CPB Netherlands Bureau for
Economic Policy Analysis

A Structural Microsimulation Model for Demand-Side Cost-Sharing in Healthcare

We developed a structural microsimulation model to predict healthcare expenditure under different demand-side cost-sharing schemes in the Netherlands, such as a deductible, a co-insurance scheme, and a shifted deductible. The parameters of the model are estimated with a Bayesian mixture model on a large Dutch administrative dataset.

We show that replacing the deductible in healthcare in the Netherlands by a shifted deductible or a co-insurance scheme can reduce healthcare expenditure without increasing the financial burden for persons with high healthcare costs.

CPB Discussion Paper

Jan Boone, Minke Remmerswaal

juni 2020

Doi: https://doi.org/10.34932/2dcx-9103

A Structural Microsimulation Model for Demand-Side Cost-Sharing in Healthcare*

Jan Boone† and Minke Remmerswaal‡

June 18, 2020

Abstract

Demand-side cost-sharing schemes reduce moral hazard in healthcare at the expense of out-of-pocket risk and equity. 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 percent reduction in healthcare expenditure and 47 percent lower out-of-pocket payments. We use administrative healthcare expenditure data and focus on the price elastic part of the Dutch population to analyze the differences between the cost-sharing schemes. The model is estimated with a Bayesian mixture model to capture distributions of healthcare expenditure with which we predict the effects of cost-sharing schemes that are not present in our data.

JEL codes: I11, I13, I14

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

*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, University of Queensland, the European Commission, 2019 ENTER Jamboree, the Dutch Ministry of Health, Tilburg University.

†Department of Economics, Tilec, Tilburg University, CPB Netherlands Bureau for Economic Policy Analysis, The Hague, the Netherlands and CEPR, London, United Kingdom

‡CPB Netherlands Bureau for Economic Policy Analysis, The Hague and Department of Economics, Tilburg University, the Netherlands

1 Introduction

Most healthcare systems and plans in the world 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. As such, it can improve efficiency. However, cost-sharing is also the topic of political debates, as it increases the 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 healthcare in the Netherlands, which we estimate on a large administrative dataset of inhabitants in the Netherlands. With the model we can show how multiple 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 for the price elastic part of the Dutch population and 47% lower out-of-pocket payments.

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. The effective price depends on the design of the cost-sharing scheme and individuals' health status (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 expenditure– 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) and highest deductible of 350 euros (in place 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: average expenditure for men are around 350 euros between age 20 and 40, whereas 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 every 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.

The top-right panel shows the probability that expenditure is positive for each gender-age category.

[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.

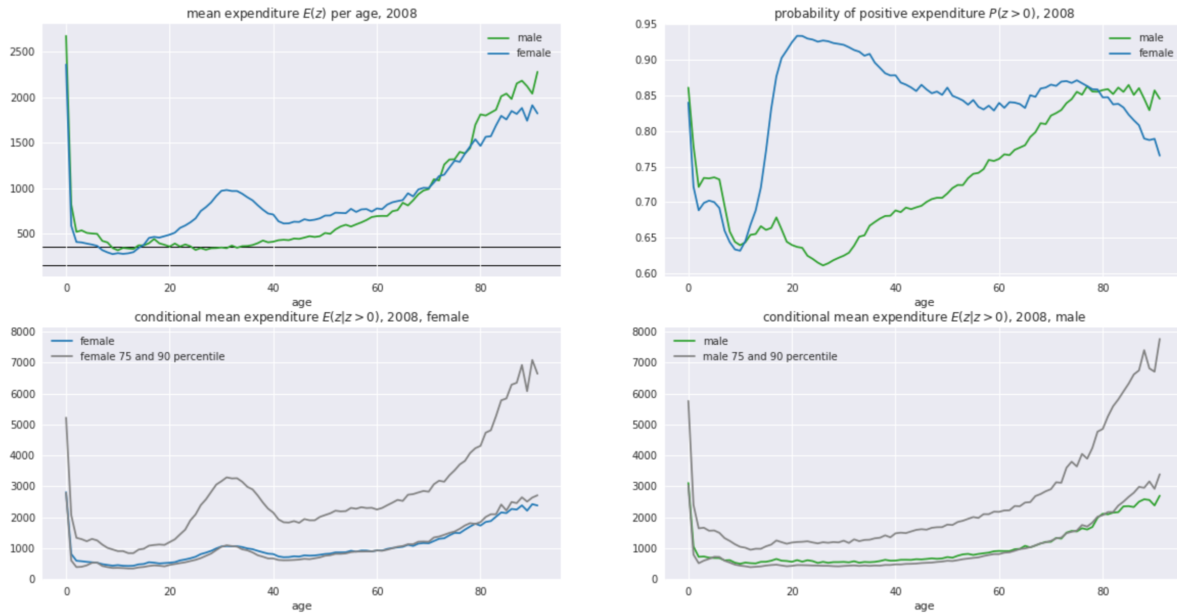

Figure 1: Healthcare expenditure across age for men and women

There is a substantial share of men and women who have no healthcare costs in a year. Also, the variation across gender and ages is large. It is unlikely that everyone would react in the same way to an increase in the deductible. The bottom row of figures shows the expected expenditure, conditional on positive expenditures. For positive expenditures, the 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 on the margin. We use a Bayesian mixture model and work with the distributions of healthcare expenditure per gender and age category. The uncertainty of estimation and simulation outcomes can be preserved in each step of the analysis with the posterior distributions. Furthermore, Bayesian methods combined with a structural model enable us to simulate the effects of cost-sharing schemes which have not been in place in Dutch healthcare (and in our data).

The paper begins with a theoretical model where an agent can face two types of healthcare expenditures: exogenous and endogenous expenditures. Exogenous expenditures are of such high value that changes in the deductible (over the Dutch deductible range) do not affect the decision to accept such a treatment. Think of breaking a leg and being plastered up in hospital. For the endogenous part of expenditures, people with a high deductible are more likely to forego the treatment. This decision is based on the agent's expected out-of-pocket expenditure of a treatment ($EOOP$) where the expectation is taken over the distribution of 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 theoretical model is literally brought to the data (not in reduced-form) and estimated as a mixture model with four components. We assume Gaussian processes for both endogenous and exogenous expenditures across age. In contrast to age fixed effects, this implies that knowing the average expenditures of for example 20 year old males, helps to predict 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 the gender-age categories. 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.

The model is estimated on administrative data which comprises healthcare expenditures of Dutch inhabitants between 2008 and 2013. The baseline model is estimated on more than half of all Dutch inhabitants: the baseline sample excludes price inelastic individuals who have been labeled in the data as chronically ill and who have had high healthcare costs in previous years as well as individuals who are labeled as chronic users of medication. As they are rather price inelastic and have high healthcare expenditures, including them in the results would obscure the differences in the effects of the cost-sharing schemes. Furthermore, their distributions of healthcare expenditures are very different from persons who are not labeled with these indicators and the number of observations per gender-age category is low. Furthermore, For the cleanest identification of the effect of the mandatory deductible, persons who chose an additional deductible voluntarily or persons who used mental health services have also been excluded from the baseline sample.

The main results of the model are presented in figure 2, with the effect of different levels of deductibles, co-insurance schemes and shifted deductibles which start at 400 euros are shown in the left panel. In this figure, we compare all results to a 300 euro deductible and therefore we normalize 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 (maximum out-of-pocket payment is 0) 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 costs 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 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 (between $[0, 1200]$) but at a lower rate of 0.25. 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.[2]

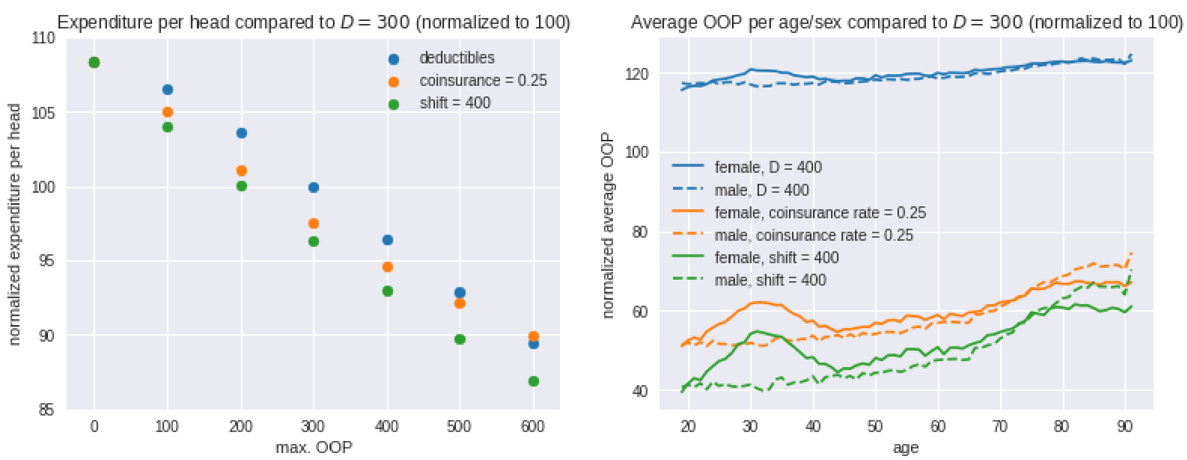

Figure 2: Average versus distribution effects of demand-side cost-sharing

There is a large body of literature on demand-side cost-sharing and moral hazard. Many models in the cost-sharing literature (or moral hazard and adverse selection 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 decisions whether to buy insurance and how generous the insurance should be. As we explain in section 3, these decisions are not relevant in our context of mandatory basic insurance in the Netherlands. Our paper is most comparable to 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

[2]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 $D = 300$ starting at 0.

by Dutch policy makers.[3] Furthermore, Einav et al. (2013) only compare a no-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 Process (GP) to model average expenditures. Moreover, we have a four component mixture model to capture both endogenous and exogenous expenditures.

