

Online Appendix to “The Socio-Economic Distribution of Choice  
Quality: Evidence from Health Insurance in the Netherlands.”  
by Benjamin Handel, Jonathan Kolstad, Thomas Minten and Johannes Spinnewijn

## A Online Appendix

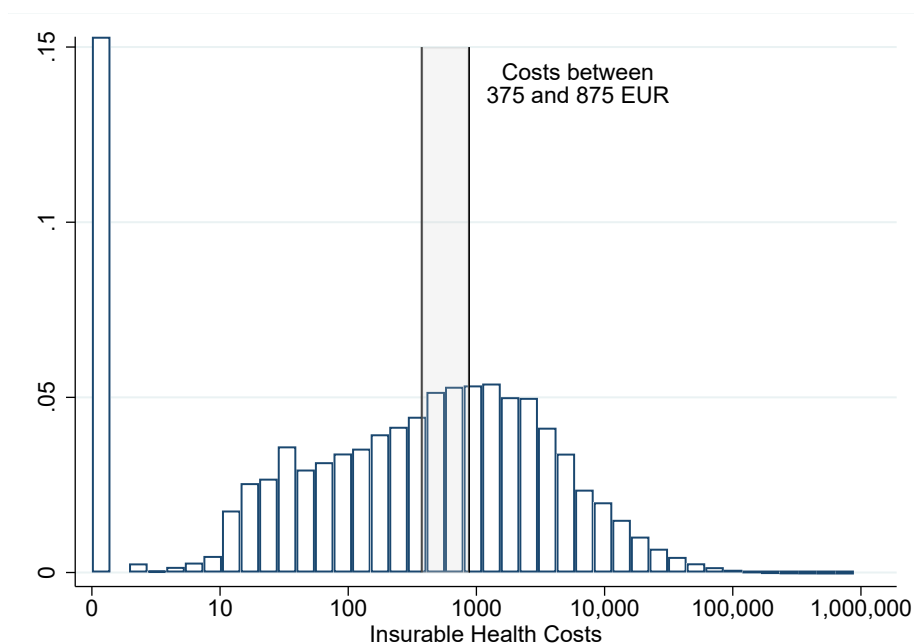
### A.1 Summary Statistics

TABLE A.1: SUMMARY STATISTICS

	Mean		Mean
<b>Demographics</b>		<b>Household Financial Status</b>	
Male	48.8%	Gross Household Income	73,289
Age	50.3	<i>10th Percentile</i>	<i>20,077</i>
Has Children	69.2%	<i>Median</i>	<i>60,358</i>
Has a Partner	62.9%	<i>90th Percentile</i>	<i>135,981</i>
<b>Education Level</b>		Household Net Worth	166,890
Less than High School	13.2%	<i>10th Percentile</i>	<i>-28,918</i>
High School	24.1%	<i>Median</i>	<i>32,694</i>
College	16.8%	<i>90th Percentile</i>	<i>403,923</i>
Further Studies	0.6%	Mortgage Debt	54.1%
Unknown	45.4%	Other Debt	34.2%
<b>Employment Status</b>		Savings > 2000 EUR	80.4%
Employee	44.3%		
Self-Employed	9.9%		
Retired	24.2%		
Student	6.3%		
Other Not Working	15.3%		
Observations			11,991,628

**Notes:** This table shows summary statistics for our full data sample in 2015, combining the prediction sample used for our cost model and the hold out sample used for our main analysis.

FIGURE A.1: DISTRIBUTION OF INSURABLE HEALTH CARE COSTS



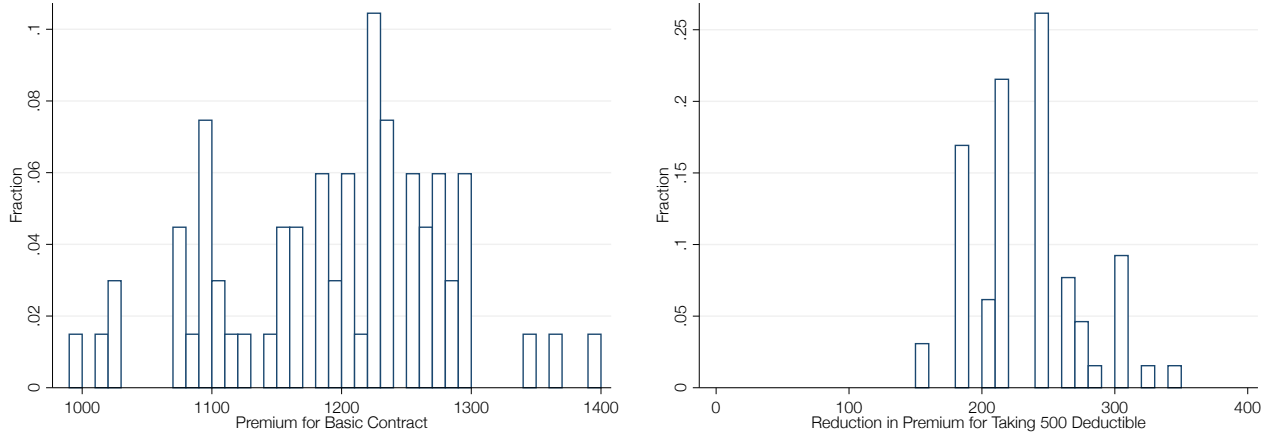
**Notes:** This figure shows the distribution of the  $\log_{10}$  of total yearly insurable health care costs in 2015, for all individuals in our baseline sample. 13.1% of individuals have health costs falling in the 375 to 875 EUR interval.

