Journal of Health Economics 97 (2024) 102900

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/jhe

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

Jan Boone [a,b,c,*], Minke Remmerswaal [a,b]

[a] Department of Economics, Tilburg University, The Netherlands
[b] CPB Netherlands Bureau for Economic Policy Analysis, The Hague, The Netherlands
[c] CEPR, London, United Kingdom

ARTICLE INFO

Keywords:
Moral hazard
Out-of-pocket
Shifted deductible
Co-insurance
Bayesian mixture model
Microsimulation model

ABSTRACT

Demand-side cost-sharing reduces moral hazard in healthcare but increases exposure to out-of-pocket expenditure. We introduce a structural microsimulation model to evaluate both total and out-of-pocket expenditure for different cost-sharing schemes. We use a Bayesian mixture model to capture the healthcare expenditure distributions across different age–gender categories. We estimate the model using Dutch data and simulate outcomes for a number of policies. The model suggests that for a deductible of 300 euros shifting the starting point of the deductible away from zero to 400 euros leads to an average 4% reduction in healthcare expenditure and 47% lower out-of-pocket payments.

1. Introduction

Healthcare demand and expenditure are high and rapidly increasing in many (Western) countries due to, among other things, ageing populations. This rapid increase presents a challenge for politicians and policy makers, as it adds further strain to the budget and the labor market, which are often already under pressure. Cost-sharing is one of the instruments available to limit healthcare demand. It reduces moral hazard and curbs healthcare expenditure by shifting part of healthcare costs to users of care (Zweifel and Manning, 2000). As such, it can improve efficiency. However, cost-sharing is a politically sensitive and highly debated topic, because it introduces out-of-pocket expenditure which is mainly borne by (chronically) ill people. Cost-sharing implies a trade-off between the effect on total health expenditure and out-of-pocket expenditure. In light of the (perceived) necessity to curb healthcare expenditure growth, sound decision-making by politicians and policy makers is valuable. For this it is important that they understand the trade-offs associated with different types and levels of cost-sharing in healthcare.

The aim of this paper is to create a structural microsimulation model for demand-side cost-sharing in healthcare which can inform policy makers about the effects of cost-sharing schemes on total expenditure and out-of-pocket expenditures for different age and gender categories. To be able to simulate the effects of cost-sharing schemes including schemes that have not been in place before, we introduce a structural model. Reduced-form estimates, that is, elasticities estimated for a given type of cost-sharing, are relatively easy to model. However, such estimates are often local and their use is, as a consequence, limited. To illustrate, in previous

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

[*] Corresponding author at: Department of Economics, Tilburg University, The Netherlands.
E-mail address: j.boone@uvt.nl (J. Boone).

https://doi.org/10.1016/j.jhealeco.2024.102900
Received 16 October 2023; Received in revised form 20 May 2024; Accepted 24 May 2024
Available online 27 May 2024

research we estimated a deductible elasticity for the Netherlands (Remmerswaal et al., 2023). This elasticity can subsequently be used to predict the financial impact of an increase, reduction or even abolishment of the deductible. A deductible elasticity cannot be used, however, for simulating the effect of a change to another cost-sharing scheme, like co-insurance instead of a deductible.

The distributions of healthcare expenditures are essential to the model in this paper, because they determine the response to a change in cost-sharing (scheme). Whether a patient decides to consume or forego a treatment does not only depend on the actual out-of-pocket payment she has to pay for a certain treatment, but also on other care and other out-of-pocket payments she expects. The introduction of a 100 euro deductible in a healthcare plan for chronically ill individuals for instance is unlikely to affect healthcare expenditure. The expected (out-of-pocket) price of a treatment offered to them is close to zero due to their likelihood of exhausting the deductible — meaning they are not at the margin and more likely to accept a treatment. However, the same 100 euro deductible in a plan for students – who tend to be young, healthy and less likely to reach the deductible – is probably more effective at reducing this group's healthcare expenditure.

In our model, we capture "whether a person is at the margin" by deriving treatments' expected out-of-pocket payments (EOOP) from the distribution of healthcare expenditures. EOOP is calculated by taking the expectation over the change in out-of-pocket expenditure with and without accepting a treatment offered by physicians. Deriving this expectation requires information on the underlying expenditure distributions which we estimate. To capture whether a change in cost-sharing leads to a change in healthcare expenditure, we compare price EOOP with the utility of treatments, where the distribution of treatment utilities is estimated as well.

The distribution of healthcare expenditure is modeled as a mixture of four underlying distributions (zero expenditure and three distributions of positive expenditure). Price EOOP affects the weights on these four components of the mixture. As healthcare distributions vary significantly among individuals, affecting their response to cost-sharing, we address this heterogeneity by explicitly modeling the mixture of distributions for each gender and age group.

After estimating the model on Dutch data, we present policy simulations which lead to interesting results for Dutch policy makers. For the Netherlands we find that the current cost-sharing scheme in our data, which is a deductible of roughly 300 euros, can be improved by shifting the starting point of the deductible by 400 euros. This shift improves the trade-off between total healthcare expenditure and out-of-pocket expenditure by (on average) reducing healthcare expenditure by 4% and out-of-pocket payments by 47% for people without a chronic condition. Although we do not present a welfare analysis in our simulations, the reduction in healthcare expenditure contributes to efficiency by limiting moral hazard. Lowering out-of-pocket payments can be seen as contributing to solidarity between people with differing health status.

To estimate the model we use a proprietary dataset covering the Dutch population over an eight year period. This dataset encompasses yearly healthcare expenditure for each individual along with various person characteristics. The methodology can be applied to other countries where similar data is available to what we use here. We use Bayesian estimation to facilitate working with distributions. This also captures the uncertainty of our results in posterior distributions. We use these posterior distributions throughout the paper, enabling us to illustrate the uncertainty surrounding our simulation results.

In the next Section we explain the model's contribution compared to other models in the literature. The model is described in a more detailed way in the theoretical framework in Section 3. In Section 4, we describe the institutional setting of healthcare in the Netherlands and the data. We explain how we parameterize and identify the model in Section 5, and discuss the estimation methodology and fit of the model in Section 6. Section 7 presents the results of simulating deductibles, co-insurance rates, and shifted deductibles with the model. We conclude in Section 8. The appendices are presented as supplementary online material.

2. Contribution to literature

This paper builds on a literature modeling demand-side cost-sharing. Our approach is comparable to Einav et al. (2013) who develop a model to show evidence of selection on moral hazard. From their paper we borrow the ideas to make a distinction between exogenous and endogenous healthcare expenditures, assume expenditure distributions are lognormal and to estimate the model using Bayesian methods with panel data. The purpose of Einav et al. (2013) is to study heterogeneity in moral hazard and to establish its relationship to health plan choice. We focus on creating a model to simulate the effects of different types and levels of cost-sharing on healthcare expenditure and individual out-of-pocket spending.

Other models for demand-side cost-sharing include Cardon and Hendel (2001) and Bajari et al. (2014). These models are based on a framework/health insurance scheme in which health insurance is not mandatory. As a result, the models include both the decision to buy health insurance as well as how much care a person utilizes given a health plan. Our model is based on a health insurance scheme in which health insurance is mandatory, as we explain in Section 4.1. Further, the coverage of basic insurance is uniform, regulated by the Dutch government and there is minimal variation in coverage across insurers. The advantage of this is that our model is relatively simple. We do not need to estimate an individual's risk aversion to understand whether or not they buy health insurance.

Our paper is also related to the literature on Bayesian estimation of healthcare demand and expenditure. Examples of papers in this literature are Deb et al. (2006), Jochmann and Leon-Gonzales (2004), and Mukherji et al. (2016). These papers model healthcare demand for different reasons than ours: Deb et al. (2006) study the effect of managed care on health care expenditure and Mukherji et al. (2016) study the effects of ageing on health demand. Mukherji et al. (2016) stress the importance of explicitly modeling the nonlinear effects of age interacted with gender on healthcare expenditure. We model these non-linear effects with Gaussian Processes (GPs).

There are a number of papers analyzing the use of mixture models to predict healthcare expenditure distributions. When comparing mixture models with more standard estimation techniques of healthcare use (like generalized linear regression and two-part models) (Jones, 2012; Deb and Holmes, 2000) find that mixture models tend to outperform the standard approaches. These papers do not use mixture models to analyze changes in demand-side cost-sharing. The main differences between our approach and the use of mixture models in this health economics literature (see also Deb and Burgess, 2003; Deb and Trivedi, 1997) are that we use Bayesian estimation (not maximum likelihood), the weights on the components are a function of the (expected) out-of-pocket price, EOOP, the components are log-normally distributed and we estimate a mixture model for each gender–age category. Bayesian estimation allows us to easily present the uncertainty of the simulation outcomes (using the posterior samples), by making the mixture weights a function of EOOP we analyze the effects of different cost-sharing schemes on the overall distribution of healthcare expenditure, assuming a log-normal distribution for each gender–age category captures the data well and simplifies the estimation of our model.

There is also a literature on demand-side cost-sharing in the Netherlands. This paper builds on previous empirical work in which we study the effect of the deductible size for 18 year olds (Remmerswaal et al., 2023). As mentioned, we used a reduced-form approach to estimate an elasticity of the change in the deductible size between 2008 and 2013. We find an average deductible elasticity of −0.09 which is close to the average deductible elasticity we find with our model in this paper (see Section 7). However, such a deductible elasticity cannot be used to simulate the effects of a change in cost-sharing scheme. This is why we introduce a structural model in this paper.

Van Kleef et al. introduced the idea of a shifted deductible and showed its potential for reducing moral hazard. Their paper mainly focuses on the idea and principles of a shifted deductible. Van Kleef et al. (2009) estimate the optimal shift, which they define as the shift which maximizes the variance of out-of-pocket expenditures. They argue that price sensitivity is highest when uncertainty is greatest. In contrast, we do not focus on finding the optimal shift of a deductible, but on simulating the effects of a number of cost-sharing designs on both total healthcare and out-of-pocket expenditures. Further, we model the effect of cost-sharing, and a shift of the deductible, on healthcare expenditure through EOOP: the expected out-of-pocket payment of a treatment offered to a person. Van Kleef et al. (2009) find that the optimal starting point is not zero (so not a traditional deductible), but positive for all individuals. This coincides with our findings, although we tend to find somewhat lower starting points.

Cattel et al. (2017) also compute the effectiveness of a traditional deductible and a shifted deductible on reducing healthcare expenditure. They argue that the effectiveness depends on two parameters: (i) the probability that an individual's healthcare expenditures end up in the deductible range and (ii) the total expected healthcare expenditures given that they end up in that range. The idea of this approach is similar to ours, however our model allows for more heterogeneity as we model the healthcare expenditures distributions more explicitly as well as separately for gender–age groups.