Recently the Congressional Budget Office published a technical description of their Medicare Beneficiary Cost-Sharing Model (Duchovny et al. (2019)). The purpose of their model is similar to ours: to estimate the budgetary effects of various cost-sharing schemes at a population level. Like this paper, they use a microsimulation model which 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 (1993)).

This paper builds on a 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 in place (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.[4] 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 (the

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

[4] 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'.

starting point is zero).

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 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.

2 A Model

The main aim of this study is to simulate average 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 other healthcare expenditures an individual expects to make.

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

We follow Einav et al. (2013) and allow for two types of treatment: high value treatments which are exogenous with regard to cost-sharing (over the Dutch deductible range), 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 2008-2013 (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 Four components

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).[5] 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 regard to the deductible and

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

consists of treatments where agents *do* take the deductible into account. For example, if a physician offers the agent extra imaging services, she may or may not accept this.[6]

Like Einav et al. (2013) 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 3.3 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.[7] As a result, $Z = 0$ if and only if $z = 0$; further $z \geq 0$. 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 (i.e. a positive draw of x), ψ_x. This treatment, if offered, 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 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 is however not entirely accurate as the model consists of four, not two, components when we also take the probabilities that x and y are positive, ψ_x and ψ_y, into account. This is shown 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
$x = y = 0$	$(1 - \psi_x)(1 - \psi_y + \psi_y F)$
$x > 0 = y$	$\psi_x(1 - \psi_y + \psi_y F)$
$y > 0 = x$	$(1 - \psi_x)\psi_y(1 - F)$
$x, y > 0$	$\psi_x\psi_y(1 - F)$

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

[6] 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.

[7] 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.

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 manifests when calculating $EOOP$. We think of $EOOP$ as follows: at the start of the period, before exogenous expenditure x has been realized, the agent is offered an endogenous treatment y. When the treatment is offered, the agent does not know exactly the price and hence the out-of-pocket payment of the treatment. Hence, 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 have expenditures above a 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 at the end of the year turn out to cost him money. Hence, 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), we assume here they have a least some idea of their health status and of the healthcare they need and the cost of those treatments, and we incorporate that in their $EOOP$. We model $EOOP$ as the expectation across the distributions of x and y.[8] 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 at maximum 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.[9] 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 the agent 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 normally distributed (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 estimation time considerably.

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

[9] Technically speaking this is feasible. We can allow for n different y distributions. If an agent is offered a y treatment, we first draw the y distribution from which the agent 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 being offered is also feasible but more complicated. In that case, we need to solve the dynamic optimization problem for the agent where she takes into account that accepting the first y treatment makes following y treatments cheaper. Both extensions are left for future research.

We use the following 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 ($\mu = 0, \sigma = 1$). 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 an agent 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 $1 - \psi_x$, the agent 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 $OOP(\mu_x + \mu_y, \sqrt{\sigma_x^2 + \sigma_y^2})$. This is the second computational gain of assuming a lognormal distribution of healthcare expenditures.

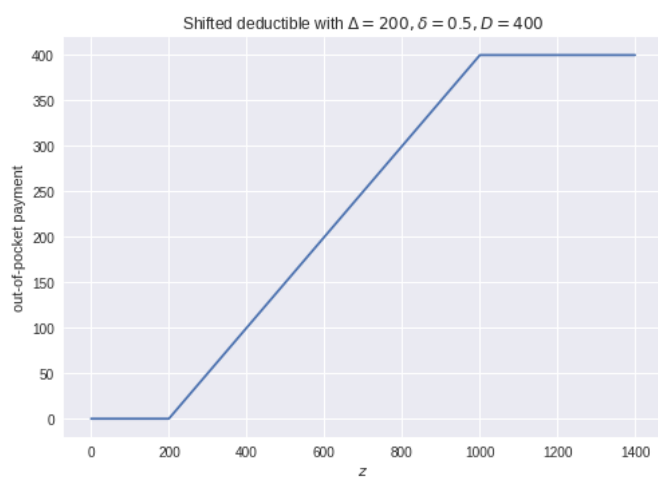

Figure 3: Out-of-pocket payments with a shifted deductible.

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

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

where z is an individual's total healthcare costs and $\Delta > 0$ 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.

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 health care 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 percent 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 $z = \Delta + D/\delta = 1000$ 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 payment schemes in the family in (5). For payment 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:

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

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

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

where D denotes the maximum OOP, Δ the shift in the starting point and δ the co-insurance 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.[10]

[10] 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. 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. Equation 7 can be used for many cost-sharing schemes, such as deductibles, co-insurance rates, shifted deductibles, et cetera.

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

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

where $\nu > 0$ and $\zeta \in \langle 0, 1]$. These parameters capture the price responsiveness of healthcare to changes in cost-sharing. This specification of F makes the parameters easy to interpret: we assume that the hazard rate is constant: $f(x)/1 - F(x) = \nu$ and ζ denotes the probability 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. We think of F as the cumulative distribution function of treatment utility or treatment value, v. A treatment is rejected if price exceeds value: $v < EOOP$.

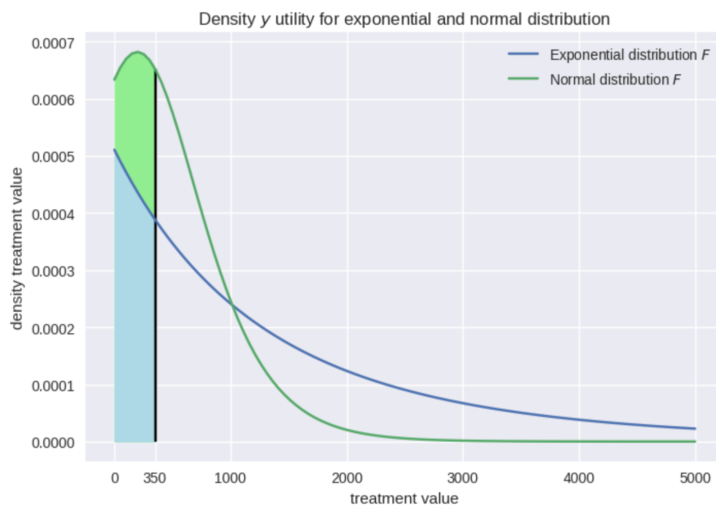

Figure 4: Two distributions of y treatment value.

This is illustrated for both the exponential distribution and a normal distribution (which we use in a robustness analysis) in figure 4.[11] 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 green area below the green curve denotes the same mass for a normal distribution of treatment value.

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.

[11] In fact, the distributions drawn are the average densities for the F distributions (averaged over the posterior distribution of parameters ν, ζ) 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 2008-2013. 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, which comprises hospital care, general practitioner care, physiotherapy, mental healthcare, et cetera.[12] Long term care and social care in the Netherlands are organized differently from curative healthcare, and therefore outside the scope of this paper. From this point onwards when we write healthcare, we refer to curative healthcare.

Curative healthcare is organised 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 face cost-sharing. The government regulates the market to ensure access to health care and protect risk solidarity. For example, the government has set up regulation that makes sure that health insurers cannot refuse persons 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 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 2008-2013

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

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

[12]For an exhaustive list, see appendix A.3.

2008, then he faced a total deductible of 650 euros. A voluntary deductible is chosen by only (roughly) 10 percent 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 health care. 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 insurances is optional and can be bought from a different insurer than insurance for the basic benefit package.[13] In this paper, we only study the basic insurance market and not the supplementary insurance market.

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 thus does not vary (much) between insurers. Consequently there is little if any selection in our data. Finally, we omit citizens from our data who chose a voluntary deductible in any year in our data.[14] This allows us to focus on the effect of cost-sharing on the decision to accept a treatment or not. In this sense, our model is simpler than some of the moral hazard models discussed in the introduction of this paper.

3.2 Data

Proprietary healthcare claims data are used to estimate this model. The data include all claims of all, approximately 17 million, Dutch inhabitants for years 2006-2013. 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 under-reporting of healthcare claims.[15] 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 A.2) 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 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 and estimations. After cleaning and excluding the years 2006 and

[13] Over 85 percent of the Dutch population bought supplementary health insurance in our data period (Nederlandse Zorgautoriteit, 2014).

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

[15] 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).

2007, the train data comprise over 58 million observations.[16]

Total healthcare expenditures of each Dutch inhabitant is one of the main variables in the data. As we are interested in the effect of cost-sharing, our expenditure variable only includes cost categories that fall under the deductible (see appendix A.3). 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.[17] Total healthcare expenditure can be separated into 21 healthcare categories, such as general practitioner care, maternity care, hospital care, and mental care.[18]

Several person characteristics are available such as gender, age, indicators of chronic use of care and 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.[19] 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 of 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.[20] Lastly, the annual choice of a voluntary deductible ranges between 0 (no voluntary deductible was chosen) to 500 euros (the maximum voluntary deductible).

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 was introduced during our data period in 2012. Furthermore, dental costs are excluded because the coverage of dental care changed for persons under 18 years old between 2008 and 2010, but not for persons above 18 years old. 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. 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 label DCG and/or PCG is relatively small and their distribution of healthcare expenditure is very different from people without it. This complicates 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 rather low.