TABLE A.2: DISTRIBUTION OF ANNUAL HEALTH CARE COSTS

	Mean	p10	p50	p90	p99
All Care	2,695	86	495	6,032	35,974
Insurable Care	2,272	0	332	5,043	31,133
Hospital Care	1,388	0	85	2,829	21,575
Medicines	320	0	53	758	3,253
Mental Care	243	0	0	0	4,801
Tools and Medical Aid	107	0	0	145	2,284
Geriatric Care	53	0	0	0	0
Transport	45	0	0	0	1,081
Multidisciplinary Care	33	0	0	124	397
Physiotherapeutic Care	32	0	0	0	1,095
Dental Care	26	0	0	0	825
Other Care	7	0	0	0	151
Sensory Handicap Care	3	0	0	0	0
Always Insured Care	423	75	121	327	8,042
Nursing Care	228	0	0	0	7,587
GP Care	157	75	119	272	659
Maternal Care	37	0	0	0	1,796
Observations					11,991,629

**Notes:** This table shows the distribution of health expenditures by subcategory, for the full sample in 2015. Expenditures are divided into insurable expenditures, that are subject to cost sharing (and to which the deductible applies) versus always insured expenditures, that are not subject to cost sharing. All values are in EUR.

FIGURE A.2: DISTRIBUTION OF PREMIA AND PREMIUM REDUCTIONS



**Notes:** This figure presents histograms of yearly premiums in 2015 for basic coverage (left-hand side) and premium reductions for those contracts when electing a maximal extra deductible of 500 for a total deductible of 875 EUR (right-hand side). Data on prices are obtained from homefinance.nl.

TABLE A.3: DISTRIBUTION OF DEDUCTIBLE CHOICES

Default Deductible	90.94%
Extra Deductible (+100 to +500EUR)	9.06%
Breakdown of Extra Deductible Choices	
+100EUR	10.64%
+200EUR	10.41%
+300EUR	6.02%
+400EUR	1.72%
+500EUR	71.21%

**Notes:** This table shows the breakdown of deductible choices in 2015. A large majority (90.94%) sticks to the default 375 EUR deductible. Of the 9.06% individuals that take an extra deductible, most individuals take the 500 EUR extra deductible.

## A.2 Data Appendix

This Data Appendix provides information on the additional datasets we linked to our health cost and insurance data at Statistics Netherlands. Datasets are linked at the individual level based on anonymized individual identifiers. Please contact the authors for additional information on accessing these data.

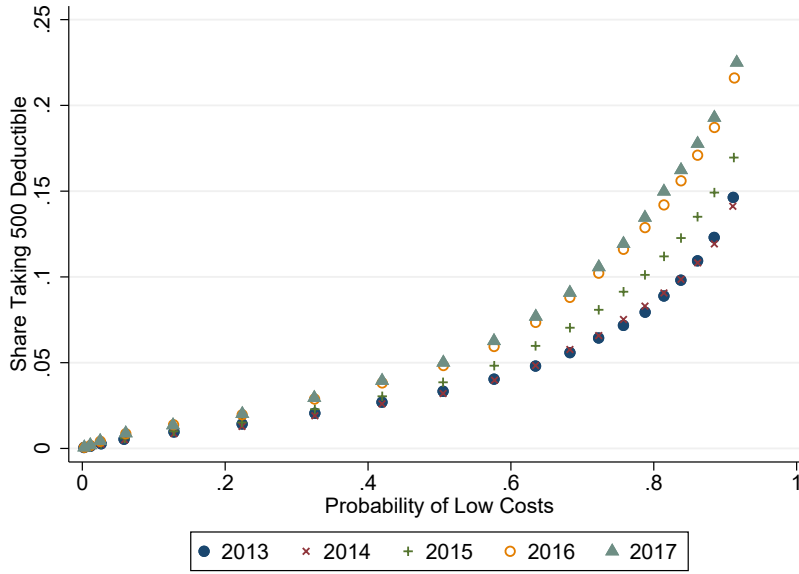
**Age and gender** Dataset *Gbapersoontab* provides an overview of all people registered living in the Netherlands at any point since 1995. These registers form a basis for the administrative records of all individuals in the Netherlands. For our purposes, *Gbapersoontab* is used to obtain age and gender, and we use this person registry as the primary dataset to match all other datasets with.

**Family and household links** Family links come from the dataset *Kindoudertab*, which contains all known legal child-parent links. Household identifiers as well as family status variables in *Ipi* and *Inpatab* allow us to identify partners and other household links. Partnerships consist of all partners who are living in the same household.<sup>9</sup>

