that we slightly underestimate healthcare expenditures for expenditures below 1,500 euros of 50 year old women in 2011 and overestimate expenditures above 1,500 euros.

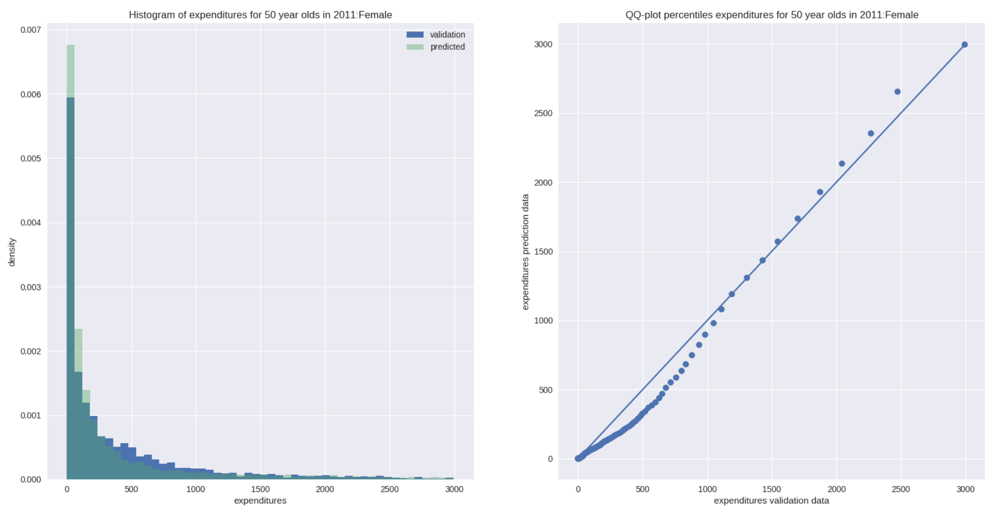

Figure 13: Predicted vs validation female log healthcare expenditures for 50 year olds in 2011

6 Simulations and policy analyses

The purpose of the structural model is to simulate healthcare expenditures for different demand-side cost-sharing schemes in order to compare the effects of these schemes on healthcare expenditure and out-of-pocket payments. In this section, we use the model to compare three cost-sharing schemes: a deductible, a co-insurance scheme, and a shifted deductible.

6.1 Simulations

Our simulations are conducted for the most recent year in our data, 2013, and maintaining the Dutch feature that people under 18 do not pay for healthcare under basic insurance. To simulate healthcare expenditures under a different cost-sharing scheme than a deductible of 350 euros, which was in place in 2013, we use the estimated posterior distributions of x and y for each gender-age category in 2013, and compute $EOOP$ given this new cost-sharing scheme. Recall that the parameters of the model ($\psi_{x,y}, \mu_{x,y}$ etc.) do not change with the cost-sharing scheme; only $EOOP$ is affected. Next, using the specification of F in Equation (7) and the estimated posterior distributions of the parameters ν and ζ, we determine how the change in $EOOP$ affects

3,000.

the probability that a treatment is rejected (F). Lastly, we compute the healthcare expenditures under the new cost-sharing scheme.[33]

We simulate and compare seven maximum out-of-pocket payment levels: 0, 100, 200, 300, 400, 500, and 600 euros, four co-insurance rates: 0.25, 0.5, 0.75, and 1.0, in combination with seven starting points: 0, 100, 200, 300, 400, 500, and 600 euros.[34] Figure 2 shows the results of the simulations for some of the cost-sharing schemes normalized on healthcare expenditures with a 300 euro deductible.

Healthcare expenditure per head decreases as the deductible level increases from 0 to 600 and out-of-pocket expenditures increase. We are not interested in this result as such. Recall that we do not estimate people's degree of risk aversion and hence cannot resolve the trade off between lower expenditure and higher out-of-pocket risk. But it is interesting to observe that these results translate into a deductible elasticity of -0.09, which is in line with Remmerswaal et al. (2019b).[35] Although, we focus on the distribution of expenditures, Remmerswaal et al. (2019b) models individual level expenditure (using individual fixed effects) as a function of the deductible level. Both methods lead to a similar elasticity but our approach here allows us to simulate a variety of cost-sharing schemes.

6.2 Intuition for the outcomes

Figure 14 shows how healthcare expenditures vary for cost-sharing schemes with four different co-insurance rates and seven different starting points, or shifts. All schemes in the figure have the same maximum out-of-pocket of 300 euros and we again normalize healthcare expenditure on the average healthcare expenditure with a 300 euro deductible, which is represented by the point (0,100) on the purple line in the figure. With a maximum out-of-pocket payment of 300 euros, healthcare expenditures are lowest with a shifted deductible which starts at 400 euros and has a 100% co-insurance rate. Among the schemes which start at zero, that is, which are not shifted, a 25% co-insurance rate leads to the biggest reduction of healthcare expenditure.

Figure 14 allows us to illustrate a number of mechanisms in the model. First, with a relatively low deductible of 300 euros, as we have in the Netherlands, quite a few individuals are not at the margin. Therefore, without a shift, it is optimal (in the sense of minimizing expenditure per

[33]See also Appendix 9.D for the full model with the specification of total (log) expenditure z in its underlying components.

[34]In total these are 7 * 4 * 7 = 196 different cost-sharing schemes.

[35]We calculate the deductible elasticity here as follows:

$$\begin{aligned} \varepsilon &= \frac{\Delta y}{\Delta D}\frac{\bar{D}}{\bar{y}} \\ &= \frac{\frac{y_{600}}{y_{300}} - \frac{y_0}{y_{300}}}{\frac{600-0}{\frac{600+0}{2}}} = \frac{0.90 - 1.08}{2} = -0.09. \end{aligned}$$

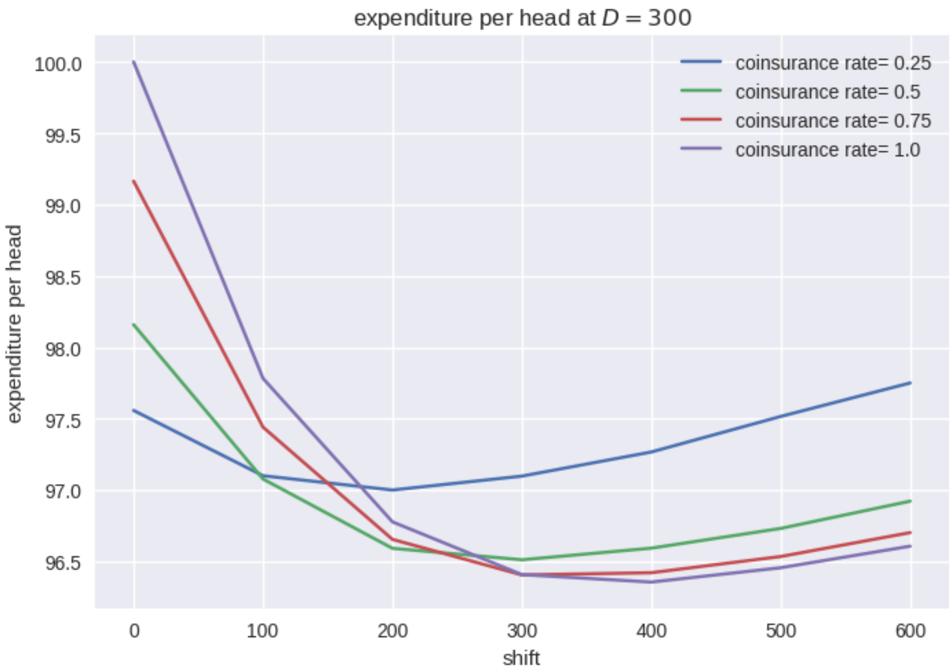

Figure 14: Healthcare expenditure per head for $D = 300$ and different levels for the shift and co-insurance rate relative to expenditures with a deductible equal to 300

head) to have a low co-insurance rate of 25%. This blunts the effect of the deductible as you only pay 25 cents out-of-pocket for every euro healthcare costs you make, but lengthens the range over which individuals face cost-sharing from $[0, 300]$ to $[0, 1200]$ euros. A 25% co-insurance rate with a maximum out-of-pocket payment of 300 euros reduces healthcare expenditure per head by 2.5% compared to $D = 300$, a 300 euro deductible. Moreover, as shown in the right panel of Figure 2, it also reduces the out-of-pocket payment on average by at least 30%.