We contribute to the work of Van Kleef et al. (2009) and Cattel et al. (2017) by analyzing more diverse cost-sharing schemes and by quantifying how much a shifted deductible can reduce healthcare expenditures and out-of-pocket payments both at an individual level as well as at the population level.

The recent working paper by Klein et al. (2023) also simulates the effects of cost-sharing schemes on healthcare expenditures. Their approach to compute the expected out-of-pocket price differs from ours. In particular, they use one year of data to analyze how expenditure varies during the year as people reach their deductible. Although they use Dutch data as well, their method tends to find higher deductible elasticities than reported here.

3. Theoretical framework

In this Section we explain how we model healthcare expenditure and how (a change in) cost-sharing affects it. A full specification of the model that we estimate, can be found in Appendix D.

The theoretical framework needs to answer two questions. First, given other healthcare expenditures an agent has, what is the (expected) out-of-pocket price of a treatment suggested by a physician? Second, given the out-of-pocket price, is the treatment worth it for the patient? To answer the latter question, we estimate the utility distribution of treatments. To answer the former, we model the distribution of healthcare expenditures per capita per year. When there is a change in cost-sharing, it changes the effective out-of-pocket price of an offered treatment, thereby affecting the probability that this treatment is accepted.

To illustrate, consider a person who is offered a medical treatment. When the deductible level increases, it generally means that she will pay more for the same treatment out-of-pocket. But if this person has an expenditure distribution with high medical expenses, such that she is likely to reach the new deductible level with other treatments regardless, then the extra amount she has to pay out-of-pocket for the offered treatment is minimal or even zero. Hence, for people with high expected costs, an increase in the deductible has a smaller effect on the out-of-pocket price of a treatment than for people with low expected costs.

We first explain the two components of healthcare expenditure that we use and how this distinction leads to a mixture model. From the mixture model we derive the expected out-of-pocket (EOOP) price of a treatment. If the utility of treatment exceeds the EOOP price, the patient accepts treatment.

3.1. Expenditures

In our model, the distribution of total healthcare expenditure per capita per year is generated by two types of treatment: exogenous treatments, denoted by X, and endogenous treatments, denoted by Y. The former are exogenous to cost-sharing because they consist of high value procedures that are always carried out in practice, irrespective of the size of cost-sharing (given the relevant Dutch policy range). One can think of setting a broken bone or immediate hospitalization following a stroke or cardiac episode. The latter are endogenous with regard to the deductible, meaning individuals consider the deductible when deciding whether to undergo

Table 1
The distribution of total log expenditure z.

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

Notes: The table lists the four components with each of the four probabilities the mixture distribution z consists of.

such treatments. For example, if a physician proposes a patient additional imaging services, individuals may choose to accept or decline based on their deductible.

It is important to note that we do not predefine certain healthcare expenditures as either X or Y. Rather the categorization arises from the observation that certain parts of the expenditure distribution vary with the deductible level, while others remain unaffected.

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.

We assume that healthcare expenditures, conditional on being positive, are lognormally distributed. We transform our observed healthcare expenditures Z from euros into logs: $z = \ln(1 + Z)$. As a result, $z = 0$ if $Z = 0$; otherwise $z > 0$. As we assume that Z conditional on Z being positive ($Z|Z > 0$) has a lognormal distribution, z conditional on z being positive ($z|z > 0$) is normally distributed. Note that we need to be careful when moving back from logs to euros to minimize bias in our simulated outcomes (Jones, 2012). We do not use a smearing factor in our calculations for two reasons. First, we actually do not transform moments of the distribution from logs into euros. Our Bayesian analysis gives us samples of the (posterior) distribution of the outcomes and these samples can be translated back to euros. Second, the results that we present in Section 7 are in relative changes (compared to a standard deductible of 300 euros) which can be derived in log space.

We model the probability of an individual being offered an exogenous x treatment as ψ_x, indicating a positive draw of $x = \ln(1 + X)$, where X is in euros. Once offered, this treatment is accepted, since it does not depend on the deductible by definition. The probability of an individual being offered an endogenous treatment $y = \ln(1 + Y)$ is represented by ψ_y. The probability that this treatment is rejected is denoted by F, as explained shortly.

We assume that the Bernoulli draws with probabilities ψ_x and ψ_y are independent: x and y are independent in logs. But there is a dependence in levels (in euros), because of the log specification $z = x + y$. Addition in logs implies multiplication in euros, which generates the intuitive effect that someone with high X expenditures in euros tends to have access to high endogenous expenditures and thus (likely) high total expenditures in euros.[1]

Below we show how the exogenous x and endogenous y treatments form total healthcare expenditure z as a mixture model. We explain two major computational benefits from the assumption of lognormal distributions, and in Section 4.4 we show that this lognormal assumption is a reasonable representation of the data.

3.2. Mixture model

Table 1 presents how exogenous expenditures x and endogenous expenditures y produce four possible outcomes of total (log) healthcare expenditure z.

A first possible outcome of z is when expenditure is zero: $z = 0$. In this case, a person has neither x nor y treatment: $x = y = 0$. The probability that this happens is determined by multiplying the probability of $x = 0$ (which is $1 - \psi_x$), by the probability of $y = 0$. The latter can transpire in two scenarios: either no y treatment was offered $(1 - \psi_y)$, or it was offered, but rejected, represented by $\psi_y F$, where F is the probability that the offered y treatment was rejected.

Another possible outcome of z is that z is positive because both x and y are positive. The probability of this outcome is the multiplication of the probability that $x > 0$, given by ψ_x, and the probability that $y > 0$ by $\psi_y(1 - F)$, meaning y is offered and accepted. Hence, the probability that both $x > 0$ and $y > 0$ is given by $\psi_x \psi_y(1 - F)$. The other two possible outcomes are variations on this. Note that the weight on the x component, ψ_x, does not vary with $EOOP$ (in this sense it is exogenous). The weight on the y component, $\psi_y(1 - F)$, varies with F which depends on $EOOP$ (see below).

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

[1] As we define our expenditure variables as $z = \ln(1 + Z)$, $z = x + y$ implies $Z = X + Y + XY$ in euros. In our data, the 1 euro is small compared to expenditure levels X, Y, Z. Hence, we think of $Z = XY$ as this is the dominant term in the expression for Z.

3.3. Out-of-pocket payments

The previous section described how we model the distribution of healthcare expenditure as a mixture model consisting of four underlying distributions. By taking the expectation over the mixture of distributions of healthcare expenditures we can we derive the expected out-of-pocket payment (*EOOP*) of a treatment offered to a person by her doctor.

EOOP represents the out-of-pocket payment one expects to pay for an offered endogenous y treatment, given other exogenous x expenditures. A more strict or literal interpretation of *EOOP* is as follows: at the start of the period, before exogenous expenditure x has been realized, a person is offered an endogenous treatment. The exact price of this y treatment is unknown at the time of offering, for example because the doctor advises the patient to see a dermatologist, but they do not know which treatment will be needed exactly. Nor are the exact exogenous x treatments known that may be needed later in the year. Therefore, we model x and y as distributions and $EOOP$ as an integral over both x and y for each gender–age category.

Broadly speaking, *EOOP* captures that the out-of-pocket price of y treatments is lower for persons expecting high medical expenditures in a given year. They are likely to reach the deductible anyway and will exhibit lower responsiveness to a deductible change.

This approach, using $EOOP$ as the effective price that individuals face, is similar, yet different, to using the end-of-year price (Keeler et al., 1977; Ellis, 1986). The end-of-year price in a deductible scheme can be estimated as the probability that a person does not reach the deductible until the end of the contract period, which is a year in the Netherlands. It captures how much one more euro of healthcare would cost for a person. Whereas the end-of-year price represents the (marginal) cost of utilizing *one more euro* of healthcare, $EOOP$ is the expected cost of accepting an *additional treatment*. *EOOP* would equal the end-of year price, if there were no distinction between exogenous and endogenous expenditures, and if a y treatment costs one euro only.

Although there is some evidence that people respond to spot prices and not to end-of-year prices because they are myopic, discount future costs, or lack information (Brot-Goldberg et al., 2017; Remmerswaal et al., 2019), we assume that they have correct expectations about the distribution of healthcare expenditures they face. Their $EOOP$ is calculated by taking the expectation across their expenditure distributions. This assumption aligns with other models, like Einav et al. (2015), that assume individuals are forward-looking. Moreover, Klein et al. (2022) have identified forward-looking behavior in the Dutch health insurance context. Additionally, in Section F, we test the validity of our assumption by allowing for behavioral bias in the specification of $EOOP$. It turns out that this does not affect the results.

We do not allow for a range of y distributions. This choice is based on the observation that the average y expenditures turn out to be relatively low. Further, the fit of our model is already quite good, as demonstrated below. Therefore, there is no need to expand the model by generalizing the way y treatments are drawn.

Below we use the following well known result:

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

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

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

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

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

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

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

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

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

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

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

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

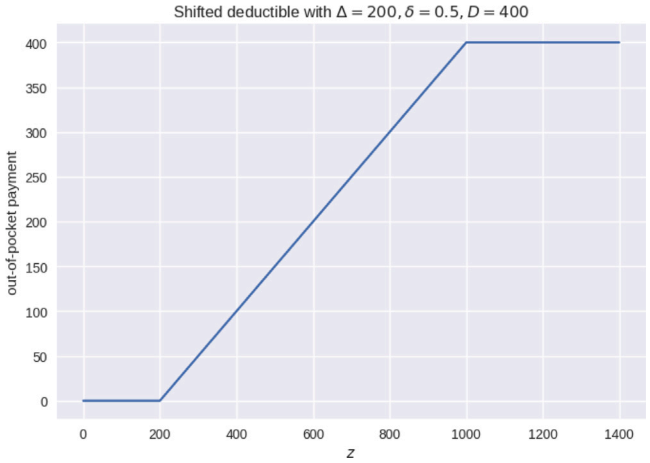

Fig. 1. Out-of-pocket payments with a shifted deductible as a function of expenditure Z. The figure shows how out-of-pocket payments (in euros) depend on expenditures Z (also in euros), given a cost-sharing scheme with shift of 200 euros ($\Delta = 200$), a co-insurance rate of 50% ($\delta = 0.5$) and a maximum out-of-pocket payment of 400 euros ($D = 400$). This cost-sharing scheme is an example of a scheme as described in Eq. (5).