**Education** *Hoogsteopltab* is a dataset that includes the highest attained educational course for each individual,

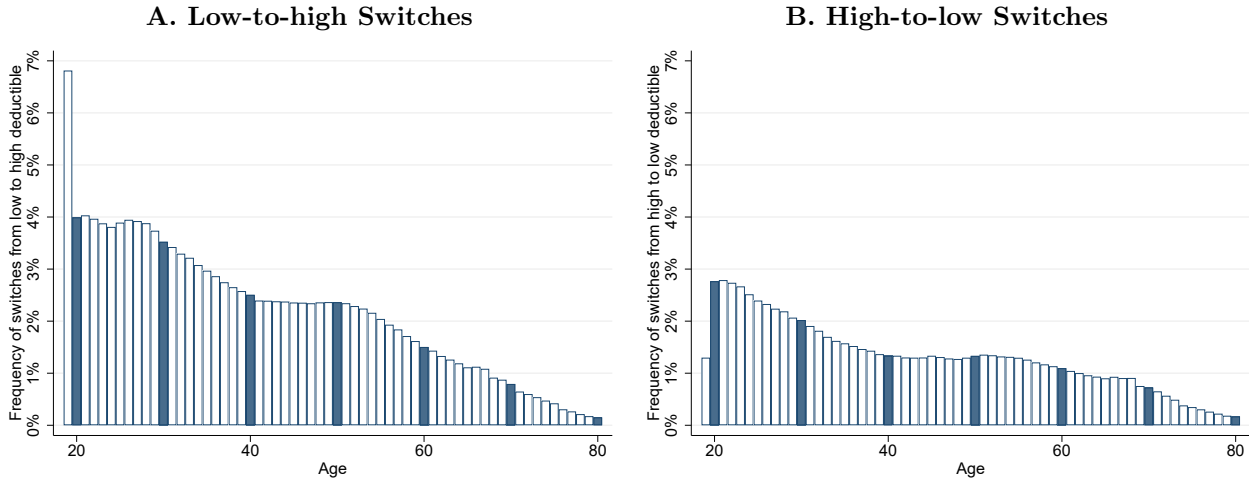
<sup>9</sup>This includes married partners, registered partners, but also partners who have not registered their partnership but are living in the same household.

FIGURE A.3: DEDUCTIBLE CHOICE GRADIENT BY YEAR



**Notes:** This figure displays the relationship between take-up of the 500 deductible and the predicted probability of low costs, separately for the five years included in our final sample.

FIGURE A.4: FREQUENCY OF DEDUCTIBLE SWITCHES BY AGE



**Notes:** This figure displays the frequency of deductible switches by age, in years 2014 to 2017. Panel A displays only switches to a higher deductible, and Panel B to a lower deductible.

and originates from several educational registers and survey data. We link each educational course to its relevant International Standard Classification of Education (ISCED) level and field of education. There is almost universal coverage for the youngest cohorts, but educational information is missing for many individuals aged over 40. Overall, we observe highest education obtained for 54.6% of our full sample.

**Income and Employment Status** Datasets *Ihi* and *Inhatab* contain information on households' income, and originates from tax authorities. Our main definition of income used in the analysis is household gross income

(called *bruto inkomen* by Statistics Netherlands). Gross income includes all labor income and capital income, as well as government transfers (e.g., UI, DI, pensions), and other transfers and income. We also use a socio-economic classification variable *seccoal1*, which classifies each individual based on where the majority of his or her personal income comes from. This variable is obtained from datasets *Ipi* and *Inpatab*.

**Wealth Dataset** *Vehtab* contains information from tax authorities on households’ assets and debts. This information is partly self-reported (on tax forms) and third-party reported. Assets include financial assets (savings, stocks, bonds, and other participations), real estate and other assets (such as cash and movable assets). Debts include mortgages, study debt and other debt. The net wealth variable in the main text equals household assets minus household debts.

**Employee-Employer links** We use the dataset *Spolisbus* to link individuals to their firms, colleagues and sector. *Spolisbus* is a highly detailed dataset with monthly information on all employment contracts in the Netherlands, collected by the tax authorities based on third-party reported data. We adopt the same definition of a firm as in the firm registry (*Algemeen Bedrijfsregister*) of Statistics Netherlands. We sum each individual’s total hours worked by year by firm. For each individual, we then select the firm at which that individual has worked the most hours in each year. The colleagues that we identify are thus all individuals who work the majority of their hours at the same firm. The sector categorization that we adopt is made by the authorities based on the collective labor agreements.

**Location** We match every individual with their yearly 6-digit postcode based on their registered residence. For this, we use datasets *Gbaadresobjectbus* and *Vslgwbtat*. Postcodes are obtained for each year on 1 October, as this is close to the period of deciding on their health insurance contract. 6-digit postcode information is at a neighbourhood level, and there are 12’116 distinct postcodes in 2015.

### A.3 Health Cost Predictions

In this Appendix, we describe the binary prediction algorithm that we use to obtain risk probabilities, and discuss its accuracy across different subgroups, and the most important predictors. We also discuss an alternative non-binary prediction algorithm and argue why the binary predictions are preferable for the analysis in this paper.

### A.4 Prediction Algorithm

We use an ensemble machine learning algorithm to predict the probability that an individual’s health costs will not exceed the mandatory deductible of 375 EUR in any given year. The prediction algorithm we use is a standard machine learning method for binary classification, an ensemble learner that consists in our case of a random forest model, gradient boosted regression trees and LASSO model. To avoid overfitting, we train and calibrate the prediction algorithm on a training sample of 1.25 million individuals. We then use this trained prediction algorithm to obtain predictions for a hold-out sample of about 12 million individuals. All the analyses and statistics in the paper are developed use only this hold-out sample.

The prediction method we use follows four steps, which closely resemble the steps used in [Einav et al. \(2018\)](#). First, we follow standard practice in machine learning by tuning key parameters that govern the prediction models by 3-fold cross-validation. Second, we train the three resulting prediction models separately. Third, we combine the three obtained predictions into one using a linear combination that we calibrate in the data. Finally, we calibrate the resulting final ensemble predictions using a linear spline. As there is some variation in the number and definition of predictors that we have across time, we repeat these four steps for all years of study (2013-2017). We describe each of the four steps in more detail here.