When we introduce a shift in the cost-sharing scheme and move to the right side in Figure 14, we see that the optimal co-insurance rate increases. With a shift of 100 euros, a scheme with a co-insurance rate of 0.25 and 0.5 lead to similar expenditures per head. For larger shifts, between 250 and 300 euros, a co-insurance rate of 0.75 is optimal, and beyond 300 euros it is optimal to have a rate equal to 1.0. In other words, shifting the starting point puts more people at the margin. Reducing the co-insurance rate below 100% to lengthen the expenditure range over which there is cost-sharing, is then no longer optimal.

Apparently, in the Netherlands in 2013, $D = 300$ together with a shift of 400 euros captures enough people at their margin that reducing the co-insurance rate is no longer useful. Not only does this minimize healthcare expenditure per head for $D = 300$, it also minimizes the out-of-pocket payment for everyone as Figure 2 shows.

At first sight this result may seem surprising as we have shown that many people in our sample have zero or relatively low expenditure. For this group healthcare becomes (almost) free with a 400 euro shift. How can this be optimal? There are two reasons for this. First, relatively many people have zero expenditures because they are not offered any treatments. For this group, the shift has no effect. Part of the people with low expenditures has exogenous (x) treatments; for them the shift has no effect either. Second, our model is additive in log (x and y) expenditures. This captures the intuitive idea that with high exogenous expenditure, there are more possibilities for (high) endogenous expenditure as well. If you are healthy, it will be more difficult to get a doctor to treat you. Once you get health problems and fall ill, you need diagnostics and treatment, and other additional treatments (like extra X-rays) will become an option. Hence, to minimize expenditure, we need people who already feature some (exogenous) expenditure to be at their margin. This is what a deductible shift helps to establish.

The result that a 400 euro shifted deductible reduces expenditure per head and out-of-pocket payments can be different for other countries with different expenditure distributions, for example due to a different age profile of the population. The best known study on the effects of cost-sharing on healthcare expenditures is Newhouse and the Insurance Experiment Group (1993). They do not analyze a shifted deductible, but they do consider co-insurance schemes, including a rate of 25%. Whereas we find that co-insurance is more effective in reducing healthcare expenditure compared to a 100% rate, their simulation model shows the opposite.[36] There can be a number of reasons for this. For example the US in the 70s had a completely different institutional setting compared to the Netherlands in 2013. But the general principle that a shift in the starting point of the deductible increases the optimal co-insurance rate will hold for all countries. In this sense, the shift and a co-insurance rate below one can be thought of as complementary policy instruments.

6.3 Policy uncertainty

Since this is a Bayesian analysis, we can also show the uncertainty surrounding our policy recommendations. This is illustrated with the kernel density plot in Figure 15. Although *on average* (across draws of the posterior distribution) average expenditure is lower with a 400 euro shift than with a 300 euro deductible without shift, this is not true for each draw of parameter values.

The figure shows on the horizontal axis average (across the population) healthcare expenditure with a 300 euro deductible with a 400 euro shift divided by average expenditure with the same deductible but no shift. For this (ratio) statistic we take the density over the 10,000 samples that we have. The cumulative density in the figure shows that there is an approximately 10% probability that this ratio is less than 0.85. In other words, there is a 10% probability that

[36] Although the differences between the effects of the different co-insurance rates are small, especially for small maximum out-of-pocket amounts.

Figure 15: Posterior distribution of average expenditure with a 300 deductible and 400 euro shift divided by average expenditure with a 300 euro deductible and no shift.

expenditure falls by at least 15% after the introduction of a 400 euro shift.

As the figure is steep around 1, for quite some mass in the posterior distribution expenditure is the same (similar) with and without the shift. Finally, the figure suggests that there is 25% probability that expenditure increases due to the 400 euro shift in deductible.

Summarizing, for policy makers the available data suggests that in expectation healthcare expenditure per head falls with the introduction of a 400 euro shift but there is 25% probability that expenditure increases slightly (by at max. 3%).

6.4 Differentiated starting points

Until now, we have explored shifted deductibles with a uniform starting point for everyone. Because we have the distribution of expenditures for each gender-age category, we can characterize the optimal shift and co-insurance rate for each gender-age category. To illustrate this, Figure 16 considers people above the age of 30. As illustrated in Figure 1, average healthcare expenditure increases (more or less) monotonically from this age onward. Hence, to keep many people at the margin, we expect that the optimal shift will increase with age. Doing this for each age separately, for example, for 40 year olds, 41 year olds, and so on, will lead to a noisy picture with an upward trend. However, this is not a serious policy option as it could mean, for example, that the shift is 300 euros when you are 40 years old, it becomes 350 euros for age 41, and then falls again to 320 euros at age 42. This would be too confusing for people to understand. Hence, we consider age-brackets of 10 years: 30 to 40 year olds, 40 to 50 year olds, and so on; within an age bracket both genders face the same deductible shift. Figure 16 plots average expenditure

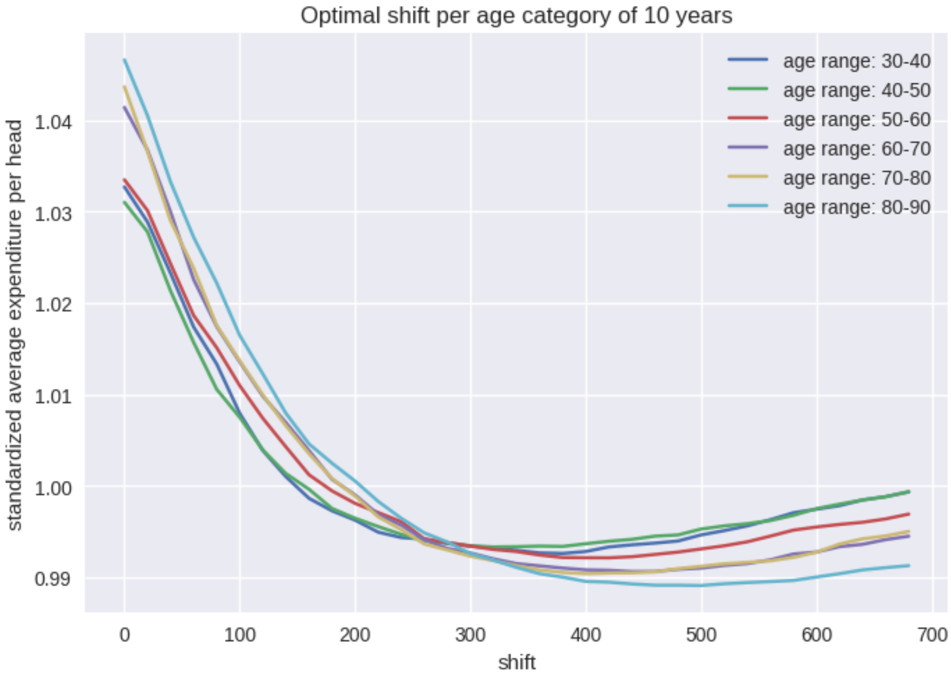

Figure 16: Optimal shift for 10-year age categories

per head for each age class as a function of the shift, standardized by the average expenditure across all shifts for that age group. In this way, the curves are comparable as normally the expenditure for the age group 80-90 would be far higher than for 30-40. The figure shows that the optimal shift indeed increases with the age bracket; varying from 320 euros (for age groups 30-40 and 40-50) to 500 euros (for 80-90). Characterizing the fully optimal system (under some sanity constraints) is beyond the scope of this paper, and we leave this for further research.