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

The generalized expression for the OOP is:

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

Corollary 1. *The generalized version of $OOP(\mu, \sigma)$ can be written as*[2]

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

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

Fig. 2 illustrates how shifting the deductible can bring people closer to the margin. The left panel of the figure shows the (estimated) density function for X expenditure for 20 and 80 year old women. As one would expect, for 20 year olds there is more mass on zero and low expenditures. The density for 80 year olds is higher from approximately 330 euros onward. Consider a 20 or 80 year old woman being offered one Y treatment with a price of 100 euro and the deductible equals 100 as well. Then only people with zero X expenditure face the full price of 100.[3] For everyone else the price is lower. For people with $X = 90$, the price equals 10 and $X \geq 100$ implies Y treatment is free.

Now suppose we shift the starting point of the deductible with 50 euros (dashed vertical line in the figure). For people with $X = 0$ this implies an OOP price of 50 and for people with $X = 50$ the OOP price equals 100. At $X \geq 150$, the Y treatment is free (150 is the second dashed vertical line). As the right panel of the figure shows, a shift of 50 maximizes the OOP price for 20 year olds. That is the reduction in price – due to the shift – for low X is compensated by the increased price for $X > 50$. As there are

[2] See Appendix A for proof of Corollary 1.

[3] To simplify the exposition, we are a bit loose with terminology here. In the model there is no person with zero expenditure; for each individual there is the distribution of X over which we take an expectation. Similarly, there is no fixed price for Y; there is a distribution for Y expenditure.

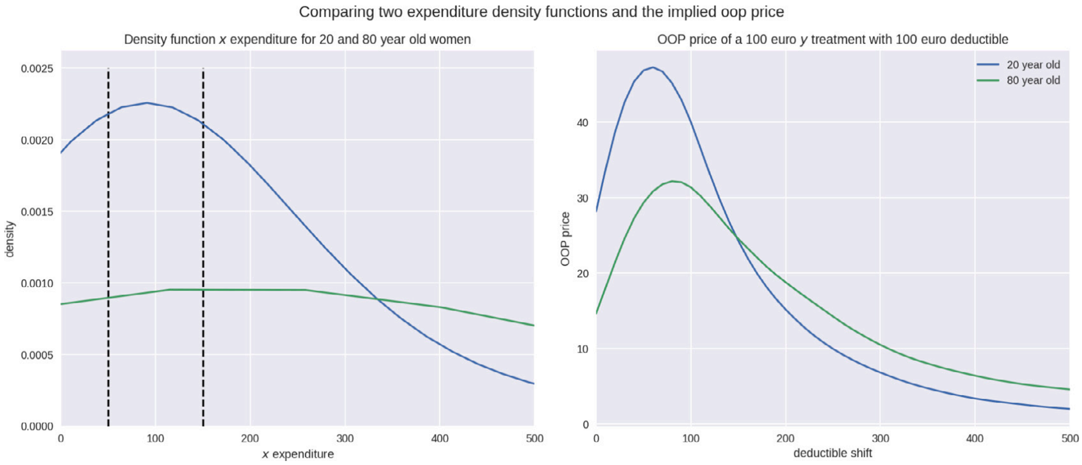

Fig. 2. Illustration of how "being at the margin" affects the out-of-pocket payment. The left panel shows the estimated density function for X expenditure (in euros) for 20 and 80 year old women. The right panel shows how the simulated OOP price of a 100 euro Y treatment given a 100 euro deductible varies for different levels of the shift of the deductible.

more people for whom the price increases due to the shift, we say that the 50 euro shift puts more people at the margin. Shifting the deductible by more than 50 euro reduces the EOOP for 20 year olds: more people get the Y treatment for free than people facing a higher price due to the shift.

As 80 year olds expect higher X expenditure, a bigger shift (around 80 euros) maximizes their OOP price. There are few people with low X and hence it is "cheaper" (compared to 20 year olds) to increase the size of the shift and more people with higher X expenditure are put at the margin of the (shifted) deductible.

Finally, manipulation of the terms in Eq. (4) allows us to introduce behavioral biases in an agent's decision making. As an illustration, Section F.3 estimates this equation with the terms $q_x\sigma_x, q_y\sigma_y$ instead of σ_x, σ_y. Allowing the "behavioral parameters" q_x, q_y to be smaller than one captures that people may under-estimate the standard deviation of a distribution. In the extreme where $q_x = q_y = 0$, people decide on the basis of mean expenditure only; ignoring the uncertainty of the expenditure distribution. Alternatively, one can consider replacing the ψ_x parameter with $q_x\psi_x$. Myopia is then captured by $q_x < 1$ where excessive weight is put on the spot price of the y treatment.

3.4. Accepting a treatment

With $EOOP$ determined, we can model F, the probability that an offered y treatment is rejected. To model this we assume that the utility of y treatments is distributed with a cumulative distribution function $F(u)$ for $u \geq 0$. For a given $EOOP$, a person rejects treatments with utility u below $EOOP$. The probability that this happens is given by $F(EOOP)$. This function captures our utility structure: $F(u)$ denotes the probability that treatment utility is less than u. F is given by a distribution from the exponential family:

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

where $\nu > 0$ and $\zeta \in \langle 0, 1]$. The parameters capture the price responsiveness of healthcare to changes in cost-sharing. The parameters in this specification of F can be interpreted as follows: we assume that the hazard rate is constant, $\nu = f(u)/(1 - F(u))$, 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 utility. This captures travel to and waiting time at a provider or side effects of a treatment. As $EOOP$ goes to plus infinity, the treatment is rejected with probability 1: no y treatment generates infinite utility.

4. Data and setting

To illustrate how the model above can be used for policy simulations, we estimate it using Dutch data. We begin this Section by explaining the institutional setting of healthcare in the Netherlands from 2008 to 2013. Following this, we delve into the specifics of our dataset.

4.1. Institutional setting

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

Table 2
Deductible levels in the Netherlands for 2008 to 2013.

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

Curative healthcare is organized at the national level and operates under a system of regulated competition, as outlined by van de Ven and Schut (2008). Here, health insurers engage in negotiations with healthcare providers and they contract care on behalf of their clients. Health insurance is mandatory for individuals aged 18 and above. Each person is required to purchase insurance from one of the health insurers at a community-rated premium (insurers are not allowed to risk rate.) Those under 18 years old are automatically insured. They do not have pay a premium nor do they face cost-sharing. The government regulates the market to ensure access to healthcare and protect risk solidarity. For example, government regulation precludes health insurers from denying coverage to any individual. An elaborate risk adjustment scheme compensates health insurers for healthcare costs of their clients. The government also dictates the coverage of the mandatory basic benefit package and sets the (minimum) level of cost-sharing, which is the same for everyone. Throughout the study period, alterations to the basic benefit package occurred, as well as some other policy changes. These are summarized by Remmerswaal et al. (2019).

Healthcare costs are funded through three sources. The first source are the health insurance premiums paid by each inhabitant aged 18 and above. These premiums range between 1000 to 1250 euros annually. Individuals with low incomes receive an income-dependent subsidy to offset these costs. The second source are taxes, with payments varying based on income levels. 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). Introduced in 2008, this deductible has been increased by the government multiple times, as outlined in Table 2.

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

Certain health services in the basic benefit package are not subject to cost-sharing. These services, which include primary care, general practitioner care, maternal care and obstetric care, constitute a small portion, approximately 8%, of total costs within curative healthcare. Cost-sharing does apply for health services such as hospital care, physiotherapy, and pharmaceutical care.

Health insurers offer supplementary insurance on top of health insurance for care in the basic benefit package. Such supplementary insurance policies cover other health services than the basic benefit package, such as contact lenses and glasses, alternative medicine, extra dental checks, and cosmetic surgery. Supplementary insurance is therefore an addition to regular insurance, not a substitute. Supplementary insurance is optional and can be bought from a different insurer than insurance for the basic benefit package. Over 85% of the Dutch population bought supplementary health insurance in our study period (Dutch Healthcare Authority, 2014). Importantly and unlike other countries, it is not possible to cover mandatory deductibles in the Netherlands through supplementary insurance nor to get faster access to treatment covered in basic insurance. In this sense, the basic and supplementary markets are orthogonal. In this paper, we only study the basic insurance market and not the supplementary insurance market.

4.2. Data

Proprietary healthcare claim data are used to estimate our model. The data include all claims of Dutch inhabitants from 2006 to 2013. The data have been collected by Dutch health insurers and assembled by Vektis. The data have been pseudonymized, are not publicly available, and do not suffer from underreporting of healthcare claims, because 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 then bill the patient (provided the deductible is not yet reached).

Our analysis employs the same dataset and cleaning procedure as outlined in Remmerswaal et al. (2019, 2023) (see also Appendix B). We exclude data for years 2006 and 2007, because another cost-sharing scheme, known as a no-claims rebate, was in place during those years. After cleaning and exclusion of the aforementioned years, the train data comprise over 58 million observations.

The primary variable of interest in our dataset is the total healthcare expenditure per Dutch inhabitant. This expenditure can be separated into 21 healthcare categories, such as general practitioner care, maternity care, hospital care, and mental care. Given our focus on the impact of cost-sharing, our expenditure variable specifically encompasses cost categories that fall under the deductible (refer to Appendix C for details). All variables in the data are only available at a yearly level. Consequently, we lack information regarding how healthcare expenditures evolve within the year.

Several person characteristics are available, such as gender, age, indicators of chronic use of care, and chronic use of medication, and a person's annual choice of a voluntary deductible. Age is available in years and registered for December 31st in every year, meaning an individual born on, for example, December 1st in 1963 is classified in our data as 50 years old in 2013, even if they were 49 years old for 11 months that year. To make sure we have enough observations per gender–age category, we pool everyone older than 90 years in age category 91. We use DCG to abbreviate diagnosis cost group ('diagnosekostengroep'); this is an indicator for chronic illness and high healthcare costs in previous years. Similarly, PCG is an abbreviation of pharmaceutical cost group ('farmaciekostengroep'), which indicates chronic use of medication.DCG and PCG are variables from the Dutch risk equalization system, which aim to identify chronic disorders that are correlated with high healthcare expenditures. Lastly, the data contain an individual's annual choice of a voluntary deductible.

Table 3
Summary of the baseline sample.

	Women	Men
Age (mean)	34.63	33.72
Number of observations	16,367,197	16,117,158
Fraction of positive expenditures	0.81	0.69
Expenditure (mean)	750.53	571.48
Expenditure (std. dev.)	2667.00	2963.61
Log expenditure (mean)	4.42	3.58
Log expenditure (std. dev.)	2.62	2.80

Notes: The train dataset was used to produce the figures in table. This train dataset will be explained in Section 4.5.