**Parameter Tuning** As the three machine learning models that we use have parameters that are choosable by the researcher, we follow standard practice and tune these parameters using 3-fold cross validation. More specifically, we tune the following parameters using 100,000 observations: minimal node size (mid.node.size), number of variables used at each node (mtry) for the random forest model, learning rate (eta) for the boosted regression trees, and the shrinkage parameter (lambda) for the LASSO.<sup>10</sup> For each of these parameters, we optimize among 5 alternatives. We tune these parameters using 3-fold cross validation, where we are optimizing the area under the receiver operating characteristic curve (AUC).<sup>11</sup> Thus, for each of the parameter values we want to test, the model is trained on 2 folds (subsets of the training sample), and then the performance is measured in the 3rd fold. The parameter values for which the AUC in the hold-out sample is highest for each prediction algorithm are: mtry = 10, min.node.size = 10, eta = 0.2, lambda = 0.0001.

**Estimating the Models** Using these tuned parameter values, all models are estimated using a training sample of 800,000 individuals.

**Obtaining Ensemble Predictor** We combine the predictions from the random forest, gradient boosting regression trees, and LASSO into one ensemble prediction. Following Einav et al. (2018), we construct the ensemble prediction to be the linear combination  $p_{ensemble} = \hat{\beta}_{rf}\hat{p}_{rf} + \hat{\beta}_{gb}\hat{p}_{gb} + \hat{\beta}_{lasso}\hat{p}_{lasso}$ , where  $\hat{p}_x$  is the prediction from algorithm  $x$  and  $\hat{\beta}_x$  is the associated weight.

We obtain estimates for the weights from a constrained linear regression (with no constant and the weights summing to one) of the dummy for having costs below 375 EUR on the three individual predicted probabilities. For this step, we use 100,000 observations that we did not use in either step of parameter tuning nor the estimation of the models. We find associated weights in 2015 that are  $\hat{\beta}_{rf} = 0.67$ ,  $\hat{\beta}_{gb} = 0.08$  and  $\hat{\beta}_{lasso} = 0.25$ .

**Calibrating Probabilities** Finally, the raw probability predictions we get from the ensemble step are calibrated to the actual observed probabilities by estimating a linear spline. This calibration is done using 350,000 observations that are used in none of the previous steps. 10 equal sized bins are created based on the ranked predicted probability. In every bin the mean probability is calibrated to the observed mean probability for these observations. The piece-wise linear spline that follows from linearly interpolating all intermediary points serves as the last step in the prediction mechanism.

**Prediction Fit** Figure A.5 below presents key results illustrating the strong fit of our cost prediction model, for the year of 2015. Figure Panel A presents a binned scatter plot of our predicted probability of having low costs against the realized share of individuals with low costs. The predictions track the realized shares almost exactly. Panel B plots the ROC curve of the different prediction methods used, showing a strong prediction of true positives relative to false positives. The bottom figures present ex-post cost realizations of individuals with predicted low (Panel C) and predicted high (Panel D) costs. These figures illustrate that our model fit is strong even at the tails and carefully distinguishes predictably healthy from predictably sick individuals.

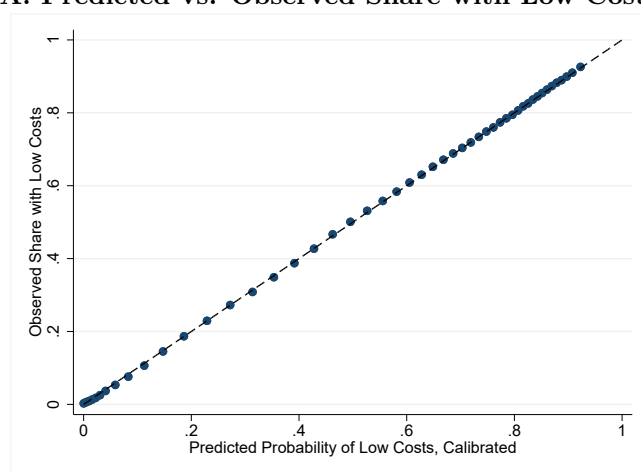
Figure A.6 presents the importance of different predictors in the random forest model, which is the model with the highest weight in our ensemble prediction. Not surprisingly, the most important predictors are different categories of past pharmaceutical spending, with  $t - 1$  values being more important than  $t - 2$  values. Hospital costs, costs to primary care visits and age are other important variables in the random forest prediction.

<sup>10</sup>We use the package CARET in R that provides a standardized way to tune parameters. The prediction models we use are RANGER (random forest), XGBLINEAR (boosted regression trees), and GLMNET (LASSO).

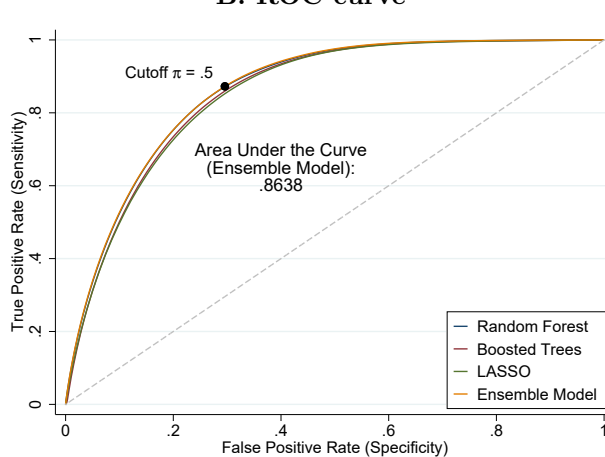
<sup>11</sup>This is a common metric used in the machine learning literature to measure the performance of a prediction model.