6.5 Quality of care

Finally, all simulations until now changed the price of healthcare through $EOOP$. Demand for healthcare is also affected by the quality of care. Governments are investing in the care sector, which leads to a direct (investment) cost but also to higher demand. Analogously, when a government reduces its healthcare budget, quality will fall and expenditures are likely to decrease as well. Because our model is microfounded on patient utility, we can simulate the effect of a, say, 10% increase in quality on healthcare expenditures.

We simulate the effect of a change in the (perceived) quality of care on healthcare spending by changing parameter ν in the function of F (Equation (7)) in the model and assume that cost-sharing does not change.[37] An advantage of our specification of F is that expected quality

[37] Alternatively, we can change the parameter ζ.

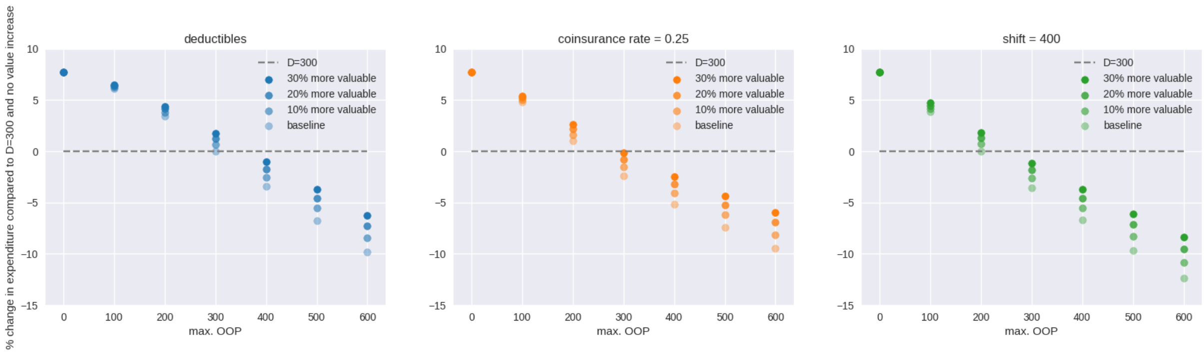

Figure 17: Healthcare expenditures after increasing the value of treatments by 10, 20, and 30%

is simply given by:

$$E(x) = \int_0^{+\infty} x f(x) dx = \frac{\zeta}{\nu} \tag{8}$$

Hence, one way in which we can increase the quality of healthcare is to compare parameter ν_0 with $\nu_1 = \frac{\nu_0}{1+g}$. This means that $E_1(x) = (1+g)E_0(x)$: expected quality has increased with growth rate $g > 0$. Figure 17 shows how healthcare expenditures change, when we increase the quality of healthcare by 10%, 20%, or 30% (the darker the colour, the larger the increase in quality). We increase the quality by lowering the estimated posterior distributions of ν, and thereby lowering the probability that a treatment is rejected, given $EOOP$. The figure shows that increasing the value of healthcare by 10%, 20%, and 30% under a 300 euro deductible, will lead to an 0.7%, 1.3%, and 1.7% increase in expenditure respectively. A 10% increase in the value of treatments does not directly translate in to a 10% increase in healthcare expenditures, because of nonlinearities in the model. We also see that healthcare expenditure is the same for no cost-sharing, regardless of the value increase (because a quality cut-off of zero is not affected by the factor $(1+g)$; this would be different if we increased quality via an increase in ζ), and as the maximum out-of-pocket payment increases, the differences become larger.

7 Robustness analyses

In this section we present analyses to show that our main prediction, a shifted deductible leads to a reduction of healthcare expenditures and out-of-pocket payments compared to a deductible, is robust to our modeling choices. Here we focus on our sample selection criteria, our functional form of F and $EOOP$.

7.1 Sample selection

The baseline sample used to estimate the model excludes individuals who chose a voluntary deductible at least once in our data, chronically ill persons, chronic users of medication (labeled with DCG or PCG) and persons who used mental health services (see Section 3.2).

We argue that excluding these groups from the baseline sample makes sense given the purpose of this paper. The aim of this paper is to model how healthcare expenditure changes under different cost-sharing schemes. As these individuals are rather price inelastic, adding them is like adding 'dead weight' to the results which blunts the differences between the different cost-sharing schemes. We illustrate this point below for people labeled with a PCG and/or a DCG.

Figure 18 shows the distributions of log expenditures conditional on being positive for our baseline sample and the labels PCG and DCG. Clearly the costs for persons labeled with a PCG and DCG are substantially higher than for our baseline sample. Moreover, although zero expenditure happens for a lot of people in our baseline, it happens for less than 1% for the PCG and DCG groups. Hence, estimating one model for both the baseline sample as well as persons labeled with a PCG and/or DCG does not make much sense. Combining the groups would imply combining their distributions of expected healthcare expenditures, but it seems unlikely that people in the baseline sample would expect healthcare expenditures as if they were chronically ill, and vice versa. Estimating the model separately for persons labeled with a PCG and persons labeled with a DCG is more appropriate, because the groups would be more homogeneous. This is possible to do for the PCG group, but for persons labeled with a DCG the number of observations per age-gender category is rather small (it can be close to zero for some gender-age-year combinations). However, as we will show below, expected healthcare expenditure of persons labeled with a PCG is so high, that changes in the maximum out-of-pocket over our range has basically no effect. This will then also be the case for the DCG group which has even higher expenditures.

First, we estimate our model for the PCG group separately. Then, to see the effect on the overall results, we mix the PCG outcomes (distributions of expenditures) with our baseline outcomes using as weights the share of each in the population. Figure 19 shows these (mixed in) results in comparison to our baseline outcomes. For both outcomes, PCG mixed in and the baseline sample, the figure plots percentage change in healthcare expenditure per head for each of the cost-sharing schemes compared to a standard 300 euro deductible (illustrated by a dashed, horizontal line). For each cost-sharing scheme we see that the results with persons with PCGs mixed in are closer to the horizontal line ($D = 300$); that is, changing the maximum out-of-pocket has a smaller effect on expenditures per head. This happens because we mix in a group which is basically inelastic with respect to changes in cost-sharing.

Figure 20 shows the results of the baseline and PCG samples combined. Even though the effects are smaller, we do see that the main results of our analysis hold: the shifted deductible leads to the biggest reduction, compared to a co-insurance scheme and a traditional deductible.

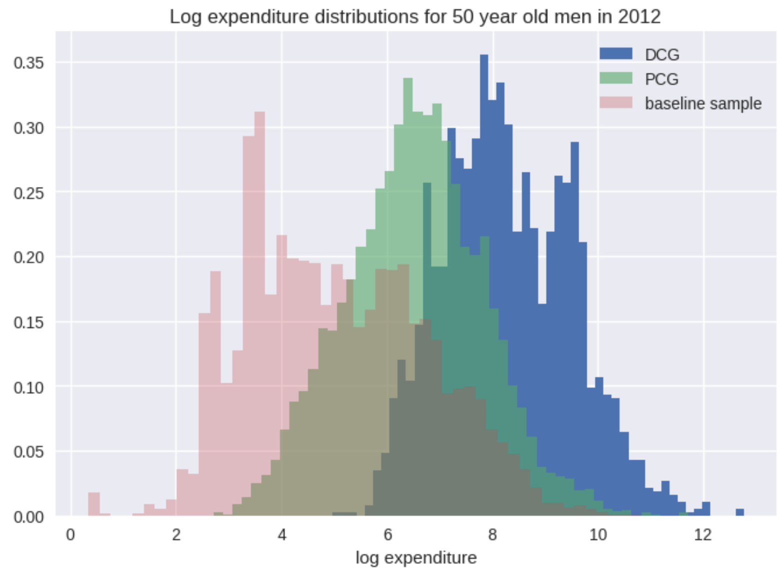

Figure 18: Distributions of log expenditures for the baseline sample, persons with FCG, and DCG

7.2 Specification of F

The probability that a treatment is rejected, F, is modeled with an exponential distribution (see Equation (7) in Section 2). To assess the impact of this specification on the results, we re-estimate the model with a normal distribution for F where ζ denotes the location (mean) and ν the precision (one over the standard deviation).