All in all, by using the baseline sample, we lose about 44 percent of the data with the selection steps described above. However, the observations that we lose, are not directly relevant for our research

[16] See section 3.4 below to read more about the train data.

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

[18] Appendix A.3 is a list of all cost categories that are available in the data.

[19] 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.

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

question because the groups are rather inelastic to the changes in the mandatory deductible. For the DCG and PCG categories, it is the case that their expected expenditures are (far) above the highest deductible in our date. 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 euro can be compensated by a reduction of 100 euro 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. Dental costs, individuals who chose a voluntary deductible at least once in our data, chronically ill persons, chronic users of medication and persons who used mental health services have been excluded.

3.3 Descriptive statistics

Table 3 summarises 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: mean healthcare expenditures for women are 750 euros compared to 570 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.

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

	Females	Males
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

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 the sum of normal distributions. This is, indeed, what our mixture model does.

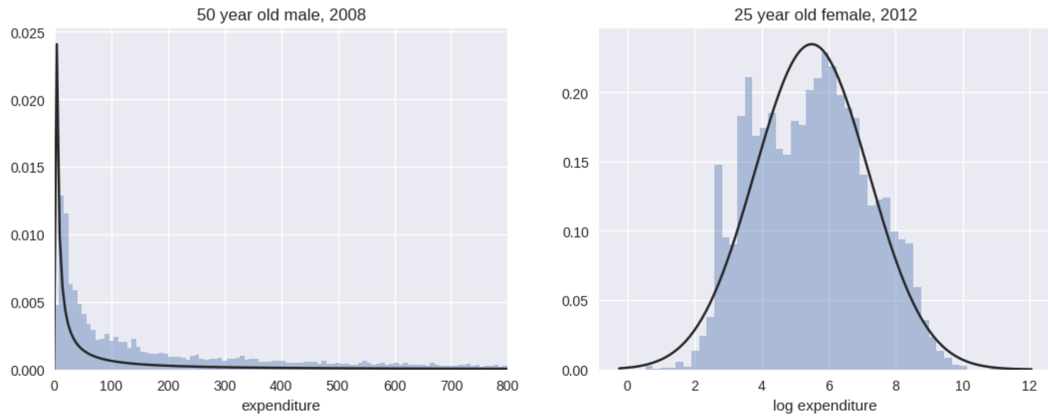

Figure 5: Two illustrative distributions for positive healthcare costs

3.4 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've 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%.

4 Econometric Specification

This section explains how we parameterise and identify and estimate the model.

4.1 Parameterisation

As mentioned, the distribution of (log) total healthcare expenditure per capita per year, z, is modeled as a mixture distribution with four components. As shown in figure 1, the distribution of expenditure by age is very different for men and women. Hence, we estimate the model separately for men and women.

We expect age to have an effect on components x and y and we model $\mu_x, \mu_y, \psi_x, \psi_y$ 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 years olds than for 20 and 50 year olds. We assume that the covariance decreases with the square of the age difference.[21]

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

To capture 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 very different from one year to the next and do not follow a coherent pattern.

For ψ_x and ψ_y we assume that the probability of being offered a treatment varies with age, but not by year.[22] That is, we assume that the probability of being offered a treatment is driven by the probability of falling ill. It does not vary over time. However the type and price of the particular treatment that you are offered can vary from year to year as the coverage changes. Furthermore, the probability that you accept a y treatment, varies across years with the deductible.

The standard deviation of expenditure z does not show a clear pattern across age.[23] 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 an agent faces a deductible

[21] 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).

[22] 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, 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.

[23] As illustrated in figure 12 below, the standard deviation for men randomly fluctuates around approximately 2.7.

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 2008-2013 (see section 3.2). These data provide three main sources of variation for identification.

The first source of exogenous variation is the change in the size of the deductible over time. As presented in table 2, the mandatory deductible was 150 euros in 2008, and 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 deductible affects the probability of positive expenditures and the distribution of expenditures z. This allows us to identify ν and ζ in the function F.

The second 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, i.e. a zero deductible. Moreover, it enables us to separate the effect of the deductible from other annual changes that affect healthcare expenditures, such as changes in coverage of the basic benefit package.[24]

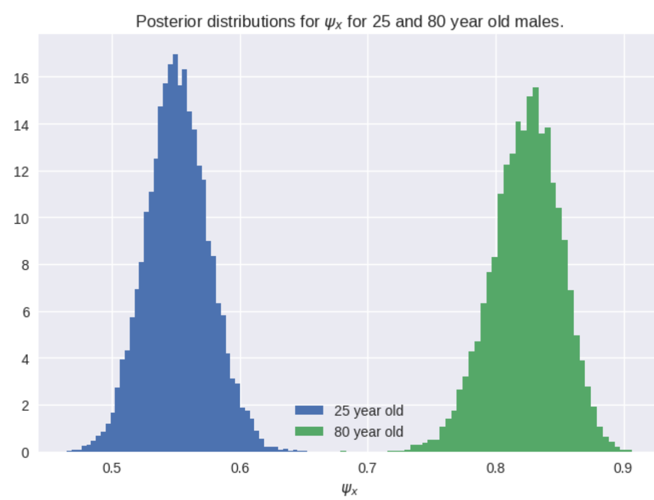

Figure 7: Distributions of ψ_x for men.

Finally, there is a lot of variation in the healthcare distributions among different gender-age categories. For example, the healthcare expenditure distribution of an 80 year old man differs greatly

[24]Remmerswaal et al. (2019a) and Remmerswaal et al. (2019b) exploit this variation for exactly this reason.

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.

These three sources of variation allow us to identify how healthcare expenditures vary with changes in demand-side cost-sharing.

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, even in our train data there are enough observations to determine the posterior distributions in a fairly robust way. The exact choice of priors can be found in appendix A.5. In the appendix we also argue that there are some bounds 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 show 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 out of 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.

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 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 usually very computationally intensive but can provide (asymptotically) exact samples from the target density (Blei et al., 2017). Variational inference can be less precise as it uses approximations, but much faster and able to deal with large datasets. That also makes it easier to estimate and compare different specifications of the model. Blei et al. (2017) state that especially for mixture models, the results of variational inference may be better than MCMC even for small datasets. Blei et al. (2017) also discuss that variational approaches tend to underestimate the variance of the posterior densities. Below we compare both first and second moments of our target distributions.

Figure 8 shows the ELBO plot for estimating the model for women.[25] The plot is skewed and flat on the right hand size, which implies convergence of the ADVI algorithm.

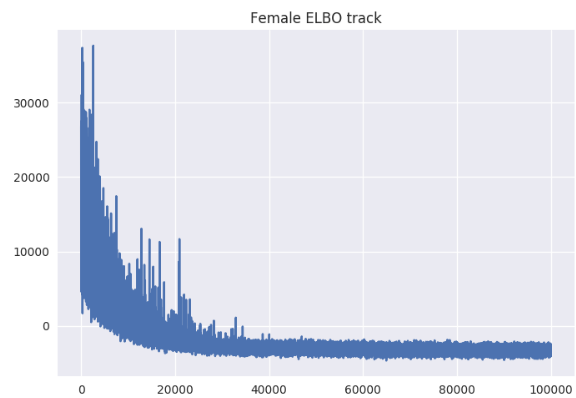

Figure 8: ELBO model estimation for women

5.2 *Parameter estimates*

Here we describe 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. Additional information is provided in appendix A.6.

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, around 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. Its posterior distribution is single peaked around 0.5.

[25] To keep this paper short, we show in most cases graphs for one gender only. For the other gender, graphs are presented in Appendix A.6.

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 suggests that 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.[26] The left figure shows that 20 year old 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. 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, next to the mean F, 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 (at $EOOP = 500$) is less than 0.65 (dashed blue line at this $EOOP$).

Table 4: Summary of posterior distributions.

Variable	Mean females	Std. dev. females	Mean males	Std. dev. males
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

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 ages 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

[26] We do not present plots of each posterior distribution here, as that is –literally– beyond the scope of this paper.

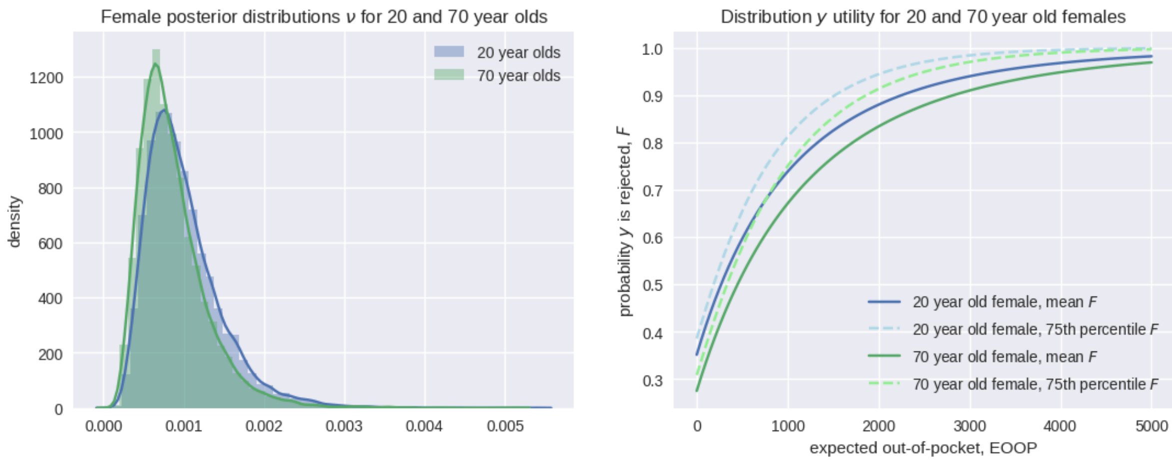

Figure 9: Posterior distributions of ν and F for women

of positive expenditures is fairly close to the realized fractions of positive expenditures for each age category. But we over-estimate this probability for men around 80 years old.