FIGURE A.5: PREDICTED VS. REALIZED COSTS

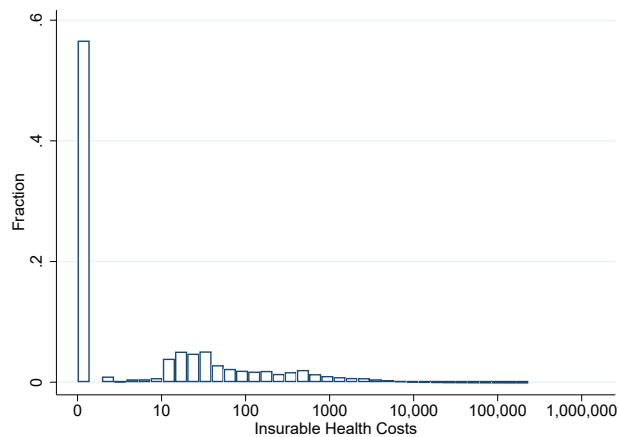
A. Predicted vs. Observed Share with Low Costs



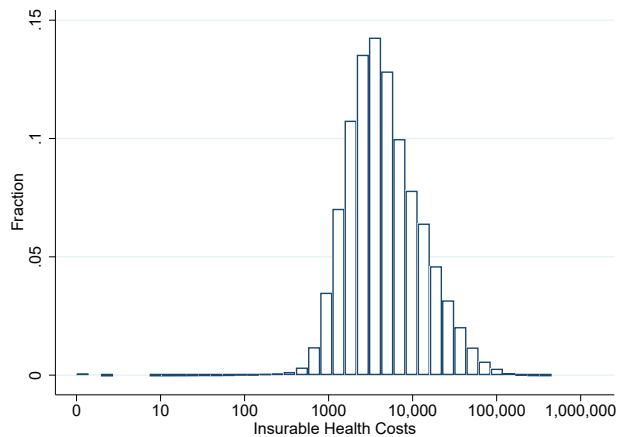
B. ROC curve



C. Top 5% Probability of Low Costs

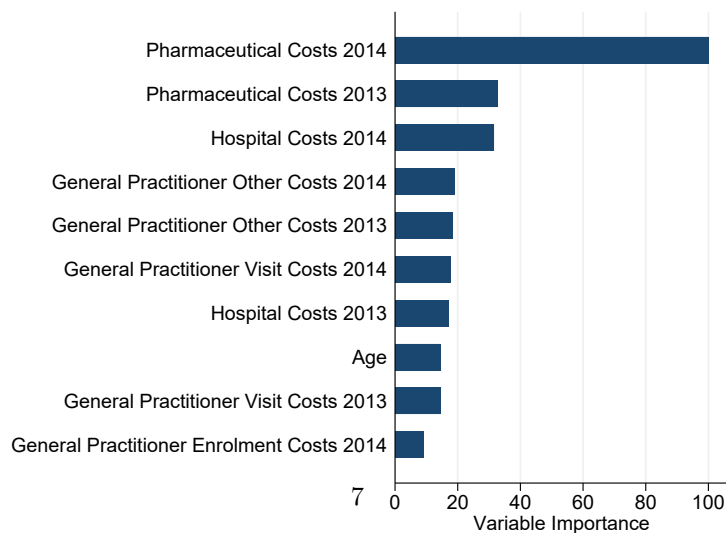


D. Bottom 5% Probability of Low Costs



**Notes:** Panel A presents a binned scatter plot of our predicted probability of having low costs against the realized share of individuals with low costs. Panel B plots the ROC curve of the different prediction methods used. The bottom figures present ex-post cost realizations of individuals with predicted low (Panel C) and predicted high (Panel D) costs. The year is 2015 for all Figures.