The results are presented in Figure 21. With a normal distribution for F the results (darker colors) are similar to the results of the baseline model (lighter colors). With a normal distribution expenditure tends to increase less steeply with a decrease in maximum out-of-pocket payment and tends to decrease faster as maximum out-of-pocket payment increases. The reason is that an exponential distribution has highest density at 0 which then falls as $EOOP$ increases. In

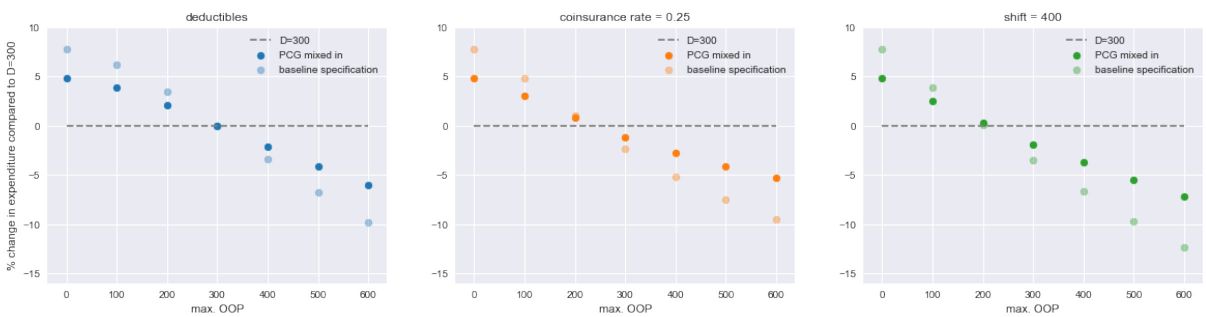

Figure 19: Comparison of the results with the baseline sample and the sample with individuals with a PCG mixed in

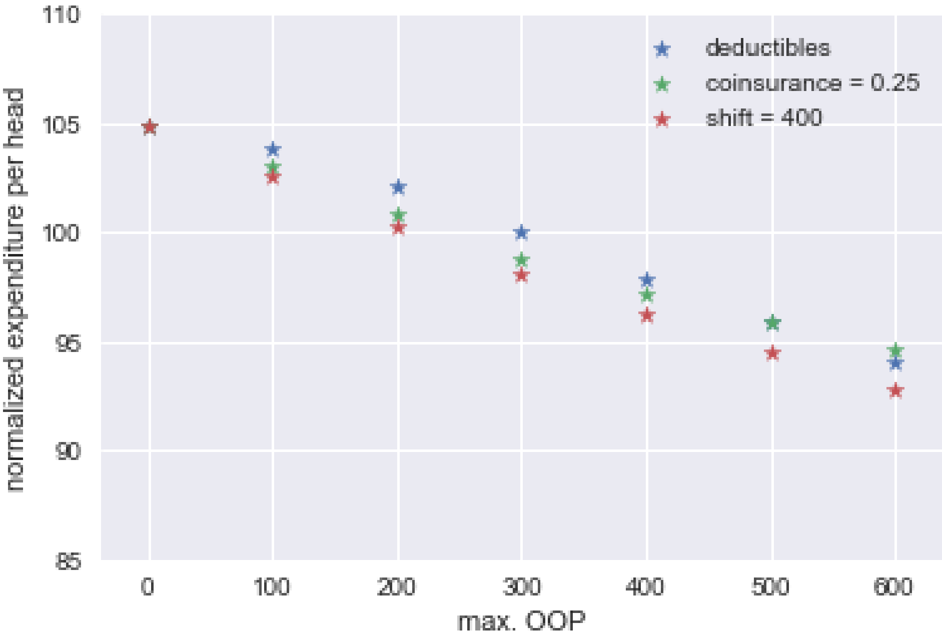

Figure 20: Results when persons with a PCG are added

this case, eliminating out-of-pocket payments increases expenditure a lot. This effect is smaller for a normal distribution. As $EOOP$ increases, the density only decreases for an exponential distribution but can increase for a normal distribution. Hence, effects of an increase in $EOOP$ with a normal distribution exceed the effects with an exponential distribution.

Also, Figure 22 confirms that the main findings of this paper hold if we change the specification of F. To illustrate, at $D = 300$, a 400 euro shift in the starting point of the deductible reduces expenditures more than a 25% co-insurance.

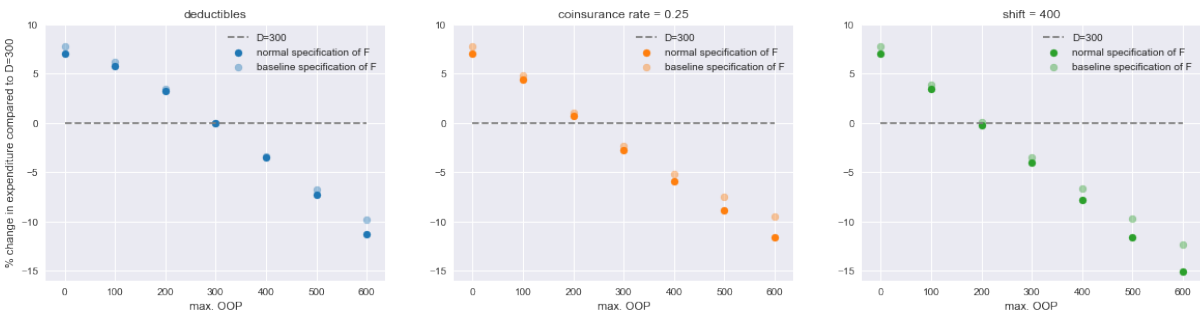

Figure 21: Comparison of the results with the baseline F-specification and the normal F-specification

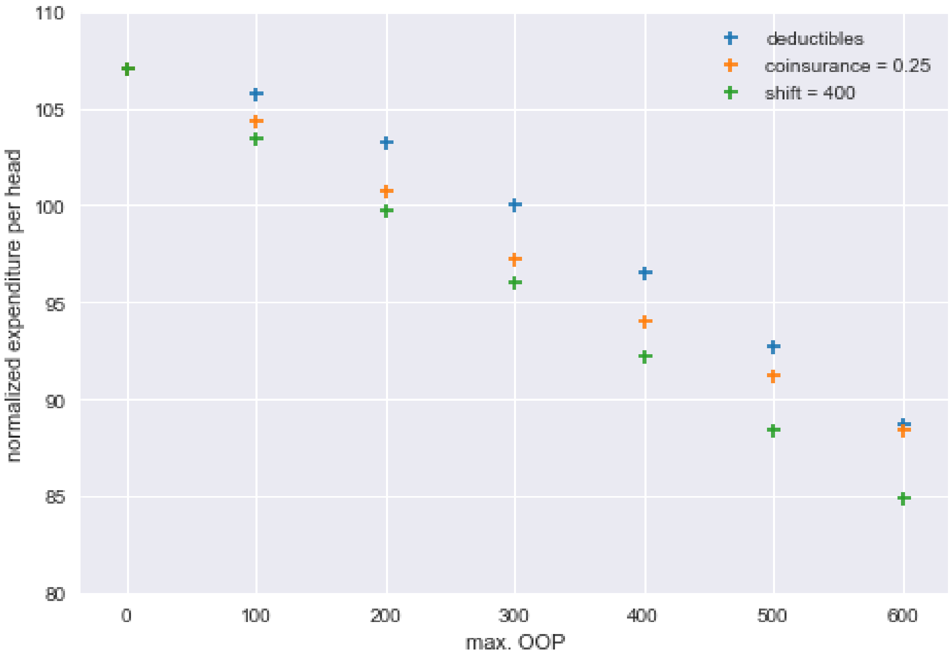

Figure 22: Results for a different specification of F

7.3 Specification of $EOOP$

In our model, we assume that individuals are fully rational and determine the expected out-of-pocket cost ($EOOP$) of a treatment using the correct underlying distributions for x and y (see Lemma 1 in Section 2.3). To check whether this is a reasonable assumption, we allow individuals to underestimate the variance of these distributions. We introduce two parameters $q_{x,y}$ –which vary by age and gender– in the model which are multiplied by the variance $\sigma_{x,y}$ in the expression for $EOOP$ only. The prior distributions for these parameters are uniform on $[0, 2]$. Hence, the prior expectation is that people are rational, but we allow for $q_x = q_y = 0$: people decide on the basis of expected values only and ignore variance. We find that the posterior distributions for these parameters have a mean of 1 and are single peaked around 1.0. Introducing these parameters does not change the results (see Figure 23).