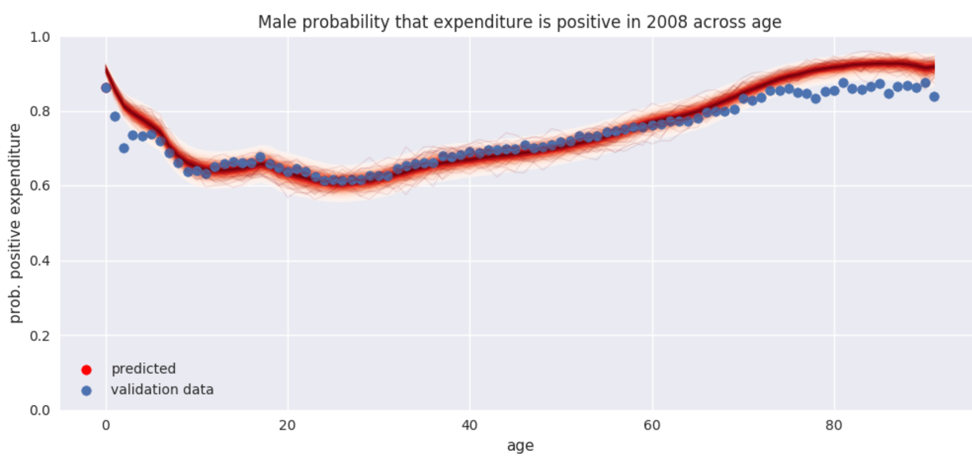

Figure 10: Predicted and realized probabilities of positive expenditures for men across age in 2008.

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.[27] 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 an offer is made (0 no offer is made).

With the samples of $\mu_{x,y}$ and $\sigma_{x,y}$ 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.

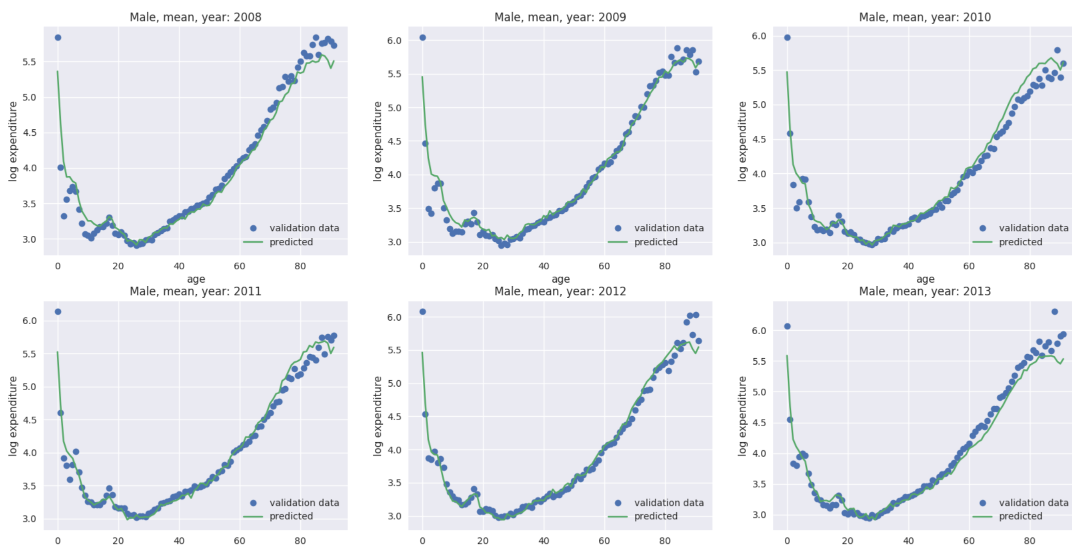

Figure 11: Average predicted vs average validation male log healthcare expenditures for 2008–2013

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.[28] The fit is quite good. Even the second moments, as shown in Figure 12 show a reasonable match.

Figure 13 uses the Kullback-Leibler (KL) divergence to measure the fit of the predicted distribution

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

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

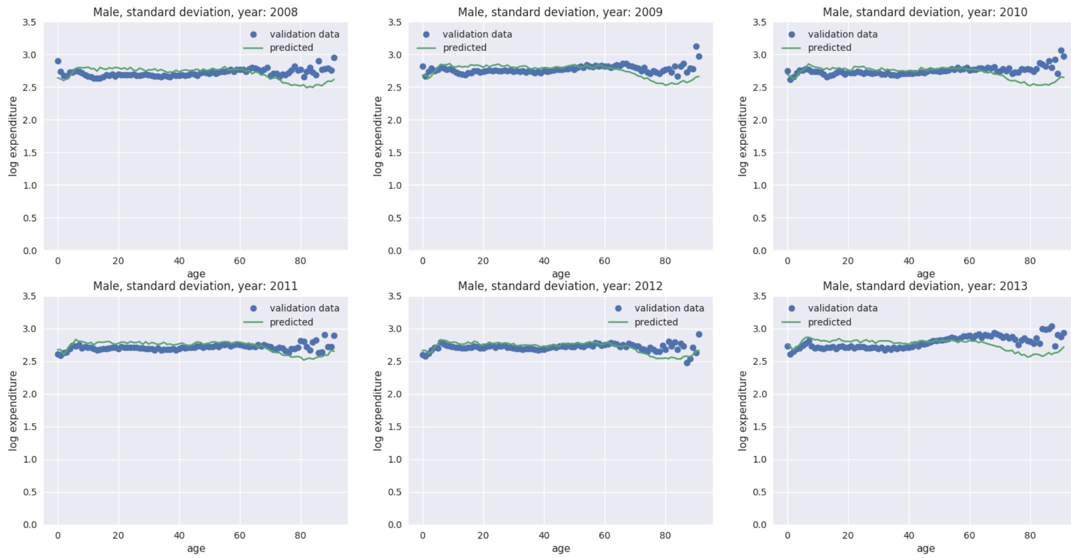

Figure 12: Standard deviation predicted vs validation male log healthcare expenditures for 2008–2013

per age-year category for men with the distribution in the validation data. A zero KL divergence implies a perfect fit. The divergence of our model (the blue dots) are close to zero, and the fit seems good. However as it can be hard to properly interpret the value of KL divergence, we compare the divergence of the model predictions with a simple (simplistic) benchmark. For our approach it is important to get the distribution per gender-age-year category right (this distribution is used to calculate $EOOP$ for the category). The green line in the figure gives the divergence between the age-year distribution of the validation data and the (true) average distribution across ages for each year for males in the train data. The latter distribution perfectly captures the data but not the differentiation across ages. For each age category, our model prediction performs either better than the green line (lower KL-divergence) and has a closer match to the validation distribution or the divergence is close to zero.

A final 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 14 compares the distribution of healthcare expenditure for 50 year old women in 2011, conditional on the expenditures being below 3,000 euros.[29] The distributions of the predicted data and the validation data are close. The right 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 over-estimate expenditures above 1,500 euros.

[29]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 3,000.