4.3. *Sample and selection*

The model is not estimated on the complete dataset, but rather on a baseline sample. Essentially, our main dependent variable of healthcare expenditure covers costs of health services for which the deductible applies, with the exception of dental costs. To estimate our model, we focus on the sub-population where we expect a behavioral response to a change in the mandatory deductible. As explained below, this implies that we drop from our data individuals with mental healthcare expenditure, with a voluntary deductible and people who are chronically ill. This is approximately 40% of our data. For our simulations we show how these individuals can be reintroduced to derive policy implications for the whole population.

First, all persons with mental healthcare expenditures are excluded, because additional co-payments for mental healthcare were introduced in 2012. As this interacts with the deductible (before and after 2012), it complicates the derivation of the expected out-of-pocket expenditure. Furthermore, dental costs are excluded from the baseline sample due to changes in the dental care coverage between 2008 and 2010, which complicates year-to-year comparisons. Given that dental costs are relatively low, omitting these costs minimally affects the analysis.

Second, we exclude from the data persons labeled with a DCG of PCG, meaning they are chronically ill or chronic users of medication. 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. In particular, 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. Even if we would include them as separate groups in our sample, we would not be able to identify the model's parameters for these groups as their $EOOP$ hardly varies over time (close to zero in each year).

Third, the baseline sample only contains individuals who never chose a voluntary deductible in our sample period. As shown in Remmerswaal et al. (2023) there are significant selection effects for this group while they hardly react to changes in the mandatory deductible which we analyze here. The former suggests that we cannot include them in the years where they go without voluntary deductible and hence we leave them out for the whole period. The latter suggests that they need not be included to estimate the response to a change in the deductible.

A potential concern could be that our policy conclusions would be invalidated because many people would drop their voluntary deductible in response to the policy change. However, our policy proposal (shift in the deductible) reduces out-of-pocket expenditure making it unlikely that people would drop their voluntary deductible in response to it (as may happen in reaction to a policy change increasing out-of-pocket expenditure).

In Section F, we assess the sensitivity of our sample selection to the results.

4.4. *Descriptive statistics*

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

Table 3 provides a first insight into the skewness of healthcare expenditure in the dataset. Further illustration is offered by Fig. 3. The figure shows how well a lognormal distribution fits healthcare expenditure (conditional on being positive) in the data. The histogram presents the raw data, displayed both in levels on the left and logarithmically on the right. The plotted line denotes an estimated lognormal distribution (for just this distribution; not our estimated model). While the fit is reasonably accurate, it is not perfect. It effectively captures the skewness observed in the distribution of healthcare expenditure. Interestingly, the right panel suggests that a better approximation could be achieved by employing a mixture of normal distributions, which aligns with our estimation approach.

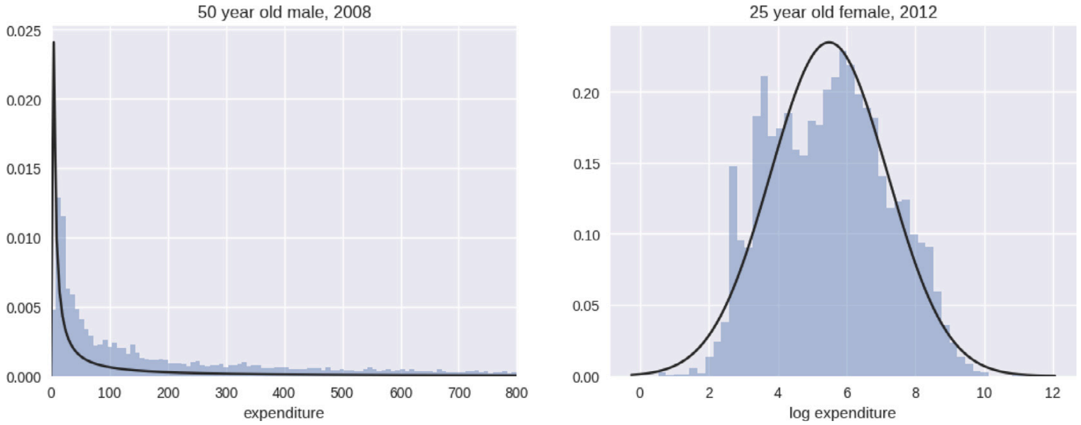

Fig. 3. Two illustrative distributions for positive healthcare costs. The left panel is a histogram of observed expenditure in euros, conditional on being positive, of 50 year old males in 2008. The right panel is a histogram of log expenditures, conditional on being positive, of 25 year old women in 2012. Both panels include a line to denote the fit of a lognormal distribution. The figures are based on raw data (in the train dataset).

4.5. Train, validation and test data

In this study, we adhere to the machine learning convention of dividing the data into a train, validation and test dataset (McElreath, 2020). Splitting the data into these datasets prevents overfitting of the model. The train dataset is used for parameter estimation of the model. With the model and the estimated coefficients of these parameters, outcomes are predicted, which are compared to the validation (and ultimately test) data. If the fit is deemed inadequate, adjustments can be made to the model, and the process can be iterated. We applied stratified sampling to make each dataset representative in terms of age, gender and years. The test and validation datasets each comprise 20% of the total data and the train dataset the remaining 60%.

5. Econometric specification

This section elaborates on the parameterization and identification of the model. Appendices D and E offer a more detailed outline of the model's specification and the choice of priors.

5.1. Parameterization

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

We expect age to have an effect on components x and y, and therefore we model parameters $\mu_x, \mu_y, \psi_x, \psi_y$ as Gaussian processes (GPs) with age. To illustrate this choice, consider Fig. 4 which shows female log healthcare expenditures (conditional on being positive) across ages and for different years. The graph reveals a clear and stable age pattern across the years. The GP captures this pattern by assuming that expenditures are more similar for, say, 20, 21, and 22 year olds, than for 20 and 50 year olds. We assume that the covariance decreases with the square of the age difference. More specifically, the covariance between, say, $\mu_{x,a}, \mu_{x,a'}$ for two different ages a, a' is – up to a constant – given by $e^{-0.5(a-a')^2}$ (Rasmussen and Williams, 2005).

To capture the relatively small annual changes in basic insurance coverage we allow for year fixed effects in μ_x and μ_y. We model these changes as year fixed effects because they can be different from one year to the next and do not necessarily follow a coherent pattern (as they would when modeled as GP). Because of these year fixed effects, we cannot assess the fit of the model by estimating the model on data from 2008 to 2012 and then test the fit on 2013.

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

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

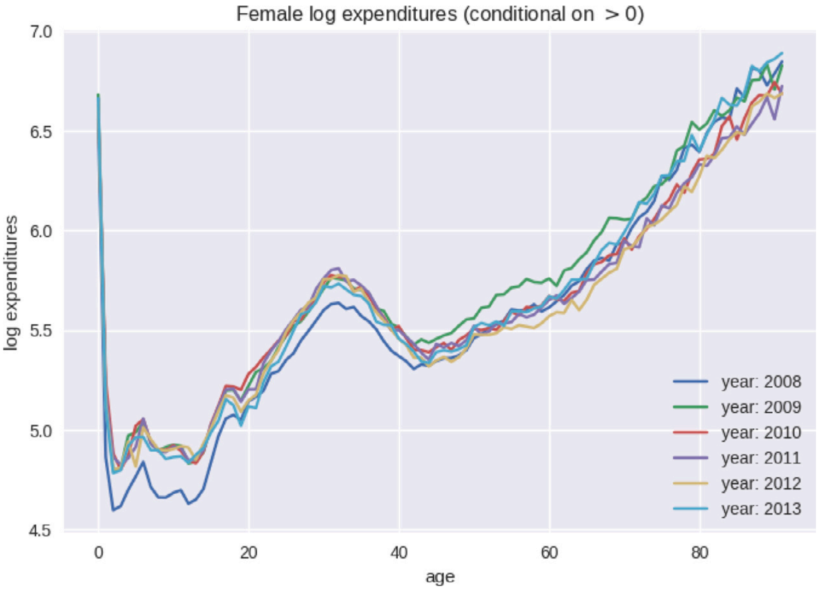

Fig. 4. Log healthcare expenditure for women conditional on being positive. The graph depicts the average log expenditure (conditional on being positive) for women at each age and in each year. The averages are computed based on raw data from the train dataset.

The standard deviation of expenditure z does not show a clear pattern across age. As illustrated in Figure G.5 in the appendix, the standard deviation for men randomly fluctuates around approximately 2.7. We model the standard deviations σ_x, σ_y as age fixed effects, not as a GP.

As mentioned, 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 weighs the effect of cost-sharing for these 18 year olds. We see α as the probability that a person faces a deductible when deciding on the y treatment. If birthdays are uniformly distributed over a year, we expect α to be around 0.5.

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

5.2. Identification

To identify the parameters described above, we use data on healthcare expenditures at the individual level for years 2008 to 2013. We start from the idea that for a given gender–age–year combination the distribution of (log) expenditures can be approximated by a mixture of four distributions: expenditure equals 0, an x distribution, a y distribution and an $(x + y)$ distribution. Such mixture models are well known (see, for instance, McElreath, 2020; Gelman et al., 2013; Jones, 2012).

What we add to this is that the weight on the y distribution varies with $EOOP$. It is this variation in $EOOP$ that allows us to identify the parameters ζ, ν of the probability F that a y treatment is rejected. We have three main sources of variation for this identification.

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

Second, there is (cross-sectional) variation in the healthcare distributions among different gender–age categories even when they face the same (mandatory) deductible per year. For example, the healthcare expenditure distribution of an 80 year old man differs greatly from the healthcare expenditure distribution of a 25 year old man. The former has higher expected expenditures than the latter and we expect him to have higher x expenditures as well. This is illustrated in Fig. 5: the (posterior) probability that ψ_x for 25 year olds exceeds ψ_x for 80 year olds is basically zero. This makes x treatments for 80 year olds more likely and hence $EOOP$ lower.