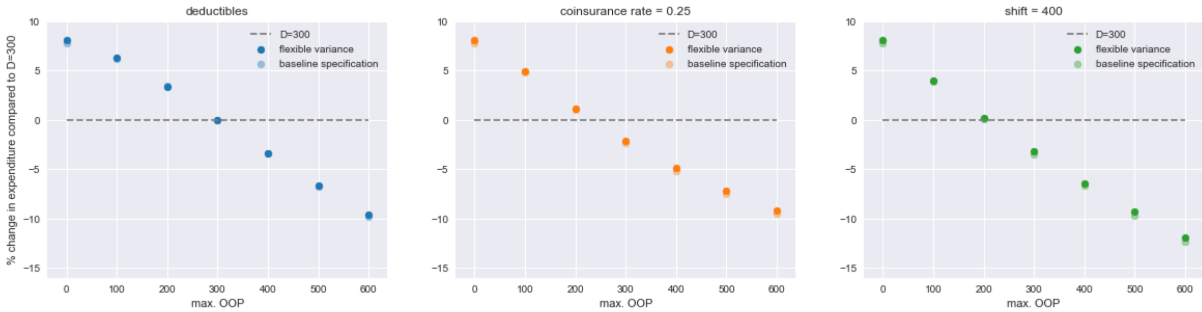

Figure 23: Comparison of the results with the baseline specification and the flexible specification of the variance in *EOOP*

8 Policy implications and discussion

In this paper, we show that a shifted deductible is an effective way to reduce healthcare expenditures without increasing out-of-pocket risk. Compared to a standard deductible, shifting the starting point alleviates the trade-off between efficiency and equity in the Netherlands. Shifting the deductible by 400 euros leads to an average 4% reduction in healthcare expenditure and 47% lower out-of-pocket payments for insured individuals who do not have a chronic condition. For the Netherlands in 2013, this shifted deductible is more effective than a co-insurance rate of, for instance, 25%.

To assess the effects of these multiple cost-sharing schemes, we use a structural model for demand-side cost-sharing focused on distributions of healthcare expenditures and of treatment utilities for different gender-age categories. As we estimate these distributions, we determine for each scheme the probability that individuals in an age-gender category are at the margin for a particular scheme. The more individuals in the category are likely to be at the margin, the more effective a scheme is to curb healthcare expenditure. We show that the effectiveness of a shifted deductible is robust to various specifications and sample selections.

The advantages of our model are the following. Because we model the distributions of expenditures, we can evaluate the effects of any cost-sharing scheme. As we model the distributions of treatment values we can also simulate the effects of changes in healthcare quality. Finally, because simulations are based on the posterior distributions of parameters, we can quantify the remaining uncertainty for each of our conclusions.

The structural model has many possibilities for extensions and additional analyses, which are not included in the paper. To illustrate, we left the derivation of the optimal demand-side cost-sharing scheme for future research. This could for example be a two-tier system with 100% co-insurance rate over the expenditure range between 0 and 300 euros and then a 50% co-insurance rate between 300 and 600 euros. The categories which determine an individual's (expected) distribution of healthcare expenditure are in this paper simply based on gender and age. However, they can be made more homogeneous by, for example, using an individual's expenditure in the previous year. A category can then be: a 25 year old male with less than 1,000 euro expenditure in the previous year. Another extension is to incorporate the voluntary deductible and its selection effects into the model. For this we need to estimate risk aversion in the model to determine an individual's decision whether to accept higher out-of-pocket risk in return for a lower premium. As we focused on mandatory insurance and deductible here, we did not need to model risk aversion yet.

The model also has some limitations and a number of these originate from the data. For example, the data comprise total healthcare expenditures per person per year, but not the exact underlying treatments, visits, scans and check-ups. As a result, we cannot simulate the effects of cost-sharing schemes such as co-payments, in which people pay for example 50 euros per visit to the hospital. Further, when simulating high levels of cost-sharing with the model, the results should be interpreted with caution. This is due to the fact that the model has been estimated on an increase in the deductible size from 150 euros to 350 euros. As we have mentioned, the model is not estimated on the entire Dutch population, but part of it. We argue that the baseline sample is the most relevant one, given the purpose of the model, because this group is elastic with respect to the cost-sharing observed in the data. Yet, this selection does reduce the applicability of the model to the entire Dutch population. Also, the Dutch healthcare system differs from healthcare systems in other countries. Thus our conclusions do not necessarily generalize to other countries. However, our framework can be used in any setting where data on individual healthcare expenditure is available.

Finally, our policy implication that a 400 euro shift in the starting point of the deductible reduces expenditure is based on the rational approach in the model: maximizing the likelihood of people being at the margin. We do want to mention two caveats here. First, there is evidence that people do not fully understand health insurance contracts (Handel and Kolstad, 2015). Arguably, a shifted deductible is harder to understand for people than a standard deductible contract. Some may not react to a shift as our rational model assumes. To illustrate, from a behavioral point of view, people may interpret a 400 euro shift as a focal point: I am expected to spend 400 euros (for free) every year. If this were the case, the effects on expenditure would be less favorable. As we do not have a shift in our data, more research is needed to test for this possibility. Second, to calculate the expected out-of-pocket price, people need to know their expenditure distribution. This is not a simple concept to fully understand. But we believe that over time people do get a sense of their healthcare expenditure and the probability of exhausting their deductible. If they for example exhaust their deductible for a number of years in a row, this will result in a low perceived out-of-pocket price, which is correct.

9 Appendix

9.A Proof of results

Proof of Corollary 1 This follows from writing

$$OOP = \delta \int_0^{\Delta+D/\delta} f_\chi(x)dx - \delta \int_0^{\Delta} f_\chi(x)dx - \delta\Delta \int_\Delta^{\Delta+D/\delta} f_\chi(x)dx$$
$$+ \left(1 - N\left(\frac{\ln(\Delta + D/\delta) - \mu}{\sigma}\right)\right) D$$

Q.E.D.

9.B Data cleaning procedure

To clean the data, observations are omitted if:

- the (pseudonymized) social security number is missing
- the postal code is not valid
- the health insurance registration period is missing or (impossibly) more than one year
- healthcare expenditures are negative
- the age sequence over time is erroneous

In total, 2,834,720 observations are excluded from the data which is equivalent to about 2% of the total number of observations.

9.C List of healthcare expenditure categories

Table 5 is duplicated from Remmerswaal et al. (2019a). Cost categories marked with X in the second column apply to the deductible. The other cost categories are exempted from the deductible. Z_{it} in the third column refers to the dependent variable in our baseline specification. The cost categories marked with an 'X' in the third column are included in Z_{it}.

Table 5: Cost types, whether they are under the deductible and whether we include them in Z_{it}.