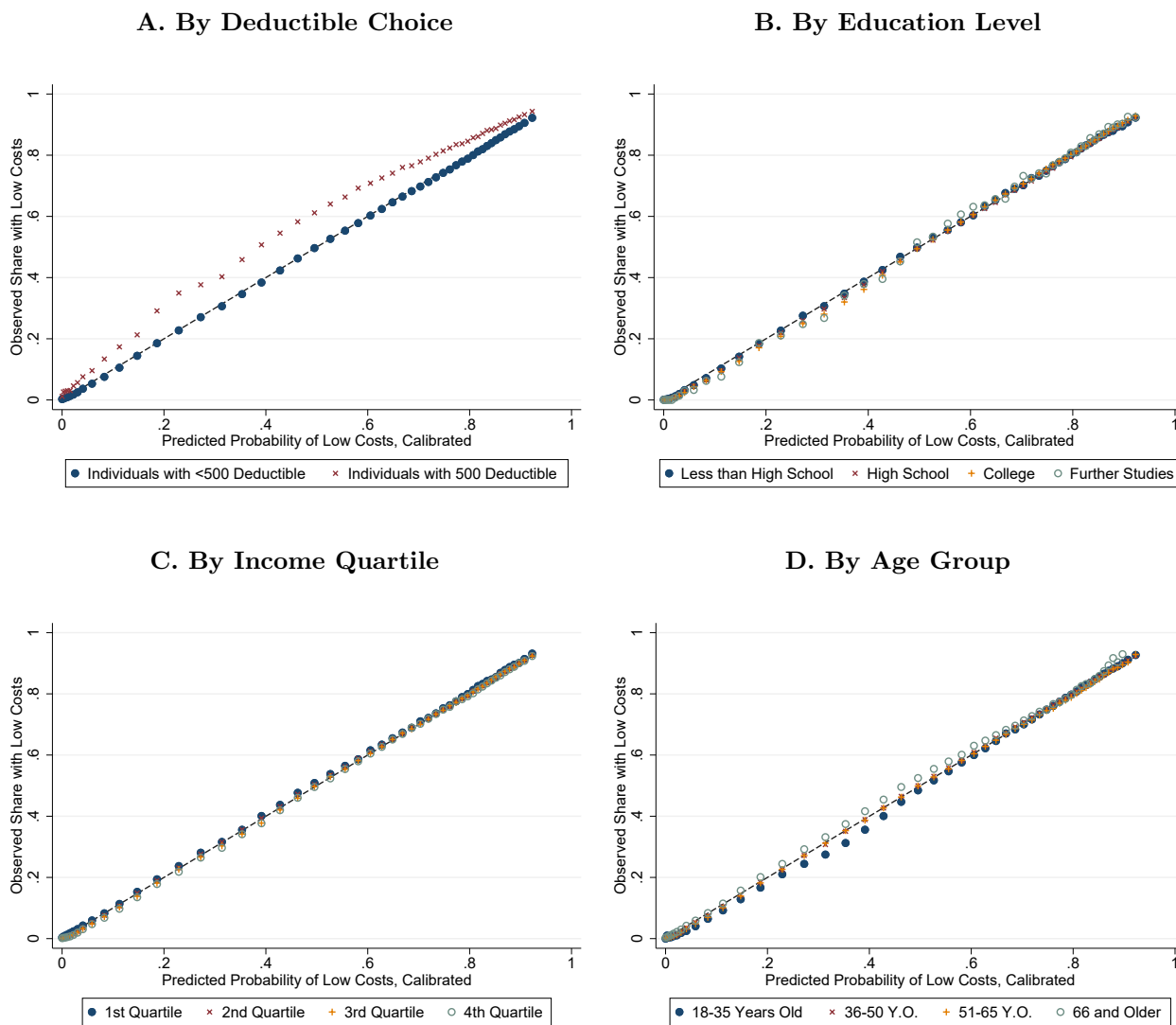
FIGURE A.6: VARIABLE IMPORTANCE IN PREDICTION WITH RANDOM FOREST



**Notes:** This figure shows the importance of selected variables in the prediction of health cost risk using only a random forest model.

**Subgroup Fit** While Figure A.5 shows a calibration plot for the entire sample, Figure A.7 shows a calibration plot for certain subgroups of the sample. As noted in the text, this is valuable to ensure that the heterogeneous choice quality we estimate is not in part related to differences in the cost prediction model by subgroup. We see from Panel B, C and D that probabilities are very precisely calibrated for distinct groups of education level, income quartile and age. This makes us comfortable that the observed differences in choice quality across these different groups are not due to differential prediction accuracy of our ensemble predictor.

FIGURE A.7: PREDICTED VS. OBSERVED SHARE OF LOW COSTS, BY SUBGROUPS



**Notes:** This figure shows the calibration plot of the predicted probability of low costs for various subgroups of the sample. Panel A plots our prediction against the observed share of people with health costs below 375 EUR, separately for people having chosen the 500 deductible and people who have not. Panel B does the same exercise splitting the sample by education level. In Panel C, the sample is split by income quartile, and in Panel D, by age group.

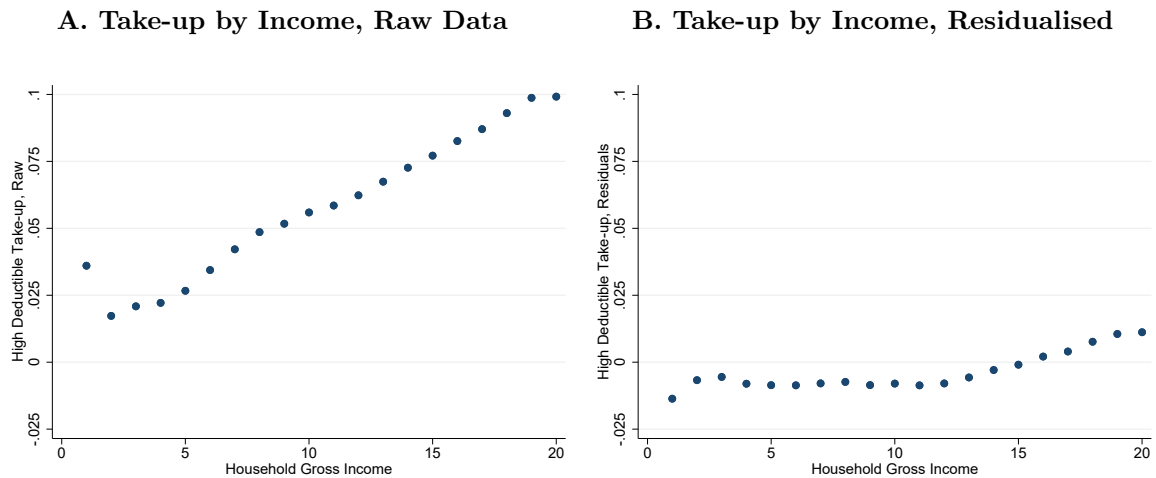
## A.5 Deductible Choice: Appendix Figures and Tables