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

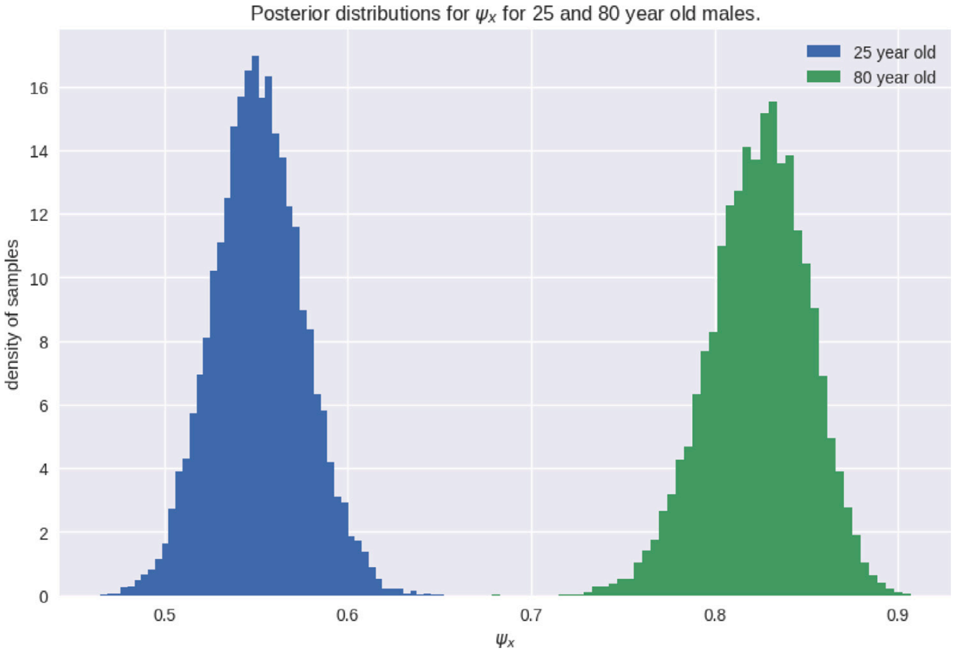

Fig. 5. Posterior distributions of ψ_x for 25 and 80 year old men. The histograms show that the probability of x treatment is clearly higher for 80 than for 25 year old men.

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

5.3. 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 trivial as the priors should not exclude plausible parameter values, but also not put (much) weight on implausible outcomes either (McElreath, 2020). Since we have data on the whole Dutch population there are enough observations to determine the posterior distributions (almost) independently of the prior choices, even in our train data. The exact choice of priors can be found in Appendix E. 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. Further, the main results are robust to different prior choices.

6. Estimation

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

6.1. Estimation methodology

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

Table 4
Summary of posterior distributions.

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

Notes: The figures in this table were produced by estimating the model described in Section 6 on the train data and by computing the means and standard deviations of 10,000 samples drawn from the estimated posteriors of the variables for each age, year and gender (see also Section 6.3).

asymptotic result. Blei et al. (2017) argue that although variational approaches tend to underestimate the variance of the posterior densities, for mixture models the results of variational inference may be better than MCMC.

Figures G.2 and G.3 in Appendix G show the ELBO (evidence lower bound) plot for estimating the model for women and men. The plots are skewed and flat on the right hand side, which implies convergence of the ADVI algorithm.

6.2. Parameter estimates

Here we summarize the posterior distributions of the directly relevant parameters for women and men: μ_x, μ_y and σ_x, σ_y, ψ_x and ψ_y, α, ν, and ζ. We do not present plots of each posterior distribution, as that would make the paper rather long. Our code repository contains the posterior samples and hence the interested reader can plot the distribution for each parameter.

Table 4 gives a first summary with mean values and standard deviations of the posterior distributions. Note that this summary provides a broad overview of the posteriors, as we have aggregated the posteriors across samples, ages, and years. We find that μ_x is on average approximately 5.0 for both men and women, while μ_y is lower, below 3.0. The standard deviations of the posterior distributions for μ_x, μ_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. Hence, the data do not suggest to deviate from symmetry here.

The value of ν is small. 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. A small ν is in line with our expectations as we do not expect a ten euro increase in $EOOP$ to have a big effect. The average value for ζ can be interpreted as women and men accepting 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, Fig. 6 gives the distribution of ν for women. 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 for a given value of $EOOP$.

Put differently, 70 year old women tend to be offered more valuable y treatments which they accept with higher probability for each level of $EOOP$ compared to 20 year olds. The effect that 70 and 20 year old women face different cost distributions is captured by $EOOP$ itself and does not affect the parameters ν and ζ of the utility distribution F. Furthermore, to give an idea of the uncertainty surrounding F for both age categories we also plot the 75th percentile of F. To illustrate the interpretation of this 75th percentile, consider an $EOOP = 500$. On average, 20 year old women will reject a treatment with $EOOP = 500$ with a probability of approximately 0.8. We are 75% sure that this rejection probability is less than 0.85, as is illustrated with the dashed blue line at this $EOOP$.

As a final illustration of our estimates, and a first step towards evaluating fit, consider Fig. 7. This figure plots the probability of positive expenditures, $1-(1-\psi_x)(1-\psi_y+\psi_y F)$ (see Table 1), for men across age in 2008. Recall that the parameters ψ_x and ψ_y are modeled as GPs. That is, we do not draw values from a distribution for each age, but chains of values across all ages from the GP. The figure illustrates this by showing these draws for ψ_y and ψ_x as thin red lines. Darker colors red indicate more draws at these values. On top of these draws, the realized fraction of positive expenditures are plotted for each category (age–male combinations) in 2008. Our predicted probability of positive expenditures is fairly close to the realized fractions of positive expenditures for each age category. But we tend to overestimate this probability for men around 80 years old. The equivalent graph for women is presented in Appendix G.

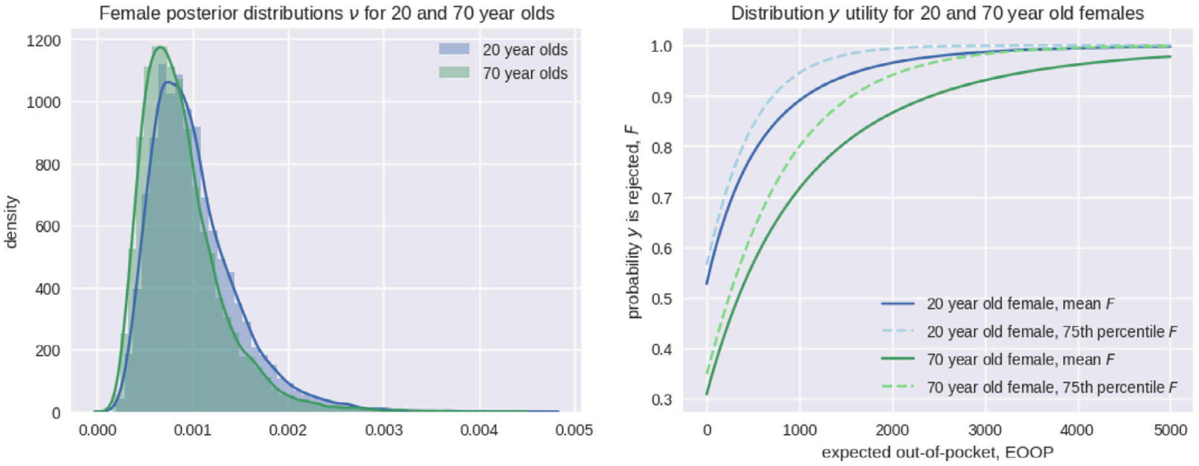

Fig. 6. Posterior distributions of ν and F for women. The left panel depicts two histograms with fitted lines of the (estimated) posterior distributions of ν for 20 and 70 year old women. The right panel shows F, the probability that y is rejected given the expected out-of-pocket payment $EOOP$ as described in Eq. (7) in Section 3.4. The mean and 75th percentile of F are computed using the estimated distributions of posteriors of ν and ζ.

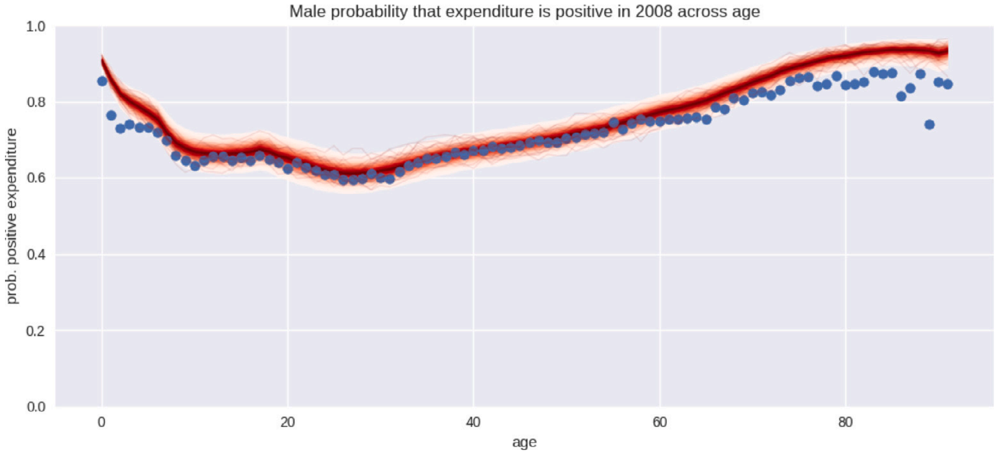

Fig. 7. Predicted and realized probabilities of positive expenditures for men across age in 2008. The blue dots are the average probability that expenditures are positive at each age for men in 2008. They are based on raw data in the train dataset. The red lines are derived from estimating the model described in Section 6, drawing samples from the estimated posterior variables (see also Section 6.3) and computing the probability of positive expenditures for men in 2008: $1-(1-\psi_x)(1-\psi_y+\psi_y F)$. Because ψ_x and ψ_y are modeled as GPs, we draw chains of values across all ages from the GP. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6.3. Model fit

To examine the fit of the model, we first predict healthcare expenditures. For each variable in the model we draw 10,000 samples from the underlying posterior distribution. 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 the samples of $\psi_x, \mu_x, \mu_y, \sigma_x, \sigma_y$ and D in a given year, we generate a value of $EOOP$ (Eq. (4)), which then combines with the ζ's and ν's to get values for F (Eq. (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.