Type of costs	Apply to the deductible	Included in Z_{it}
General practitioner registration		
General practitioner visits		
Other costs of general practitioner care		
Pharmaceutical care	X	X
Dental care	X	
Obstetrical care		
Hospital care	X	X
Physiotherapy	X	X
Paramedical care	X	X
Medical aids	X	X
Transportation for persons lying down	X	X
Transportation for seated persons	X	X
Maternity care		
Care that is delivered over the Dutch borders	X	X
Primary healthcare support		
Primary mental healthcare support		
Mental healthcare with (overnight) stay	X	
Mental healthcare without (overnight) stay:		
- at institutions	X	
- by self-employed providers	X	
Other mental healthcare costs	X	
Geriatric revalidation	X	X
Other costs	X	X

9.D The full model

Here we specify the structure of the model; the priors are given in the next section. To model total (log) expenditure we use a mixture model with 4 distributions (first list) and corresponding probabilities (second list). These probabilities follow from Table 1 above.

$$z \sim Mixture([0, f, g, f+g], [(1-\psi_x)(1-\psi_y+\psi_y F), \psi_x(1-\psi_y+\psi_y F), (1-\psi_x)\psi_y(1-F), \psi_x\psi_y(1-F))])$$

We mix over four distributions, the first of which has all mass on zero, the other three distributions are defined as:

$$\begin{aligned} f &\sim Normal(\mu_x, \sigma_x) \\ g &\sim Normal(\mu_y, \sigma_y) \\ f+g &\sim Normal(\mu_x+\mu_y, \sqrt{\sigma_x^2+\sigma_y^2}) \end{aligned}$$

The probabilities ψ_x, ψ_y that treatments are offered are each modeled as a Gaussian Process (GP) with age. For each of the GPs that we use, the covariance between two age categories x, x' (of the same gender) is given by

$$K(x, x') = \eta^2 (e^{-\frac{(x-x')^2}{2\ell^2}})$$

which is called the squared exponential or Gaussian kernel. Hence, we specify the corresponding GP as $GP(\eta, \ell)$.

We use the following expression for ψ_y which guarantees that $\psi_y \in [0, 1]$:

$$\begin{aligned} \psi_y &= \frac{1}{1+e^{-gp_{psi_y}}} \\ gp_{psi_y} &\sim GP(\eta_{psi_y}, \ell_{psi_y}) \end{aligned}$$

The expression for ψ_x for women is slightly more involved:

$$\begin{aligned} \psi_x &= \frac{1}{1+e^{(-gp_{psi_x}-I_{f_{21}}*decay)}} \\ gp_{psi_x} &\sim GP(\eta_{psi_x}, \ell_{psi_x}) \end{aligned}$$

where the indicator variable $I_{f_{21}} = 1$ for women aged 21 and over in the years 2011 and after;

and 0 otherwise. Recall that in 2011 contraceptives were removed from the basic package for women aged 21 and over. The `decay` term captures that older women are less likely to use contraceptives and hence their treatment probability is less affected by this removal from the basic package. We model `decay` as varying with age but constant over time (after 2011).

Consequently, ψ_y varies with age and gender, ψ_x varies with gender, age and years (but the latter only for women).

To calculate EOOP, we use Equation (3) where the deductible D is now a parameter as well:

$$EOOP = (I_{18+} + \alpha I_{18})(\psi_x(OOP(D, \mu_x + \mu_y, \sqrt{\sigma_x^2 + \sigma_y^2}) - OOP(D, \mu_x, \sigma_x)) + (1 - \psi_x)$$
$$(OOP(D, \mu_y, \sigma_y)))$$

where $I_{18+} = 1$ for age equal to 19 and higher and 0 otherwise, $I_{18} = 1$ for age 18 and zero otherwise and D varies over the years. Then we write the probability that a y treatment is rejected as[38]

$$F = F(EOOP, \zeta, \nu)$$

where ζ, ν vary with age fixed effects and gender.

The parameters of the distribution of x expenditure can be written as:

$$\begin{aligned} \mu_x &= gp_x + \text{year\_fixed\_effect\_x} \\ gp_x &\sim GP(\eta_x, \ell_x) \\ \sigma_x &= \sigma_{x_{age}} + \sigma_{x_{year}} \end{aligned}$$

where gp_x denotes the GP with age; `year_fixed_effect_x` and $\sigma_{x_{year}}$ denote year fixed effects – for the mean and standard deviation resp.– and $\sigma_{x_{age}}$ age fixed effects for the standard deviation. We have an identical structure for μ_y, σ_y.

The python specification of our model can be found in the file `estimation.py` in the `code` folder of our repository.

9.E Priors

In this section we present the priors that we use in the main analysis and motivate the choices that we made.

As we have no a priori information about whether x or y expenditures tend to be higher and/or more prevalent, we use the same set-up for the priors of μ_x, μ_y, and ψ_x, ψ_y respectively. For each of these variables, the age effects are modeled as a Gaussian Process (GP). That is, the

[38] Note that this is potentially confusing in terms of notation. We use f to denote the distribution of x and F as the cumulative distribution function of the value of y treatments. We could use another letter to denote the latter, but readers would not immediately associate e.g. L with a cumulative distribution function.

function of each of these variables w.r.t. age is drawn from a multivariate normal distribution with co-variance matrix K, where the element ij is given by:

$$K_{ij} = \eta^2 (e^{-\frac{(i-j)^2}{2\ell^2}}) \tag{9}$$

where η captures the variance and ℓ the smoothness of the GP. In words, in the prior the draws for ages i and j are correlated, but the correlation is lower as i and j are further apart. Intuitively, expenditure at age 20 is more closely correlated with expenditure at 21 than at 49. The priors for η and ℓ are HalfNormal which implies that these values are positive. Figure 6 illustrates why we expect a relatively smooth function between expenditures and age.

For the standard deviation of expenditures, this relation turns out to be less smooth. Hence, we model it as age fixed effects without correlation between ages. Also ν, ζ in Equation (7) are modeled as age fixed effects. We expect ν to be small as it "translates" $EOOP$ into a probability of rejection. The prior is assumed to be HalfNormal with $sd = 0.003$. The probability that a free treatment is accepted ζ is a priori assumed to be uniformly distributed between 0 and 1.

Year effects capture, for instance, policy changes and the introduction of new treatments which are more expensive than the previous ones (or cheaper, e.g. due to the introduction of generic drugs). We do not expect average expenditures to be correlated across calendar years and hence we model this as year fixed effects drawn from a normal distribution with $\mu = 4.5, sd = 0.7$. The value of μ is based on mean log expenditure in Table 3. Hence, we allow both x and y to capture total expenditure. The age fixed effects are modeled as offsets from the year fixed effects.

The values for x and y are drawn from a Normal distribution with expectation equal to the sum of the age Gaussian process and the year fixed effects and standard deviation equal to the sum of age and year fixed effects. The priors for the age and year fixed effects standard deviations are HalfNormal with sd equal to sd_x, sd_y. These sd_x, sd_y are drawn from a HalfNormal distribution with parameter $sd = 0.15$. We assume a hierarchical structure here which implies that data on sd_x is informative about sd_y and the other way around.

In 2011 there was a shock in the basic package for women above 21. We model the impact of the shock `package_f_21_2011` as being drawn from an Exponential distribution with parameter 2.0. As the shock concerns contraceptives, we expect the effect to wear off across age. This is captured by the variable `package_decay` which has age fixed effects drawn from a Uniform distribution on $[-1.0, 1.0]$. These effects feed into ψ_x as: `k(mu_psi-(package_f_21_2011* package_decay)*dummy_f_2190)`, where $k(x) = e^x/(1+e^x)$ denotes the inverse logit function, `mu_psi` the Gaussian Process of ψ_x w.r.t. age, `package_decay` depends on age and `dummy_f_2190` equals 1 for women, year 2011 and later and age 21 and higher.

Finally, the prior for the probability α that an 18 year old has had her/his birthday is modeled as Uniform on $[0, 1]$. Table 6 summarizes these priors.