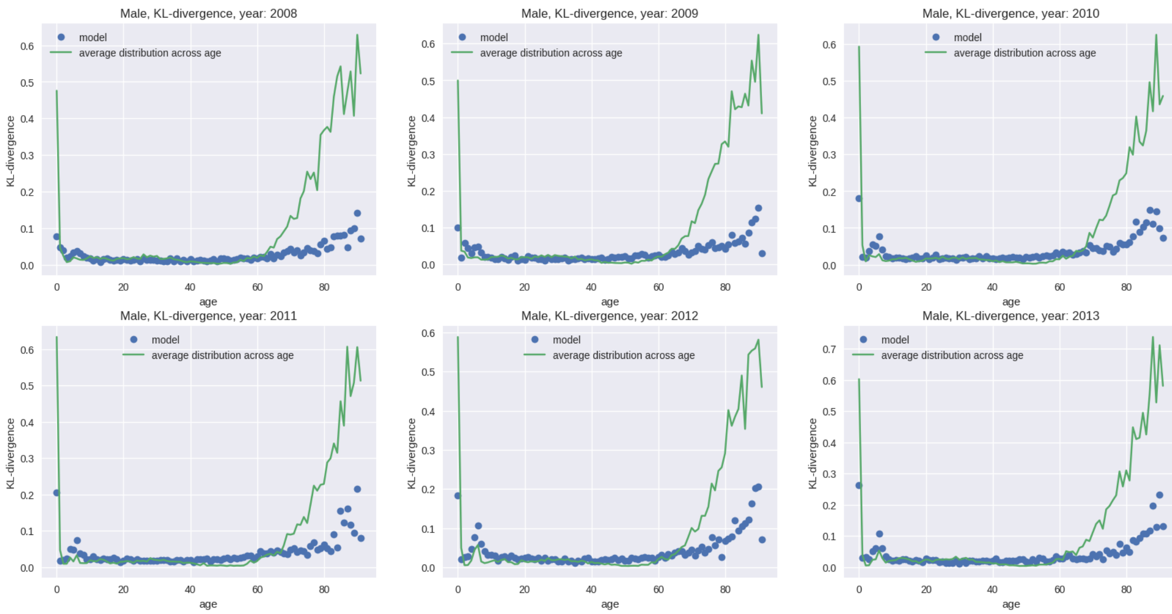

Figure 13: KL divergence expenditure distribution males for 2008–2013

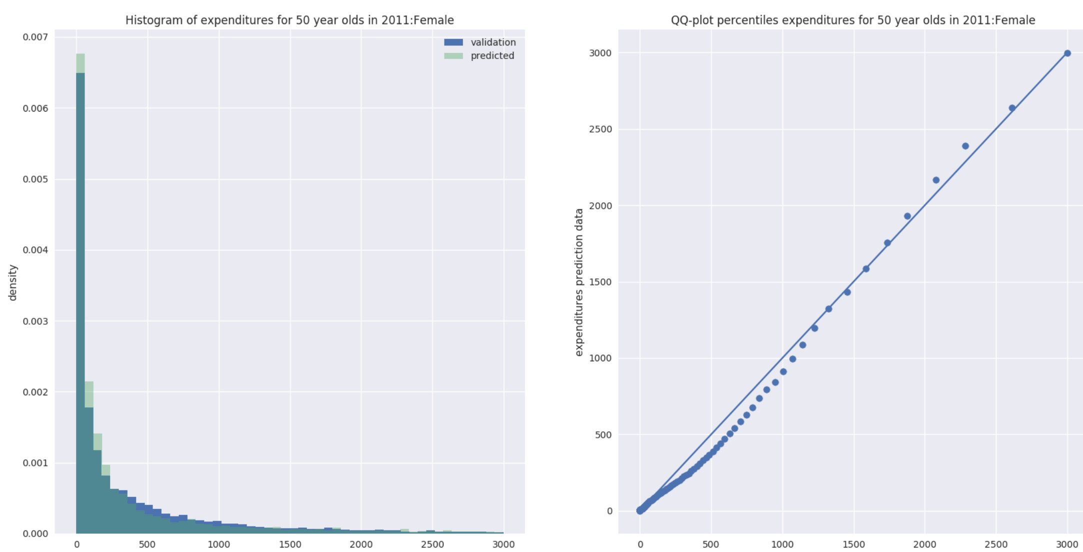

Figure 14: Predicted vs validation female log healthcare expenditures for 2011.

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. We do the simulations for the most recent year in our data: 2013. 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. Next, using the exponential function in equation (7) and the estimated posterior distributions of the parameters ν and ζ, we determine how the change in $EOOP$ affects the probability that a treatment is rejected (F). Lastly, we compute the healthcare expenditures under the new cost-sharing scheme.[30]

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, 1.0 in combination with seven starting points: 0, 100, 200, 300, 400, 500, and 600 euros.[31] Figure 2 shows the results of the simulations for some of the cost-sharing schemes normalized on expenditures for a 300 euro deductible. Expenditure per head decreases as the deductible level increases from 0 to 600. These results translate into a deductible elasticity of -0.09, which is the same as Remmerswaal et al. (2019b). We calculate the deductible elasticity here as follows:

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

Higher levels of D lead to lower expenditure per head but also higher out-of-pocket payments for people with high expenditure. The right panel in figure 2 illustrates that people have on *average* about 20% higher out-of-pocket payments with $D = 400$ compared to $D = 300$. On an individual level, this percentage will be even higher for persons with high health care use.

Figure 15 shows how healthcare expenditures vary for four different co-insurance rates and seven different starting points, or shifts. All schemes in the figure have the same out-of-pocket risk of 300 euros and we normalize expenditure on average expenditure with a 300 deductible (point (0,100) on the purple line in the figure). With a deductible 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 (i.e. no shift), a 25% co-insurance rate leads to the biggest reduction of healthcare expenditure.

Figure 15 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 to have a low co-insurance rate of 25 percent. This blunts the

[30] See also appendix A.4 for the full model with the specification of total (log) expenditure z in its underlying components. A change of cost-sharing scheme leads to a change in $EOOP$, which leads to a change of F.

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

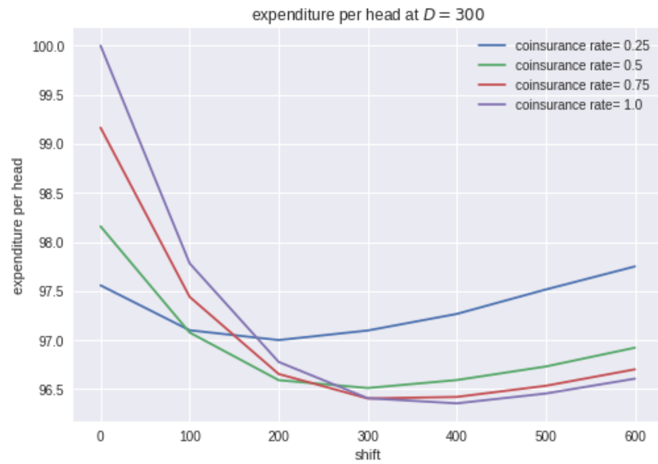

Figure 15: Expenditure per head for $D = 300$ and different levels for the shift and co-insurance rate relative to expenditures with a deductible equal to 300.

effect of the deductible (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 percent co-insurance rate with a maximum out-of-pocket payment of 300 euros reduces 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 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 in figure 15, 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 it is optimal to have a rate equal to 1. In other words, by introducing a shift, more people are 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 euro captures enough people on their margin that reducing the co-insurance rate is no longer useful. Not only does this minimize expenditure per head for $D = 300$, it also minimizes the out-of-pocket payment for everyone as figure 2 shows.

However, we cannot conclude that this is a Pareto improvement, because persons with low healthcare expenditure could be worse off under a co-insurance scheme or shifted deductible than under a deductible for the following reason: as mentioned before, cost-sharing schemes shift a part of healthcare expenditures to users of care. Under a deductible, a relatively large share of healthcare expenditures is paid out-of-pocket, by users of care, whereas under a co-insurance scheme or a shifted deductibe, a larger share of healthcare expenditure is financed out of premiums. Even though a shifted deductible which starts at 400 euros leads to lower overall healthcare expenditure, this reduction in total expenditure is smaller than the reduction in out-of-pocket expenditures; hence this scheme features a higher premium. For someone with low expected expenditure, the deductible scheme may be preferred over the other two schemes because of the lower premium. As we do not estimate the degree of risk aversion, we cannot resolve this trade off between lower premium and higher out-of-pocket uncertainty.

The results can be different for other countries with different expenditure distributions (e.g. due to a different age profile of the population). The best known analysis of cost-sharing on healthcare expenditures is Newhouse (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 (although the differences between the effects of the different co-insurance rates are small, especially for small maximum out-of-pocket amounts). 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 the co-insurance rate can be thought of as complementary policy instruments.

Figure 16 shows how the effects of two cost-sharing schemes –relative to a 300 euro deductible– vary across ages and gender.[32] The percentage difference does not vary smoothly with age. It is roughly constant and appears to fall slightly with age (becomes more negative) for the shifted deductible. As expenditure itself rises with age, the distribution of healthcare expenditures for older individuals is such that shifting the deductible by 400 euros puts more people at the margin compared to a traditional deductible which starts at zero. This effect is not so strong for younger people and hence the reduction in expenditure smaller for this age group.

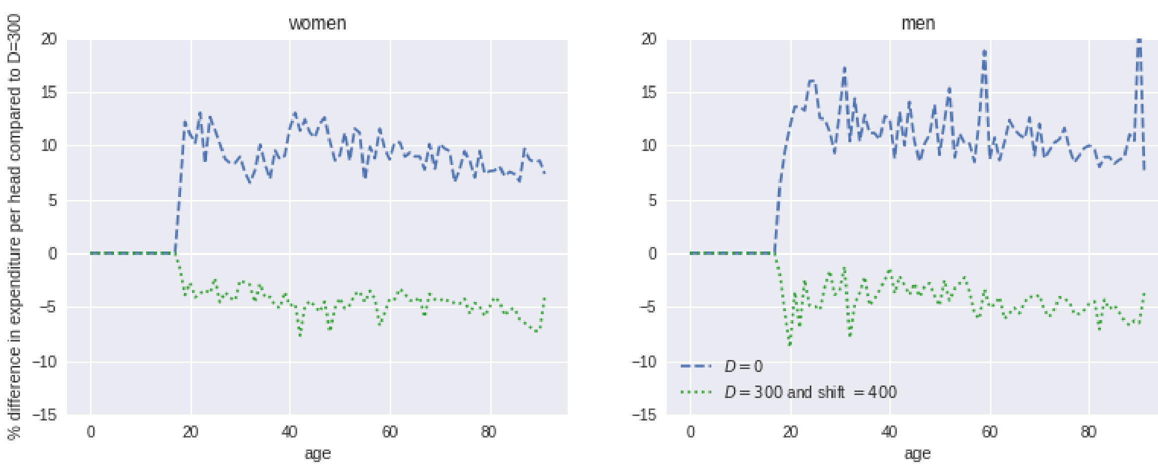

Figure 16: Percentage difference between expenditure across ages and gender

Until now, we have presented average outcomes for different gender-age categories. The advantage

[32]Note that the percentage difference can be slightly different from figure 2 as we compare with average healthcare expenditure for a 300 euro deductible for each age and gender, not with healthcare expenditure averaged across ages and gender.

of using a Bayesian model is that the posterior samples allow us to present the uncertainty associated with these outcomes. To illustrate this, figure 2 shows that compared to expenditure per head with a traditional 300 euro deductible (starts at zero, 100% co- insurance rate) expenditure is lower with either 25% co-insurance or a shifted deductible which starts 400 euro. The figure shows that this is true *on average*. Figure 17 shows that this is not true in each posterior sample. The figure gives the inverse cumulative distribution function of expenditure per head relative to expenditure with a 300 euro deductible.[33] Normalized expenditure below 1 means that expenditure per head is lower than with a 300 euro deductible. Hence, with a zero deductible there is a probability of 1 that normalized expenditures exceed 1: in all samples expenditure with $D = 0$ exceeds expenditure with $D = 300$. With a 400 euro shift there is only a 30% probability that normalized expenditure exceeds 1.0. With 70% probability $(1 - 0.3)$, a 400 euro shift reduces expenditure compared to $D = 300$. With 25% co-insurance, it is highly unlikely that expenditure exceeds expenditure at $D = 300$, but the relative reductions in expenditures tend to be smaller than with a 400 euro shift. In this sense, whereas the shift leads to lower expected expenditures, co-insurance is the safer option in the sense that it basically always reduces expenditures compared to $D = 300$.