Fig. 8 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. The vertical lines present the 95% interval for mean log expenditures per age category. This interval is generated by bootstrapping from the 10,000 samples that we have for (individual) z draws (per gender–age–year category) and calculating the mean. The fit of the model is quite good. Even the second moments, as shown in Figure G.5

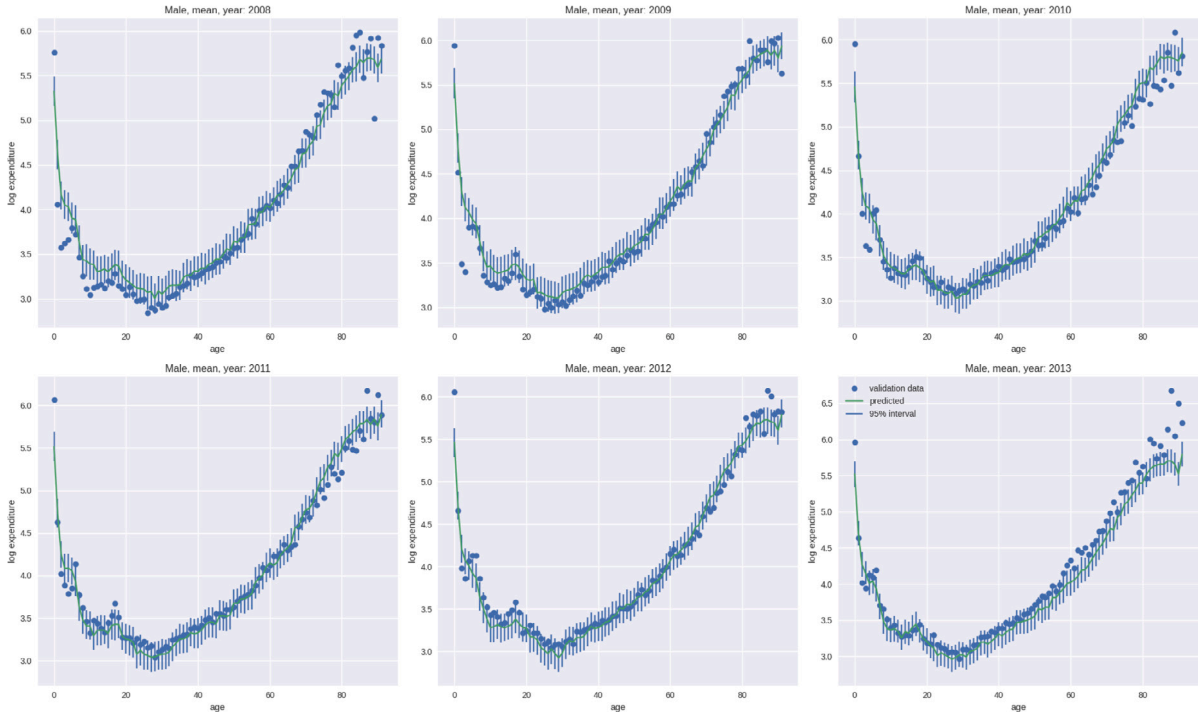

Fig. 8. Average predicted vs. average validation male log healthcare expenditures for 2008 to 2013. The six panels show for each year available, 2008–2013, how well the model predicts expenditures (conditional on being positive) for men at each age. To ensure a 'clean' comparison, the model predictions are compared to mean values and a 95% interval based on the validation dataset (on which the model was not estimated). Section 6.3 further describes how the predicted values and 95% interval were procured.

in the appendix, show a reasonable match. We over-estimate the standard deviation a bit till age 60 and then under-estimate the standard deviation around age 80.

7. Simulations and policy analyses

The goal of this paper is to develop a model that can be used to simulate different demand-side cost-sharing schemes. Now that we have estimated the model, we illustrate how it can be used for policy simulations by applying it to the Dutch healthcare system. Specifically, we compare three types of cost-sharing schemes: a deductible, a co-insurance scheme, and a shifted deductible. We simulate and compare seven maximum out-of-pocket payment levels: 0, 100, 200, 300, 400, 500, and 600 euros, four co-insurance rates: 0.25, 0.5, 0.75, and 1.0, in combination with seven starting points: 0, 100, 200, 300, 400, 500, and 600 euros. In total these add up to 196 different cost-sharing schemes.

The simulations suggest that shifting the starting point of the deductible is beneficial in that this reduces both expenditure per head and out-of-pocket expenditure.

Our simulations are conducted for the most recent year in our data, 2013, and maintaining the Dutch feature that people under 18 years old do not pay for healthcare under basic insurance. To simulate healthcare expenditures under different cost-sharing schemes we use the estimated posterior distributions of x and y for each gender–age category in 2013, and compute $EOOP$ given a cost-sharing scheme.

7.1. Results of the simulations

The main results of the simulation with the model are presented in Fig. 9. The left panel presents the effect of different levels of deductibles, co-insurance schemes and shifted deductibles. In this figure, we compare all results to a 300 euro deductible by normalizing expenditure on the expenditure with a 300 euro deductible. The vertical bars illustrate the uncertainty by showing the (bootstrapped) 95% interval for each simulated outcome. With the (obvious) exception of $D = 0$, the 95% intervals for the by 400 euros shifted deductible lie below the intervals for a standard deductible.

As expected, schemes with higher maximum out-of-pocket payments lead to lower healthcare expenditures. Having no demand-side cost-sharing, that is, the maximum out-of-pocket payment is zero, leads to approximately 12% higher costs compared to a 300 euro deductible. With a deductible equal to 600 euros, expenditures decrease by 10%. Increasing the deductible reduces expenditures per head, but at the cost of higher out-of-pocket spending for insured. The latter is illustrated in the right panel of the figure. With

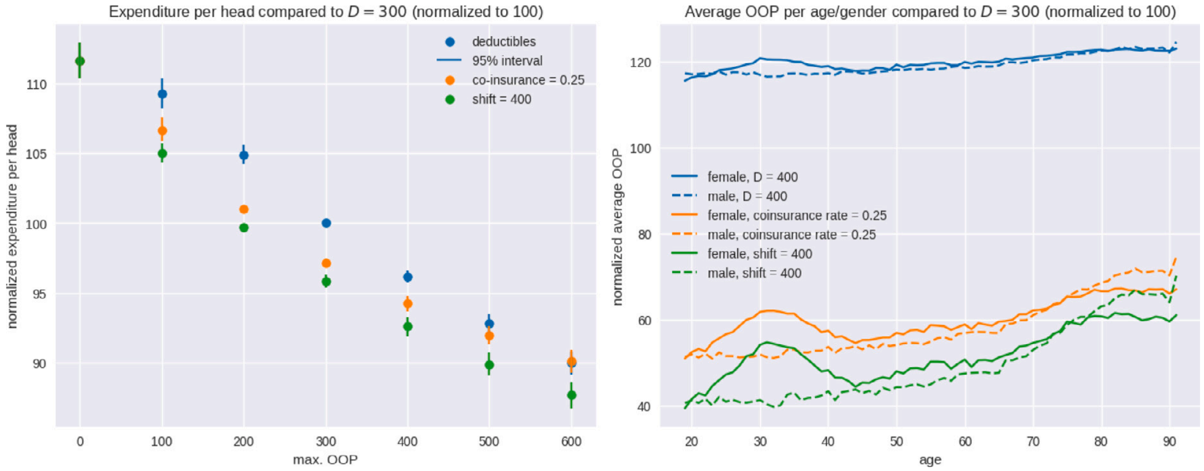

Fig. 9. Effects of demand-side cost-sharing schemes on expenditure per head and out-of-pocket payments. The left panel shows simulated healthcare expenditure per head for different cost-sharing schemes (deductible, co-insurance rate and shifted deductible) with a maximum out-of-pocket payment ranging from zero to 600 euros. Healthcare expenditure per head is compared to healthcare expenditure per head given a 300 euro deductible (normalized to 100). The vertical bars represent the 95% probability interval. The right panel shows the average simulated out-of-pocket payments given different cost-sharing schemes, across age and for men and women separately. Out-of-pocket payments are compared to a 300 euro deductible (normalized to 100). The figure is based on simulations for the year 2013 using the estimated posterior distributions of x and y for each gender–age category to compute $EOOP$.

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.

These results translate into a deductible elasticity of −0.11, which is in line with Remmerswaal et al. (2023).[4] Although we focus on the distribution of expenditures whereas Remmerswaal et al. (2023) model individual level expenditure (using individual fixed effects) as a function of the deductible level, both methods lead to a similar elasticity. Our approach here however allows us to simulate a variety of cost-sharing schemes.

Fig. 9 furthermore shows that both co-insurance schemes and shifted deductibles, each featuring a maximum out-of-pocket payment of 300 euros, effectively reduce healthcare expenditure and the average out-of-pocket payment per head compared to a standard 300 euro deductible. This occurs because both schemes raise the effective price of healthcare for individuals, given their expected healthcare expenditures. Consequently, there are more persons at the margin with these schemes. The reduction in healthcare expenditures and out-of-pocket payments alleviates the trade-off between total healthcare expenditure and out-of-pocket expenditure. Such an outcome is not observed when increasing the deductible from, for instance, 300 to 400 euros, as it would reduce healthcare expenditure while also increasing out-of-pocket expenditures. Below, we provide further insights into these findings.

7.2. Understanding the outcomes

Fig. 10 shows how healthcare expenditures vary for cost-sharing schemes with four different co-insurance rates and seven different starting points, or shifts. All schemes depicted in the figure have the same maximum out-of-pocket of 300 euros. Once again, we normalize healthcare expenditure on the average healthcare expenditure with a 300 euro deductible (and no shift), represented by the point (0,100) on the purple line in the figure.

Healthcare expenditures are minimized with a shifted deductible that starts at 400 euros and has a 100% co-insurance rate. Among the schemes that start at zero, i.e. those not shifted, a 25% co-insurance rate results in the most substantial reduction in healthcare expenditure. The 400 euro shift and the 25% co-insurance have the following beneficial effects compared to a traditional deductible (starting at 0). These schemes (i) reduce moral hazard, (ii) with the same maximum out-of-pocket (of 300 euro) and (iii) lower out-of-pocket payments in each state of nature compared to the standard deductible. Although we assume in estimating the baseline model that the chronically ill (PCG, DCG) do not react to changes in demand-side cost-sharing over the relevant Dutch policy range, these implications also apply to them. Due to the reduction in moral hazard the premium falls and in each state of the world they pay (weakly) less out-of-pocket due to a shift or co-insurance.