Table 6: Priors used in the main analysis

Parameter	Distribution	Prior
Gaussian Processes for age		
x and y		
ℓ	HalfNormal	$sd = 2.0$
η	HalfNormal	$sd = 0.2$
ψ_x and ψ_y		
ℓ	HalfNormal	$sd = 2.0$
η	HalfNormal	$sd = 0.2$
Age fixed effects		
sd_x, sd_y	HalfNormal	$sd = 0.15$
σ_x, σ_y (hierarchical)	HalfNormal	$sd = sd_{x,y}$
ν	HalfNormal	$sd = 0.003$
ζ	Uniform	$[0.0, 1.0]$
Year fixed effects		
x and y		
`year_fixed_effect`	Normal	$mu = 4.5, sd = 0.7$
σ_x, σ_y (hierarchical)	HalfNormal	$sd = sd_{x,y}$
Year-age-gender adjustment		
`package_f_21_2011`	Exponential	2.0
`package_decay`	Uniform	$[-1.0, 1.0]$
Separate parameter		
α	Uniform	$[0.0, 1.0]$

As expenditures are distinguished into two components (x, y) which are not directly observable, it is not straightforward to gauge whether our priors are reasonable. One way to judge whether our priors are reasonable is by generating expenditures directly from the priors and compare these prior predictive outcomes with metrics we know about healthcare expenditures, without using our data. For example, we know that on average the nominal premium paid in the Netherlands is roughly 1,000 euros. This is supposed to cover 50% of costs, the other half comes from income dependent contributions collected by employers (see Section 3.1). Hence, average total expenditures per head per year will be roughly 2,000 euros. This is an upper-bound on the expenditures in our sample selection because: (i) we focus on expenditures under the deductible (only) while the 2,000 euros covers all expenditures, (ii) we select relatively healthy individuals and exclude the chronically ill who tend to have higher expenditures. Hence we expect average expenditures to be lower and about 1,500 euros per person per year. To get an idea of the maximum healthcare costs in the data, suppose that for our sample selection, the best treatment would provide the patient 10 additional years in full health. If we value a QALY (quality adjusted life year) at 100,000 euros per year, this treatment can maximally cost 1,000,000 in a year. Lastly, to determine an expected standard deviation: within the sample of a gender-age category of 10,000 individuals, we expect to have maximally 1 individual with expenditure exceeding 500,000 euros. Using Chebyshev's inequality, we then have a standard deviation within

Table 7: Sensitivity of outcomes w.r.t. priors

Priors changed	Average expenditure	Standard deviation	Maximum expenditure
None	1,256	7,108	1,158,945
$sd_{x,y} = 0.2$	1,482	9,639	53,785,019
Year fixed effects: $sd = 1.0$	1,932	14,452	4,515,761

a gender-age category of approximately 5,000 euros.[39] As shown in Table 7, with the priors specified above we find from the model that has not yet fitted the data, average expenditures of around 1,200 euros, a standard deviation of 7,000 euros and maximum expenditures of one million. These numbers are in the ballpark of the numbers mentioned above.

At first sight, it may seem that the prior standard deviations are rather small and hence the priors on the mean may seem tight. However, this is not the case. First, the prior standard deviation, as shown in Table 7, is already 7,000 euros which is above our expectation of 5,000 euros. Hence, the overall standard deviation on expenditures per category turns out not to be very tight. Second, if we increase the $sd_{x,y}$ components of the standard deviations from 0.15 to 0.20, the standard deviation per category increases to almost 10,000 euro. Moreover, the maximum expenditure then exceeds 53 million euros which is far above what we would expect. Third, if we alternatively would increase the prior standard deviation on the year fixed effects from 0.7 to 1.0, the prior standard deviation per category becomes 14,500 and the maximum expenditures become 4.5 million euros. Both are way beyond what one would expect from the data.

We have estimated the model using these different priors. The results are very similar to the results when using our baseline priors. This robustness to prior choices is not surprising given the size of our data set. With more than 32 million observations (see Table 3) one would expect the data to supersede the priors.

9.F Additional figures

This section presents plots that are in the main text for the opposite sex (only). For men we also find a clear pattern in log expenditures (conditional on being positive) across age which is stable across calendar years 2008 to 2013 (see Figure 24). The ELBO plot, as shown in Figure 25, also suggests convergence for men when using the ADVI algorithm. Plotting the predicted probability of positive expenditures for women in 2008 across age against the validation data suggests a fairly good fit of this GP (see Figure 26). Figures 27 and 28 show that the fit for women is quite good both in terms of first and second moments, although for 2013 we under estimate expenditures for women above 60.

[39] Roughly speaking, $Prob((x - 2,000) > 5,000 * 100) \leq 1/100^2$ for $x > 500,000$.

Figure 24: Log healthcare expenditure for men conditional on being positive.

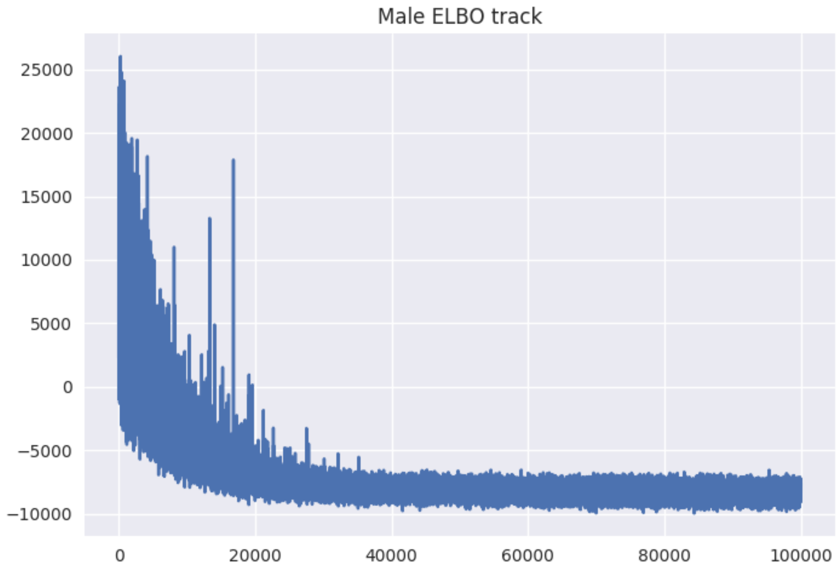

Figure 25: ELBO model estimation for men

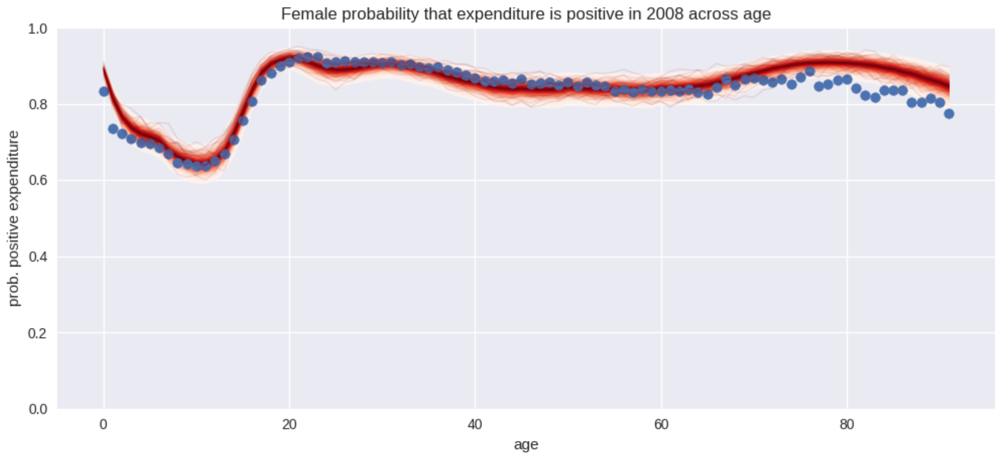

Figure 26: Predicted and realized probabilities of positive expenditures for women across age in 2008.