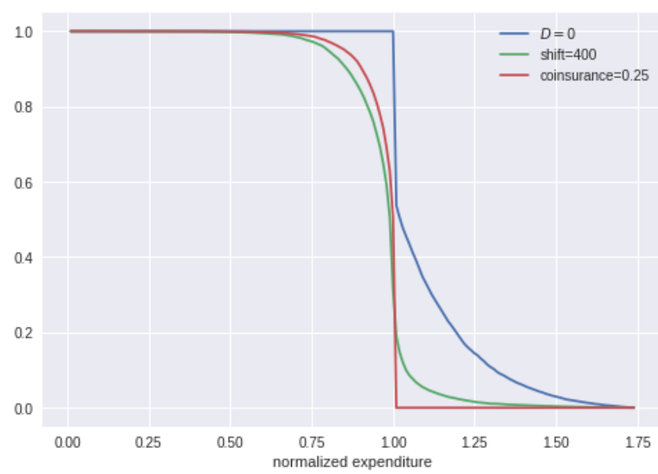

Figure 17: Inverse cumulative distribution functions of expenditures per head relative to a 300 euro deductible.

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 18 considers people above the age of 40. As illustrated in figure 1, average healthcare expenditure increases monotonically from this age onward. Hence, to keep many people at the margin, we expect that the optimal shift increases with age. Doing this for each age separately, ie. for per 40 year olds, 41 year olds, and so on, will

[33]To be precise, for each of the 10,000 samples we first divide the outcome, say, with a 400 euro shift by the outcome with a standard 300 euro deductible. Then we average across samples. Because division is not a linear operation, this is not the same as: dividing the average outcome with a 400 euro shift by the average outcome of a 300 euro deductible.

lead to a noisy picture with an upward trend.[34] However, this is not a serious policy option: when you are 40, the shift is 300 euro, at 41 it becomes 350 and then falls again to 320 at 42. Hence, we consider age-brackets of 10 year: 40 to 50 year olds, 50 year olds to 60 year olds, and so on. Figure 18 shows that the optimal shift increases with the age-bracket; varying from 320 euros to 440 euros and constant at 380 euros for 50 to 70 year olds. Characterizing the fully optimal system (under some sanity constraints) is beyond the scope of this paper and we leave this for further research.

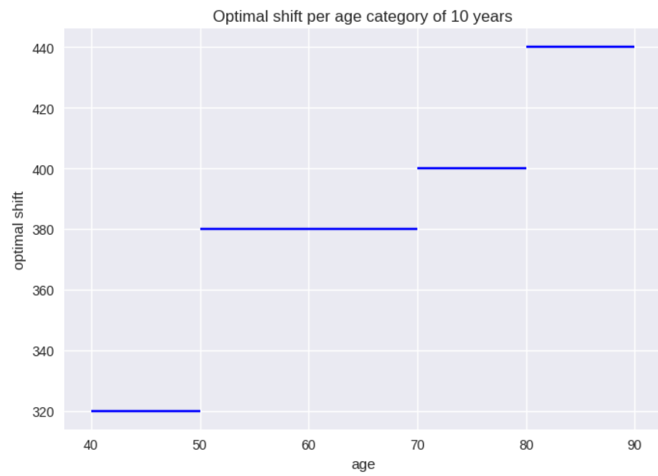

Figure 18: Optimal shift for 10-year age categories

Finally, all simulations until now changed the price of healthcare through $EOOP$. Demand for healthcare is however also affected by the quality of healthccare. 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 micro founded 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[35] in the function of F (equation 7) in the model and assume that cost-sharing does not change. An advantage of our specification of F is that expected quality 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 19 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

[34] As shown in figure 15, around the expenditure minimizing shift of 400 euro, the curve is rather flat. This implies that small changes in parameters can lead to rather big increases or decreases in the optimal shift per age category.

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

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 $(1+g)$), but as the maximum out-of-pocket payment increases, the differences become larger.

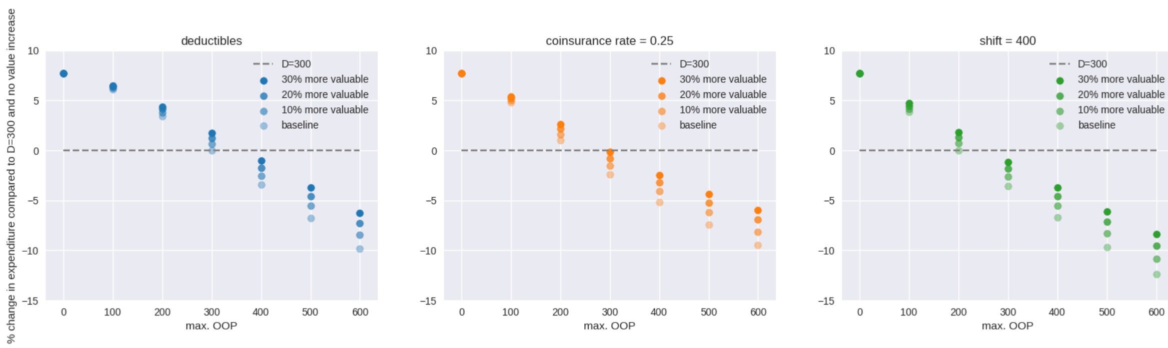

Figure 19: Healthcare expenditures after increasing the value of treatments by 10, 20, and 30 percent

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 excluding 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 have been excluded (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.

Figure 20 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

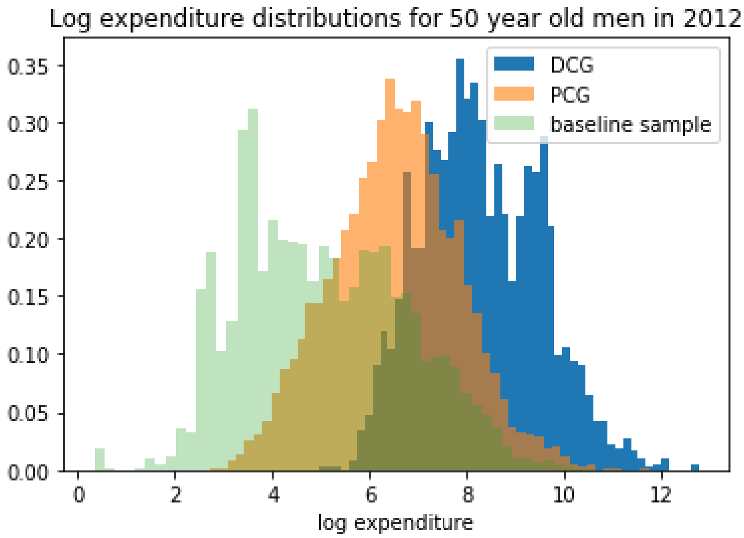

Figure 20: Distributions of log expenditures for our baseline sample, people with FCG and DCG

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 seemsly 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 21 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 mixed in results 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.

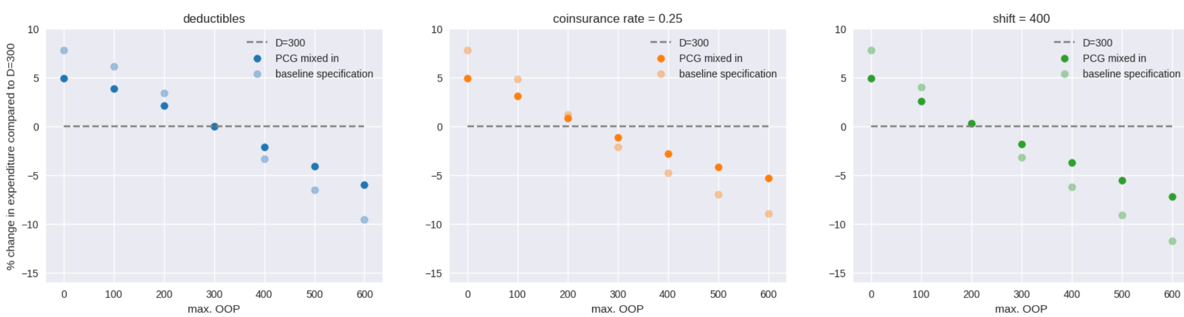

Figure 21: Comparison of the results with the baseline sample and the sample extended with individuals with a PCG

Figure 22 shows the results of estimating the model and simulating separately for persons with a PCG and after mixing the results with the baseline sample. 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.