Fig. 10 allows us to illustrate a number of mechanisms in the model. First, a significant part of the sample is not at the margin under a relatively low deductible of 300 euros. Therefore, without a shift, the optimal strategy in terms of minimizing expenditure

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

$$\varepsilon = \frac{\Delta y}{\Delta D}\frac{\bar{D}}{\bar{y}} = \frac{\frac{y_{600}}{y_{300}} - \frac{y_0}{y_{300}}}{\frac{600-0}{\frac{600+0}{2}}} = \frac{0.90 - 1.12}{2} = -0.11.$$

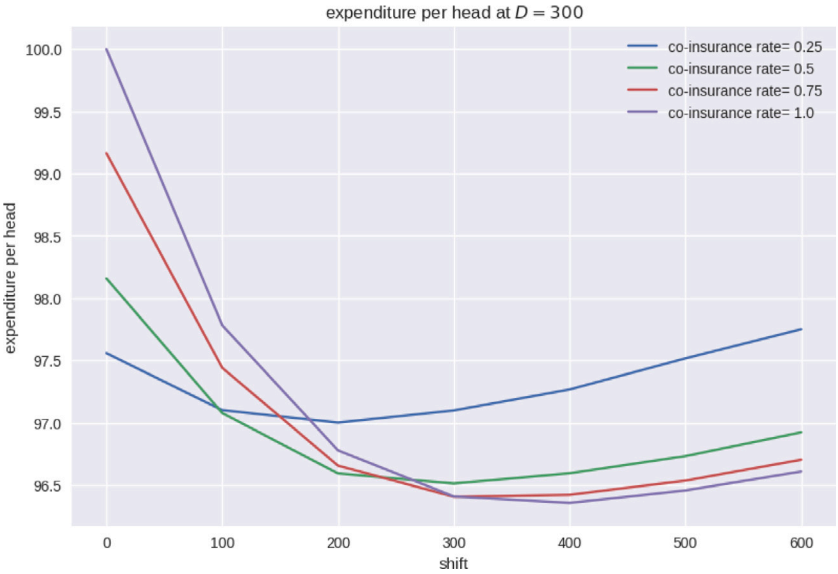

Fig. 10. The effect of varying the shift and co-insurance rate. The graph shows simulated healthcare expenditure per head for cost-sharing schemes with varying levels of the shift and co-insurance rate, each with a maximum out-of-pocket payment of 300 euros. Healthcare expenditure per head is compared to healthcare expenditure per head given a 300 euro deductible (normalized to 100). The figure is based on simulations for the year 2013 using the estimated posterior distributions of x and y for each gender–age category to compute $EOOP$.

per head involves a low co-insurance rate of 25%. This co-insurance rate blunts the effect of the deductible, as you only pay 25 cents out-of-pocket for every euro healthcare costs you make. However, it extends the range over which individuals face cost-sharing from $[0, 300]$ to $[0, 1200]$ euros. A 25% co-insurance rate with a maximum out-of-pocket payment of 300 euros reduces healthcare expenditure per head by 2.5% compared to a 300 euro deductible. Moreover, as shown in the right panel of Fig. 9, it also reduces the out-of-pocket payment on average by at least 30%.

When we introduce a shift in the cost-sharing scheme and move to the right in Fig. 10, 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 for shifts beyond 300 euros it is optimal to have a rate equal to 1.0. In other words, shifting the starting point puts more people at the margin. Reducing the co-insurance rate below 100% to extend the expenditure range subject to cost-sharing, is then no longer optimal.

Apparently, $D = 300$ together with a shift of 400 euros captures enough people at their margin that reducing the co-insurance rate is no longer useful. Not only does this minimize healthcare expenditure per head for $D = 300$, it also reduces the out-of-pocket payment for everyone as Fig. 9 shows.

At first sight, this outcome may seem surprising, as we have shown that many people in our sample have zero or relatively low expenditure. For this group healthcare becomes (almost) free under a deductible with a 400 euro shift. How can this be optimal? There are two underlying reasons for this. First, a considerable proportion of individuals with zero expenditures are those who are not offered any treatments, rendering the shift inconsequential for them. Additionally, part of the people with low expenditures has exogenous (x) treatments, rendering the shift equally ineffective for them. Secondly, as previously mentioned, our model is additive in log (x and y) expenditures. This implies that for individuals with high exogenous expenditure there are more possibilities for high endogenous expenditure (in euros). Essentially, individuals with existing health issues necessitate diagnostic and treatment procedures, thereby opening up possibilities for additional treatments, such as supplementary X-rays. Therefore, to minimize overall expenditure, it is necessary for individuals with some existing (exogenous) expenditure to be at their margin. This is precisely what a shifted deductible helps to establish.

The result that a 400 euro shifted deductible reduces expenditure per head and out-of-pocket payments can be different for other countries with different expenditure distributions, for example due to a different age profile of the population. The best known study on the effects of cost-sharing on healthcare expenditures is Newhouse and the Insurance Experiment Group (1993). They do not analyze a shifted deductible, but they do consider co-insurance schemes, including a rate of 25%. Whereas we find that co-insurance is more effective in reducing healthcare expenditure compared to a 100% rate, they show the opposite, although the differences between the effects of the different co-insurance rates are small, especially for small maximum out-of-pocket amounts. Several factors could explain this difference. For instance, the levels of healthcare expenditure and the institutional settings of the United States in the 1970s differ significantly from that of the Netherlands in 2013. However, the general principle that a shift of the deductible increases the optimal co-insurance rate will hold for all countries. In this sense, the deductible shift and the co-insurance rate can be viewed as complementary policy instruments.

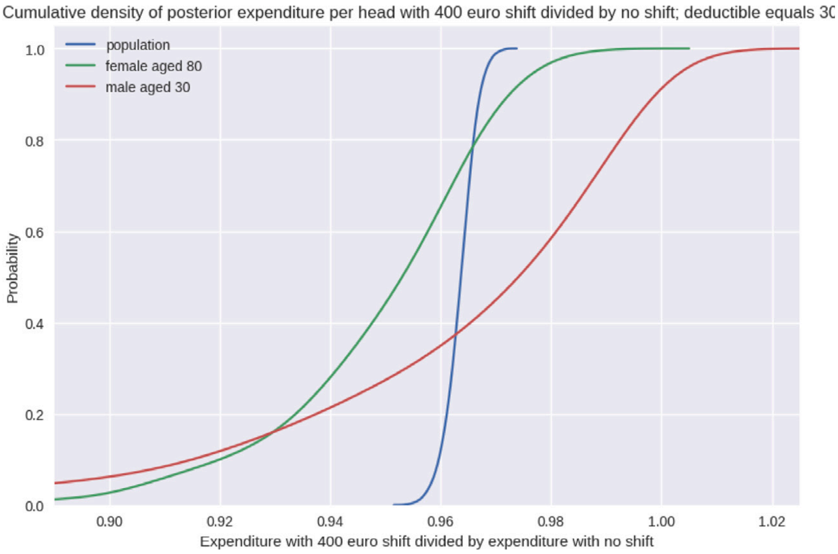

Fig. 11. Uncertainty of simulation results for a shifted deductible with a shift of 400 euros and a maximum out-of-pocket payment of 300 euros. The graph depicts the cumulative posterior distribution of expenditure per head for the full sample, 80 year old women and 30 year old men. Healthcare expenditure per head is compared to healthcare expenditure per head given a traditional (not shifted) 300 euro deductible (normalized to 100). The figure is based on simulations for the year 2013 using the estimated posterior distributions of x and y for each gender–age category to compute $EOOP$.

7.3. Policy uncertainty

As our model employs Bayesian methodology, we can show the uncertainty surrounding our policy recommendations. This is illustrated in the cumulative distribution plot in Fig. 11. The horizontal axis of the graph depicts average (across a population) healthcare expenditure with a 300 euro deductible with a 400 euro shift, divided by average expenditure with the same deductible, but without a shift.

For the population as a whole, it is likely that the deductible shift reduces expenditure per head by at least 3%. However, we do not expect the reduction to exceed 5%. As we aggregate over 2*92 categories (using the population weights), there is relatively little uncertainty surrounding this effect. If we focus on one particular gender–age category, there is more uncertainty on the effect of the 400 euro shift. But for 80 year old females we are still pretty sure that the shift will reduce expenditure per head. Given that expenditure per head is quite high for this category, it is not surprising that the shift brings more agents at the margin. However, this is not clear for 30 year old males. Indeed, the figure shows that there is a 10% probability that the 400 euro shift increases expenditure per head for this group by reducing the number of people at the margin.

7.4. Differentiated starting points

Until now, we have explored shifted deductibles with a uniform starting point for everyone. Because we have the distribution of expenditures for each gender–age category, we can characterize the optimal shift and co-insurance rate for each gender–age category. To illustrate this, Fig. 12 considers people above the age of 30. We expect that the optimal shift will increase with age. Doing this for each age separately, for example, for 40 year olds, 41 year olds, and so on, will lead to a noisy picture with an upward trend. However, this is not a serious policy option as it could mean, for example, that the shift is 300 euros when you are 40 years old, it becomes 350 euros for age 41, and then falls again to 320 euros at age 42. This would be too confusing for people to understand.

We consider $D = 300$ and age-brackets of 10 years: 30 to 40 year olds, 40 to 50 year olds, and so on; within an age bracket both genders face the same deductible shift. Fig. 12 plots average expenditure per head for each age class as a function of the shift, standardized by the average expenditure per age class with a (uniform) 400 euros shift. In this way, the curves are comparable as normally the expenditure for the age group 80–90 would be far higher than for 30–40. With this standardization, all curves have the point $(400, 1.0)$ in common.

The figure shows that the optimal shift indeed increases with the age bracket. For the younger ages standardized expenditure is minimized with shifts to the left of 400 and for the older ages the minimum is to the right of 400. Age-range 80–90 minimizes expenditure with a shift of 500 euros.

Note that the relative reduction in expenditure per head with differentiated starting points is modest. The main reduction is due to the 400 euro shift; differentiating the shift per age category yields small further gains.

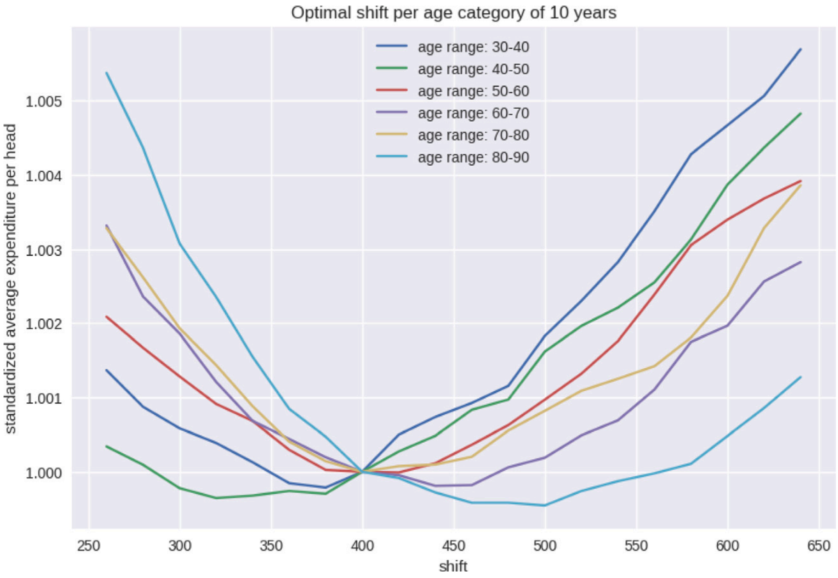

Fig. 12. Optimal shift given different age categories. The graph shows simulated healthcare expenditure per head for cost-sharing schemes with varying levels of the shift and each a maximum out-of-pocket payment of 300 euros. Healthcare expenditure per head is compared to healthcare expenditure per head given shifted deductible with a maximum out-of-pocket payment of 300 euros and a 400 euro shift (normalized to 100). The figure is based on simulations for the year 2013 using the estimated posterior distributions of x and y for each gender–age category to compute $EOOP$.