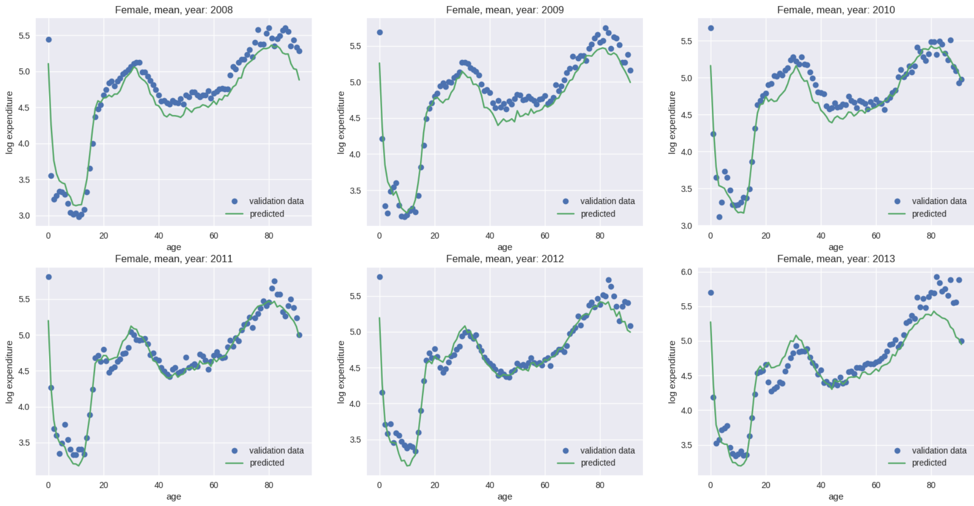

Figure 27: Average predicted vs average validation female log healthcare expenditures for 2008 to 2013

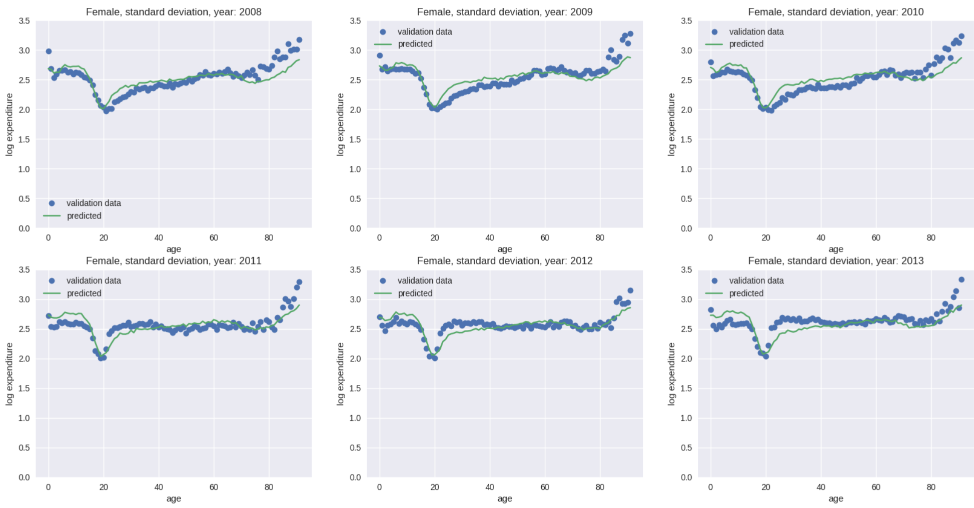

Figure 28: Standard deviation predicted vs validation female log healthcare expenditures for 2008 to 2013

Bibliography

P. Bajari, C. Dalton, H. Hong, and A. Khwaja. Moral hazard, adverse selection, and health expenditures: A semiparametric analysis. *The RAND Journal of Economics*, 45(4):747–763, 2014.

D.M. Blei, A. Kucukelbir, and J.D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877, 2017.

Z. Brot-Goldberg, A. Chandra, B. Handel, and J. Kolstad. 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, 2017.

J.H. Cardon and I. Hendel. Asymmetric information in health insurance: Evidence from the national medical expenditure survey. *The RAND Journal of Economics*, 32(1):408–427, 2001.

D. Cattel, R.C. van Kleef, and R.C.J.A. van Vliet. A method to simulate incentives for cost containment under various cost sharing designs: an application to a first-euro deductible and a doughnut hole. *European Journal of Health Economics*, 18(8):987–1000, 2017.

P. Deb, K.M. Munkin, and P.K. Trivedi. Bayesian analysis of the two-part model with endogeneity: Application to health care expenditure. *Journal of Applied Econometrics*, 21:1081–1099, 2006.

N. Duchovny, J. Kowalczyk, S. Laughery, and S. Sajewski. Cbo's medicare beneficiary cost-sharing model: A technical description. Congressional Budget Office Working Paper 2019-08, 2019.

Dutch Healthcare Authority. Marktscan en beleidsbrief zorgverzekeringsmarkt: weergave van de markt 2010-2014. Dutch Healthcare Authority, 2014.

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

L. Einav, A. Finkelstein, and P. Schrimpf. The response of drug expenditure to non-linear contract design: Evidence from medicare part d. *Quarterly Journal of Economics*, 130(2): 841–899, 2015.

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

Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

John Geweke, Gautam Gowrisankaran, and Robert J. Town. Bayesian inference for hospital quality in a selection model. *Econometrica*, 71(4):1215–1238, 2003.

B.H. Hamilton. Estimating the demand for health care with panel data: a semiparametric bayesian approach. *Health Economics*, 8:403–414, 1999.

B.R. Handel and J.T. Kolstad. Health insurance for "humans": Information frictions, plan choice, and consumer welfare. *American Economic Review*, 105(8):2449–2500, 2015.

M. Jochmann and R. Leon-Gonzales. Estimating the demand for health care with panel data: a semiparametric bayesian approach. *Health Economics*, 13:1003–1014, 2004.

E. B. Keeler, J. P. Newhouse, and C. E. Phelps. 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, 1977.

T.J. Klein, M. Salm, and A. Hayen. Does the framing of patient cost-sharing incentives matter? the effects of deductibles vs. no-claim refunds. CEPR Discussion Paper No. 12908, 2018.

T.J. Klein, M. Salm, and S. Upadhyay. The response to dynamic incentives in insurance contracts with a deductible: Evidence from a differences-in-regression-discontinuities design. IZA Discussion Paper Series No 13108, 2020.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D.M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18:1–45, 2017.

Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* Chapman and Hall/CRC, second edition, 2020.

A. Mukherji, S. Roychoudhury, P. Ghosh, and S. Brown. Estimating health demand for an aging population: A flexible and robust bayesian joint model. *Journal of Applied Econometrics*, 31: 1140–1158, 2016.

J.P. Newhouse and the Insurance Experiment Group. *Free for All? Lessons from the RAND Health insurance experiment.* Harvard University Press, Cambridge, Massachusetts, 1993.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press Cambridge, 2005.

M.C. Remmerswaal, J. Boone, M. Bijlsma, and R. Douven. Cost-sharing design matters: A comparison of the rebate and deductible in healthcare. *Journal of Public Economics*, 170: 83–97, 2019a.

M.C. Remmerswaal, J. Boone, and R. Douven. Selection and moral hazard effects in healthcare. *CPB Discussion Paper 393*, 2019b.

John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2, 2016.

W.P.M.M Van de Ven and F.T. Schut. Universal mandatory health insurance in the Netherlands: A model for the United States? *Health Affairs*, 27(3):771–781, 2008.

R.C. Van Kleef, van de Ven W.P.M.M., and van Vliet R.C.J.A. Shifted deductibles for high risks: More effective in reducing moral hazard than traditional deductibles. *Journal of Health Economics*, 18:198–209, 2009.

P. Zweifel and W.G. Manning. *Handbook of Health Economics*, volume 1, chapter Moral Hazard and Consumer Incentives in Health Care. Elsevier, Amsterdam, 2000.