Whereas figure 18 shows that the optimal shift increases as people have higher expected healthcare expenditures, this is not the case for the PCG group. Although their expenditures are higher, they are also less elastic. But this result can be due to the deductible in our data being rather small compared to the PCG expenditure distribution.

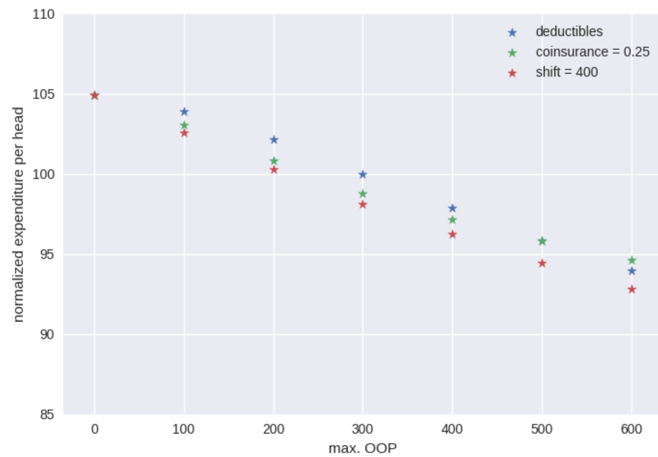

Figure 22: Results when persons with a PCG are added

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 23. With a normal distribution for F the results (darker colors) are similar to the results of the baseline model (lighter colors).[36] With a normal distribution expenditure tends to increase less steeply with a fall in maximum out-of-pocket payment and falls 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 this case, eliminating out-of-pocket payments tends to increase expenditure a lot. This effect is smaller for a normal distribution. As $EOOP$ increases, the density only falls for an exponential distribution but can increase for a normal distribution. Hence, effects of an increase in $EOOP$ with a normal distribution exceeds the effects with an exponential distribution.

Also, figure 24 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.

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

[36]The average deductible elasticity equals -0.09.

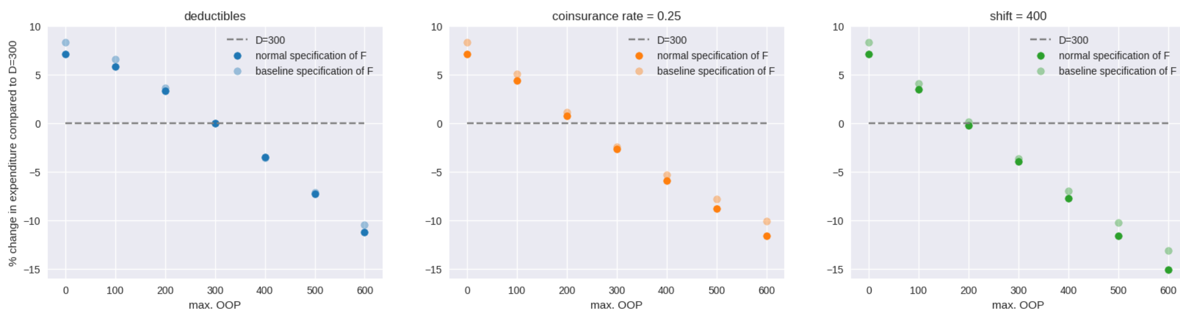

Figure 23: Comparison of the results with the baseline F-specification and the normal F-specification

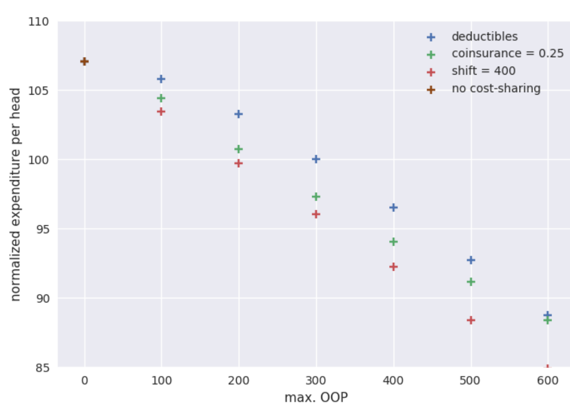

Figure 24: Results for a different specification of F

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. Furthermore, introducing these parameters does not change the results (see figure 25).

Figure 25: 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 are elastic over the policy range. For the Netherlands in 2013, this shift is more effective than a co-insurance rate of 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 for different gender-age categories. As we estimate these distributions, we determine for each scheme the probability that individuals in a 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 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 over the expenditure range between 0 and 300 euro 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 now simply based on gender and age. The categories can be made more homogeneous by 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 euro 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 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. However, from a behavioral point of view, people may interpret the shift as a focal point: I am expected to spend 400 euro (for free) every year. If this were the case, the effect on expenditures would be less favorable. As we do not have a shift in our data, more research is needed to test for this possibility.

References

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. Technical report, Congressional Budget Office Working Paper 2019-08, 2019.

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.

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.

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.

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

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.

Nederlandse Zorgautoriteit. Marktscan en beleidsbrief zorgverzekeringsmarkt: weergave van de markt 2010-2014. , Nederlandse Zorgautoriteit, 2014.

J.P. Newhouse. *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, UNITED STATES, 2005. ISBN 9780262256834.

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 (Februay):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(nil):e55, 2016. doi: 10.7717/peerj-cs.55. URL `https://doi.org/10.7717/peerj-cs.55`.

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.

A Appendix

A.1 Proof of results

Proof of Corollary 1 This follows from writing

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

Q.E.D.

A.2 Data cleaning procedure

To clean the data, observations are omitted when:

- 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 percent of the total number of observations.

A.3 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

A.4 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).

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

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

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

Hence, we specify the corresponding GP as $GP(\eta, \ell)$.

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

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. For ψ_y we do not have this correction for women.

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

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:

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

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[37]

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

where ζ, ν vary with age fixed effects and gender. The parameters of the distribution of x expenditure can be written as:

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

where `year_fixed_effect_x` and $\sigma_{x_{year}}$ denote year fixed effects and $\sigma_{x_{age}}$ age fixed effects. We have an identical structure for μ_y, σ_y.

A.5 Priors

In this section we first present the priors that we use in the main analysis and motivate the choices that we made. Then we show that our main results are robust to different choices of the priors.

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 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 such 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 $\mu = 4.5, sd = 0.7$.

[37]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.

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 $sd = 0.15$; that is, we assume a hierarchical structure here.

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 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" 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 will 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 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 a gender-age category of approximately 5,000 euros.[38] 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 euro 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.

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

priors changed	average expenditure	standard deviation	max. 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

Figures 26 and 27 show how the results of the simulations change when the priors of $sd_{x,y}$ are

[38] Roughly speaking, $Prob((x - 2000) > 5,000 * 100) \leq 1/100^2$ for $x > 500,000$.

increased to 0.2 and when the priors of the standard deviation of the year fixed effects are increased to 1.0 respectively. The results are very similar to the results when using our baseline priors which suggests that these results are robust to different choices of priors.

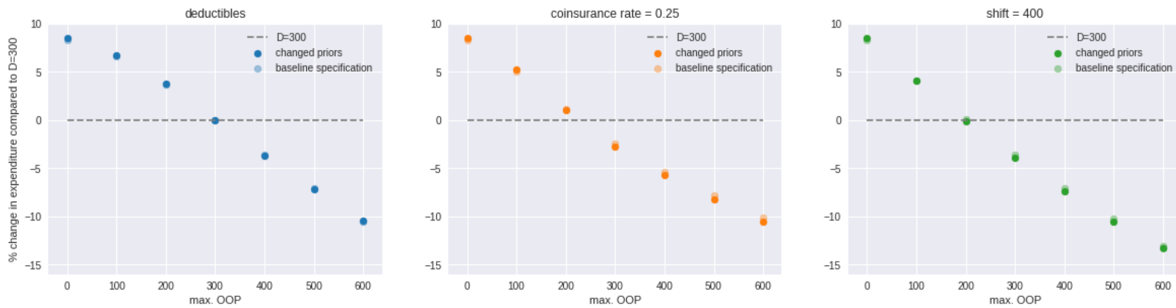

Figure 26: Comparison of the results with the baseline priors and the priors changed to $sd_{x,y} = 0.2$

Figure 27: Comparison of the results with the baseline priors and the priors for year fixed effects changed to $sd = 1.0$

A.6 Additional figures

This section presents plots that are not in the main text.

Also for men we find a clear pattern in log expenditures across age which is stable across calendar years 2008-2013.

The ELBO plot for men also suggests convergence.

Plotting the predicted probability of positive expenditures for women in 2008 against the validation data suggests a fairly good fit.

Also the fit for women is quite good, although for 2013 we under estimate expenditures for women above 60. The KL-divergence is close to zero and for most ages lower for our model than for the true distribution averaged across age in the train set.

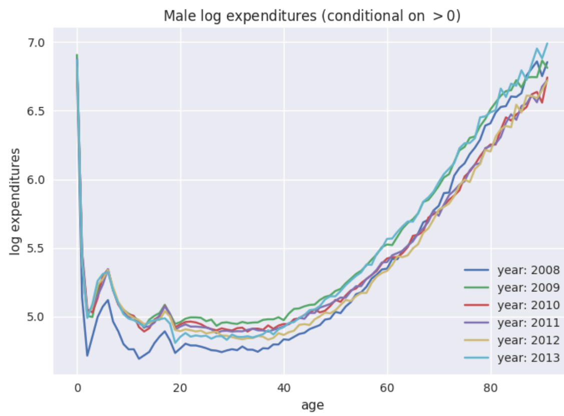

Figure 28: Log healthcare expenditure for men conditional on being positive.