7.5. Quality of care

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

We simulate the effect of a change in the (perceived) quality of care u on healthcare spending by changing parameter ν in the function of F (Eq. (7)) in the model and assume that cost-sharing does not change. An advantage of our specification of F is that expected quality/utility is simply given by:

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

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 $E_1(u) = (1+g)E_0(u)$: expected utility has increased with growth rate $g > 0$. Fig. 13 shows how healthcare expenditures change, when we increase the quality of healthcare by 10%, 20%, or 30% (the darker the color, the larger the increase in quality). We increase the quality by lowering the estimated posterior distributions of ν, and thereby lowering the probability that a treatment is rejected, given $EOOP$. The figure shows that increasing the value u of healthcare (by 10%, 20%, and 30%) under a 300 euro deductible increases expenditure. A 10% increase in the utility of treatments does not directly translate into 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 utility increase (because a quality cut-off of zero is not affected by the factor $(1+g)$; this would be different if we increased quality via an increase in ζ), and as the maximum out-of-pocket payment increases, the differences become larger.

7.6. Robustness checks

In Appendix F we present three robustness checks. First, we consider our selection of the baseline sample where we dropped the chronically ill from the sample. We argue that our baseline sample is logical, because we want to estimate how expenditure varies with the relatively low deductible levels in our Dutch data. It is unlikely that these groups with expenditure (far) exceeding these deductible levels would react to deductible changes in this range. However, when we estimate the model for chronically ill persons separately and mix the outcomes with the outcomes of the baseline sample to see the effect on the overall results, our findings are indeed unaffected.

Second, we consider a normal distribution to capture the probability that a treatment is rejected, F in Eq. (7). Also here the results are robust to this adaptation of the model.

Finally, we allow for the case where agents are not fully rational in their perception of the cost distributions that they face. In particular, we allow them to put more weight on the first moments of the distributions compared to the second moments. Intuitively, they base their decisions more on expected expenditure than on the distribution as a whole. Also here our results are robust.

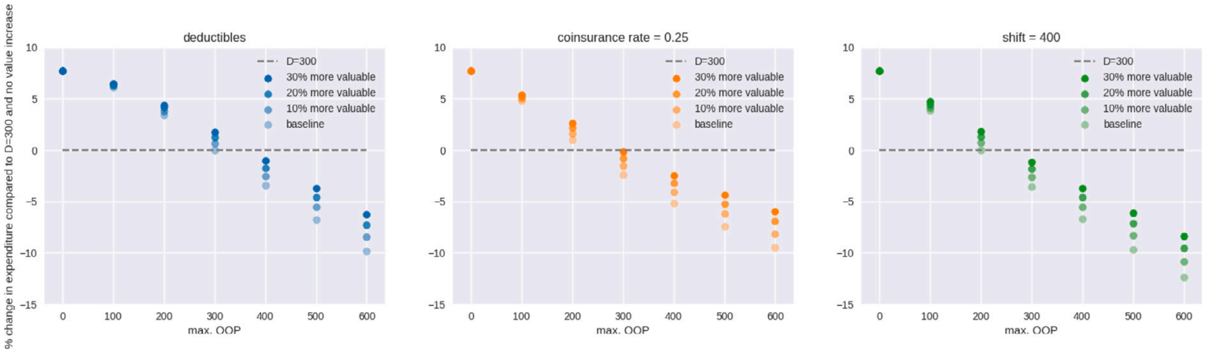

Fig. 13. Healthcare expenditures after increasing the value of treatments by 10, 20, and 30%. The left panel shows the percentage change in simulated healthcare expenditure per head when increasing the value u of healthcare by 10%, 20%, and 30% under a 300 euro deductible. The percentage change is depicted for deductibles ranging between zero and 600 euros. Healthcare expenditure per head is compared to healthcare expenditure per head given a 300 euro deductible and without an increase of u (the baseline normalized to 100). This is also depicted by a horizontal dashed line. The middle and right panel are identical to the left panel but for schemes with a co-insurance rate of 25% (middle panel) and a shifted deductible with a shift of 400 euros (right panel). The figure is based on simulations for the year 2013 using the estimated posterior distributions of x and y for each gender–age category to compute $EOOP$.

8. Discussion

In this paper, we create a structural microsimulation model to simulate the impact of different types and levels of cost-sharing on healthcare expenditure and individual out-of-pocket spending. To apply the model to the Dutch healthcare system, it is estimated on Dutch data and we present a number of policy simulations. The methodology can be applied to other countries.

To assess the effects of different cost-sharing schemes, we estimate the distributions of healthcare expenditures and of treatment utilities for different gender–age categories. With these distributions, we can determine for each scheme the probability that individuals in a gender–age category are at the margin for a particular scheme. The more individuals in the category are at the margin, the more effective a scheme is to curb healthcare expenditure.

The model illustrates how different cost-sharing arrangements affect the trade-off between the effect on total expenditure and out-of-pocket expenditure. To illustrate, for the Netherlands a shifted deductible is an effective way to reduce healthcare expenditures without increasing out-of-pocket expenditures.

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

The structural model has many possibilities for extensions and additional analyses, which are not included in the paper. To illustrate, we left the derivation of the optimal demand-side cost-sharing scheme for future research. This could for example be a two-tier system with 100% co-insurance rate over the expenditure range between 0 and 300 euros and then a 50% co-insurance rate between 300 and 600 euros. The categories which determine an individual's (expected) distribution of healthcare expenditure are in this paper simply based on gender and age. However, they can be made more homogeneous by, for example, using an individual's expenditure in the previous year. A category can then be: a 25 year old male with less than 1000 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 to determine an individual's decision whether to accept higher out-of-pocket expenditure in return for a lower premium. As we focused on mandatory insurance and mandatory deductible, we did not need to model risk aversion yet.

The model also has some limitations and a number of these originate from the data. To illustrate, the data comprise total healthcare expenditures per person per year, but not the exact underlying treatments, visits, scans and check-ups. As a result, we cannot simulate the effects of cost-sharing schemes such as co-payments, in which people pay for example 50 euros per visit to the hospital. Further, when simulating high levels of cost-sharing – say, around 2000 euros – 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. Related to this, because the deductible levels in our data are relatively low we cannot identify the parameters of the model for the chronically ill. Their expenditure tends to be (far) higher than the maximum deductible (350 euros) and hence observed changes in the deductible are not expected to affect their treatment choices.

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

CRediT authorship contribution statement

Jan Boone: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Minke Remmerswaal:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jhealeco.2024.102900.

References

Bajari, P., Dalton, C., Hong, H., Khwaja, A., 2014. Moral hazard, adverse selection, and health expenditures: A semiparametric analysis. Rand J. Econ. 45 (4), 747–763.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: A review for statisticians. J. Amer. Statist. Assoc. 112, 859–877.

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

Cardon, J.H., Hendel, I., 2001. Asymmetric information in health insurance: Evidence from the national medical expenditure survey. Rand J. Econ. 32 (1), 408–427.

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

Deb, Partha, Burgess, Jr., James F., 2003. A Quasi-experimental Comparison of Econometric Models for Health Care Expenditures. Economics Working Paper Archive at Hunter College 212, Hunter College Department of Economics, URL https://ideas.repec.org/p/htr/hcecon/212.html.

Deb, Partha, Holmes, Ann M., 2000. Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. Health Econom. 9 (6), 475–489.

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

Deb, Partha, Trivedi, Pravin, 1997. Demand for medical care by the elderly: a finite mixture approach. J. Appl. Econometrics 12 (3), 313–336.

Dutch Healthcare Authority, 2014. Marktscan en Beleidsbrief Zorgverzekeringsmarkt: Weergave van de Markt 2010–2014. Dutch Healthcare Authority.

Einav, L., Finkelstein, A., Ryan, S.P., Schrimpf, P., Cullen, M.R., 2013. Selection on moral hazard in health insurance. Amer. Econ. Rev. 103 (1), 178–219.

Einav, L., Finkelstein, A., Schrimpf, P., 2015. The response of drug expenditure to non-linear contract design: Evidence from medicare part D. Q. J. Econ. 130 (2), 841–899.

Ellis, R.P., 1986. Rational behavior in the presence of coverage ceilings and deductibles. Rand J. Econ. 17 (2), 158–175.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2013. Bayesian Data Analysis. Chapman and Hall/CRC.

Geweke, J., Gowrisankaran, G., Town, R.J., 2003. Bayesian inference for hospital quality in a selection model. Econometrica 71 (4), 1215–1238.

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

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

Jones, Andrew M., 2012. Models for health care. In: The Oxford Handbook of Economic Forecasting. Oxford University Press, pp. 625–654, (Chapter 23).

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

Klein, T.J., Salm, M., Upadhyay, S., 2022. The response to dynamic incentives in insurance contracts with a deductible: Evidence from a difference-in-regression-discontinuities design. J. Public Econom. 210.

Klein, Tobias, Salm, Martin, Upadhyay, Suraj, 2023. Patient Cost-Sharing and Risk Solidarity in Health Insurance. Technical Report, Tilburg University.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., Blei, D.M., 2017. Automatic differentiation variational inference. J. Mach. Learn. Res. 18, 1–45.

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

Mukherji, A., Roychoudhury, S., Ghosh, P., Brown, S., 2016. Estimating health demand for an aging population: A flexible and robust Bayesian joint model. J. Appl. Econometrics 31, 1140–1158.

Newhouse, J.P., the Insurance Experiment Group, 1993. Free for All? Lessons From the RAND Health Insurance Experiment. Harvard University Press, Cambridge, Massachusetts.

Rasmussen, C.E., Williams, C.K.I., 2005. Gaussian Processes for Machine Learning. MIT Press Cambridge.

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

Remmerswaal, M.C., Boone, J., Douven, R.C.M.H., 2023. Minimum generosity levels in a competitive health insurance market. J. Health Econom. 90, 102782.

Salvatier, J., Wiecki, T.V., Fonnesbeck, C., 2016. Probabilistic programming in python using Pymc3. PeerJ Comput. Sci. 2.

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

Van Kleef, R.C., van de Ven, W.P.M.M., van Vliet, R.C.J.A., 2009. Shifted deductibles for high risks: More effective in reducing moral hazard than traditional deductibles. J. Health Econom. 28 (1), 198–209.

Zweifel, P., Manning, W.G., 2000. In: Culyer, A.J., Newhouse, J.P (Eds.), Handbook of Health Economics, vol. 1, Elsevier, Amsterdam, (Chapter Moral Hazard and Consumer Incentives in Health Care).