TABLE A.4: DEDUCTIBLE TAKE-UP: IMPACT OF HEALTH AND INCOME CHANGES

	(1)	(2)	(3)	(4)	(5)
	No FE	Individual FE	First difference	First difference	First difference
Probability of Low Costs	0.115***	0.0570***	0.0422***		
Prob. Low Costs, Positive $\Delta$				0.00691***	
Prob. Low Costs, Negative $\Delta$				-0.0670***	
$\Delta$ Prob. Low Costs > +2 Deciles					0.0102***
$\Delta$ Prob. Low Costs = +2 Deciles					0.00685***
$\Delta$ Prob. Low Costs = +1 Decile					0.00342***
$\Delta$ Prob. Low Costs = -1 Decile					-0.00277***
$\Delta$ Prob. Low Costs = -2 Deciles					-0.00636***
$\Delta$ Prob. Low Costs < -2 Deciles					-0.0202***
Income ('000 EUR)	6.06e-05***	1.57e-05***	6.63e-06***	6.65e-06***	6.85e-06***
Number of Individuals	12,317,248	12,317,248	12,074,444	12,058,624	12,074,444
Observations	47,685,794	47,685,794	35,368,540	35,216,196	35,368,540

**Notes:** This table presents the result of an OLS regression of take-up of the 500 EUR extra deductible on probability of low costs, changes in probability of low costs, income, and changes in income. In column (1), take-up of the high deductible is regressed on the probability to have health costs lower than 375 EUR, and on income in thousands of EUR. Column (2) adds individual fixed effects. Column (3) regresses the first difference of deductible take-up on the first difference of the probability of low costs and the first difference of income. Column (4) splits the first difference in two distinct variables, one containing only positive shocks, the other only negative shocks. Column (5) creates six dummies capturing shocks of various magnitudes: positive and negative shocks of one, two, and strictly more than two deciles. In Columns (4) and (5), income first difference remains unchanged compared to Column (3). All regressions include year fixed effects. \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$  with robust standard errors.

FIGURE A.8: DEDUCTIBLE TAKE-UP AS A FUNCTION OF INCOME



**Notes:** These figures plot the relationship between household gross income and the take-up of the 500 EUR extra deductible. Panel A plots take-up of 500 deductible by household income percentile. Panel B plots the residuals of an OLS regression of take-up of 500 EUR extra deductible on risk probability, four levels of education dummies, four age dummies, and indicators for gender, having a partner, and having children.

TABLE A.5: ROBUSTNESS CHECK

	Baseline			500 vs. 0 Deductible			0 vs. >0 Deductible		
	Without	With Interaction		Without	With Interaction		Without	With Interaction	
	Interaction	<i>intercept</i>	<i>slope</i>	Interaction	<i>intercept</i>	<i>slope</i>	Interaction	<i>intercept</i>	<i>slope</i>
High School	0.017***	-0.011***	0.057***	0.018***	-0.012***	0.061***	0.025***	-0.014***	0.077***
College Degree	0.065***	-0.034***	0.165***	0.071***	-0.038***	0.181***	0.089***	-0.037***	0.210***
Further Studies	0.091***	-0.047***	0.226***	0.099***	-0.052***	0.250***	0.123***	-0.044***	0.275***
2nd Income Quartile	-0.003***	0.004***	-0.007***	-0.003***	0.004***	-0.007***	0.002***	0.009***	-0.005***
3rd Income Quartile	0.004***	0.004***	0.007***	0.005***	0.003***	0.009***	0.014***	0.011***	0.013***
4th Income Quartile	0.024***	0.002***	0.039***	0.026***	0.001***	0.045***	0.041***	0.015***	0.048***
36 to 50 years old	-0.011***	0.020***	-0.045***	-0.010***	0.022***	-0.046***	-0.006***	0.024***	-0.042***
51 to 65 years old	-0.004***	0.029***	-0.047***	-0.004***	0.030***	-0.048***	0.003***	0.036***	-0.045***
65+ years old	-0.001***	0.034***	-0.082***	0.000**	0.036***	-0.085***	0.007***	0.043***	-0.092***
Male	0.011***	-0.004***	0.025***	0.012***	-0.004***	0.028***	0.017***	-0.001***	0.030***
Has Partner	0.003***	-0.002***	0.013***	0.003***	-0.002***	0.014***	0.002***	-0.005***	0.018***
Has Children	-0.010***	0.004***	-0.028***	-0.011***	0.004***	-0.031***	-0.014***	0.004***	-0.035***
Self-employed	0.009***	-0.006***	0.026***	0.009***	-0.007***	0.028***	0.013***	0.000	0.023***
Constant	-0.042***	-0.041***		-0.045***	-0.043***		-0.055***	-0.044***	
Prob. Low Costs	0.122***		0.098***	0.129***		0.100***	0.169***		0.124***
Year and Insurer FE	YES	YES		YES	YES		YES	YES	
Observations	57,100,388	57,100,388		55,335,880	55,335,880		57,100,388	57,100,388	