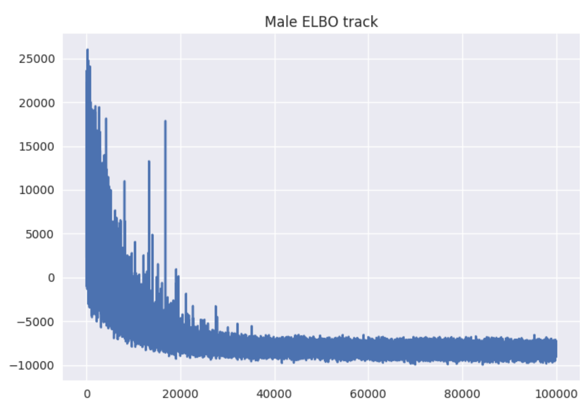

Figure 29: ELBO model estimation for men

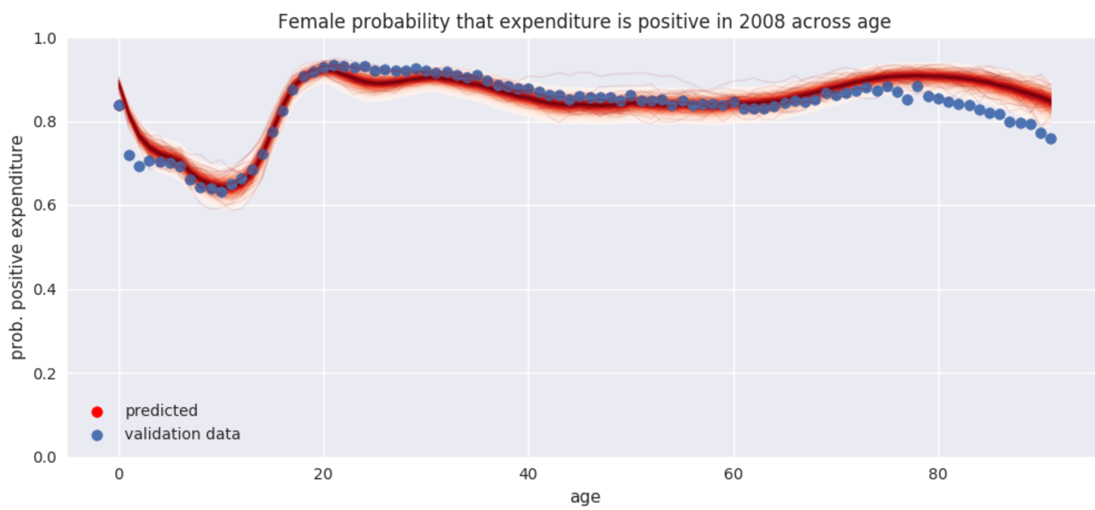

Figure 30: Predicted and realized probabilities of positive expenditures for women across age in 2008.

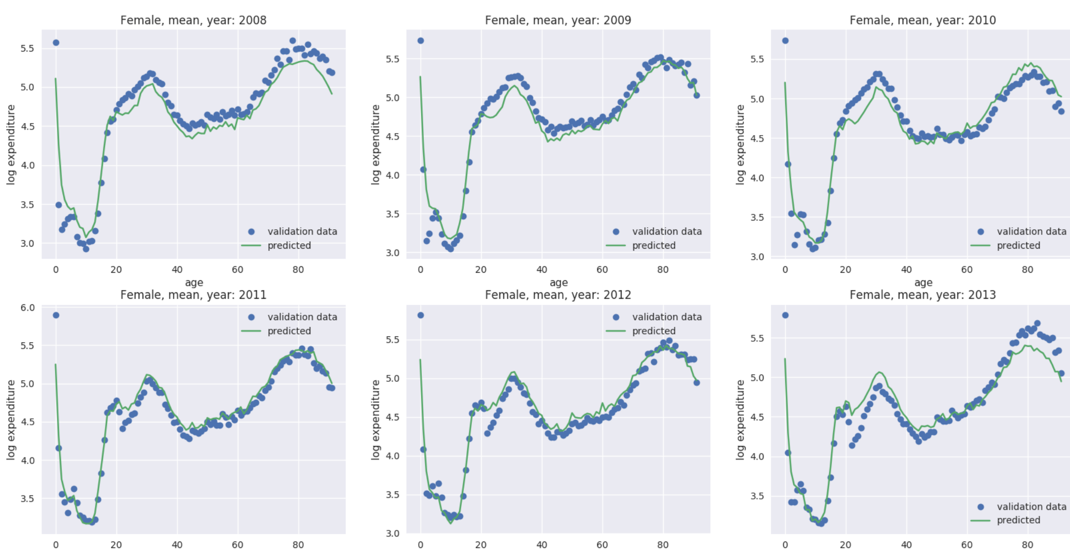

Figure 31: Average predicted vs average validation female log healthcare expenditures for 2008–2013

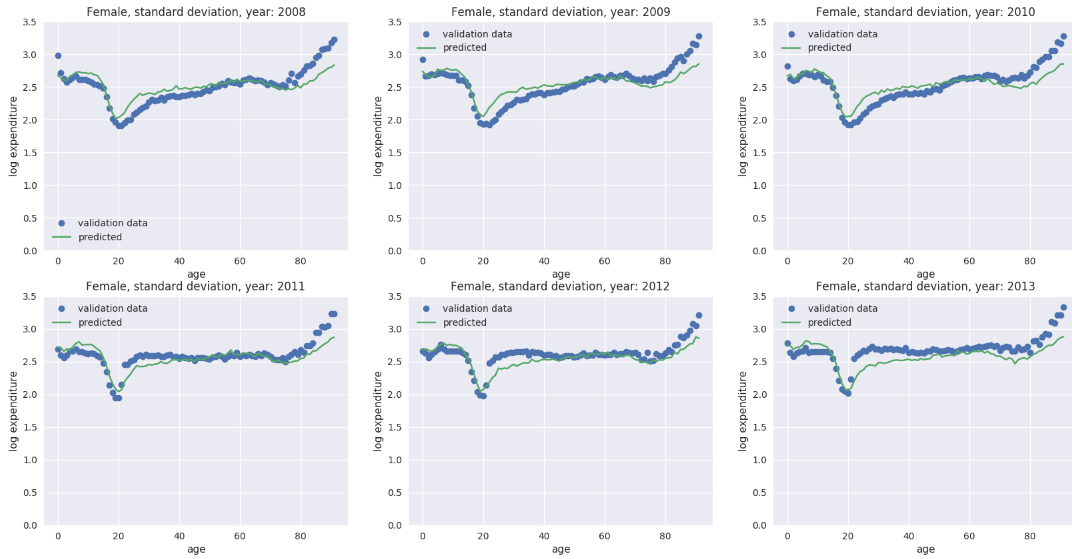

Figure 32: Standard deviation predicted vs validation female log healthcare expenditures for 2008–2013

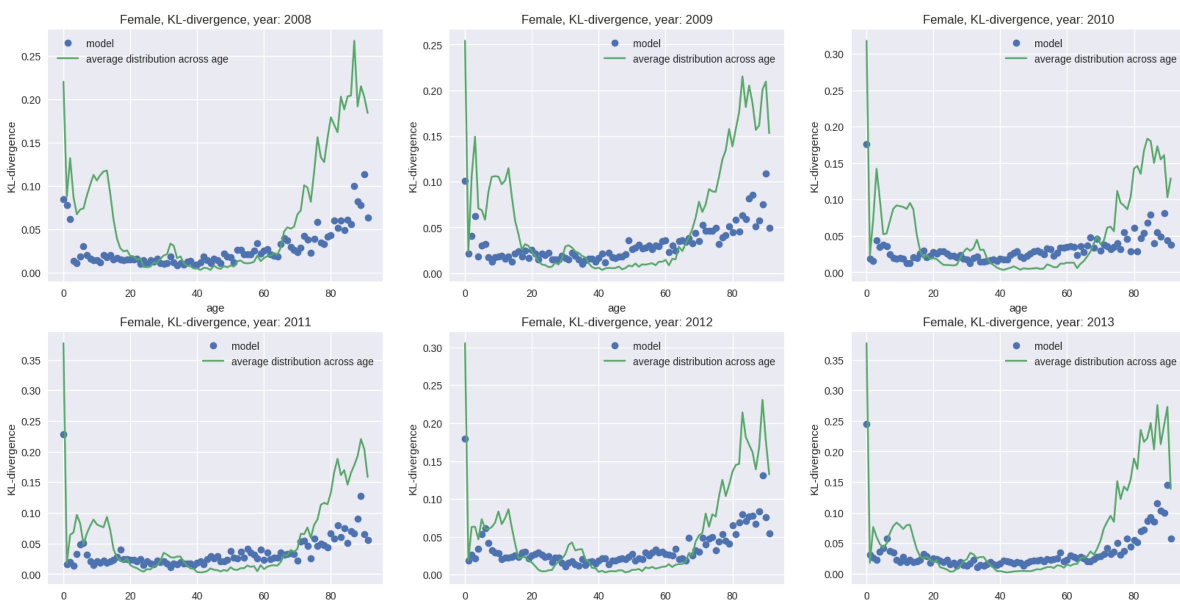

Figure 33: KL Divergence expenditure distribution females for 2008–2013