	Baseline			Probit			Binary Pred. Low Costs		
	Without	With Interaction		Without	With Interaction		Without	With Interaction	
	Interaction	<i>intercept</i>	<i>slope</i>	Interaction	<i>intercept</i>	<i>slope</i>	Interaction	<i>intercept</i>	<i>slope</i>
High School	0.017***	-0.011***	0.057***	0.022***	0.006***	0.023***	0.019***	0.002***	0.032***
College Degree	0.065***	-0.034***	0.165***	0.051***	0.014***	0.051***	0.068***	0.013***	0.081***
Further Studies	0.091***	-0.047***	0.226***	0.063***	0.005**	0.081***	0.093***	0.019***	0.105***
2nd Income Quartile	-0.003***	0.004***	-0.007***	0.003***	0.030***	-0.040***	-0.001***	0.003***	-0.005***
3rd Income Quartile	0.004***	0.004***	0.007***	0.010***	0.041***	-0.044***	0.007***	0.008***	0.002***
4th Income Quartile	0.024***	0.002***	0.039***	0.024***	0.057***	-0.048***	0.027***	0.017***	0.016***
36 to 50 years old	-0.011***	0.020***	-0.045***	-0.008***	-0.007***	-0.001	-0.013***	-0.005***	-0.009***
51 to 65 years old	-0.004***	0.029***	-0.047***	0.000	0.001**	-0.002**	-0.012***	-0.005***	-0.006***
65+ years old	-0.001***	0.034***	-0.082***	-0.011***	-0.008***	-0.004***	-0.017***	-0.008***	-0.025***
Male	0.011***	-0.004***	0.025***	0.006***	0.007***	-0.002***	0.015***	0.001***	0.023***
Has Partner	0.003***	-0.002***	0.013***	0.005***	0.007***	-0.003***	0.003***	0.000***	0.008***
Has Children	-0.010***	0.004***	-0.028***	-0.009***	-0.000	-0.011***	-0.011***	-0.000	-0.020***
Self-employed	0.009***	-0.006***	0.026***	0.008***	0.018***	-0.013***	0.011***	0.005***	0.009***
Constant	-0.042***	-0.041***					-0.014***	-0.003***	
Prob. Low Costs	0.122***		0.098***	0.169***		0.191***			
Pred. Costs <375							0.062***		0.034***
Year and Insurer FE	YES	YES		YES	YES		YES	YES	
Observations	57,100,388	57,100,388		57,100,388	57,100,388		57,100,388	57,100,388	

**Notes:** This table performs a range of robustness checks on our baseline results. In the top panel, we compare our baseline regression with alternative definition of take-up of the high deductible. In the baseline, we define take-up as choosing the 500 deductible, as opposed to choosing any other deductible. In the second top panel, we keep only choices that are the 500 or the 0 deductible, and drop intermediate choices. In the third top panel, we instead define take-up as choosing any deductible strictly greater than 0. In the second bottom panel, we compare our baseline OLS regression with a probit specification. Finally, in the third bottom panel, we replace our linear probability of low costs with a binary indicator taking value one if the individual is predicted to have health costs lower than 375 EUR. In each panel, we present a regression with and without interacting our regressors with the probability of low costs.

TABLE A.6: DEDUCTIBLE TAKE-UP AND PREDICTED HEALTH BY FIELD

Education Field	(1) Take-up of 500 Deductible	(2) Probability Low Costs	(3) Take-up of 500 Ded.   Being Predictably Healthy
1 <b>Statistics</b>	29%	87%	34%
2 Mathematics	21%	85%	27%
3 Physics	21%	91%	26%
4 Architecture and town planning	18%	88%	21%
5 Physical science	18%	82%	22%
6 Earth science	18%	88%	21%
7 <b>Philosophy and ethics</b>	17%	82%	21%
8 Medicine	17%	83%	20%
9 Chemistry	16%	87%	20%
10 Biology and biochemistry	16%	83%	20%
11 Science, Mathematics and Computing	16%	85%	19%
12 Computer science	15%	87%	18%
13 Environmental protection	15%	86%	18%
14 Political science and civics	15%	85%	18%
15 Design	15%	85%	18%
16 Sociology and cultural studies	14%	82%	18%
17 Mining and extraction	14%	91%	17%
18 Economics	14%	84%	17%
19 Humanities and Arts	14%	84%	18%
20 Dental studies	14%	76%	18%
21 History and archaeology	13%	82%	16%
22 Business and administration	13%	82%	16%
23 Pharmacy	13%	73%	17%
24 Health	13%	79%	16%
25 Environmental protection technology	13%	84%	15%
26 Medical diagnostic and treatment technology	13%	81%	16%
27 Religion	13%	80%	17%
28 Law	13%	80%	16%
29 Psychology	12%	77%	16%
30 Management and administration	12%	81%	16%
31 Engineering and engineering trades	12%	87%	15%
32 Forestry	12%	86%	14%
33 Therapy and rehabilitation	12%	78%	15%
34 Finance, banking, insurance	12%	80%	15%
35 Social and behavioural science	12%	79%	15%
36 Health and Welfare	12%	80%	15%
37 Fisheries	12%	94%	15%
38 Journalism and reporting	12%	80%	14%
39 Training for teachers w. subject specialisation	11%	79%	14%
40 Education science	11%	75%	14%
41 <b>Accounting and taxation</b>	11%	78%	14%
42 Agriculture, forestry and fishery	10%	81%	13%
43 <b>Marketing and advertising</b>	10%	80%	13%
44 Chemical and process	10%	85%	12%
45 Arts	10%	80%	13%
46 Electronics and automation	10%	86%	12%

TABLE A.6: DEDUCTIBLE TAKE-UP AND PREDICTED HEALTH BY FIELD (CONT'D)

47 Music and performing arts	10%	81%	12%
48 Training for teachers of vocational subjects	10%	81%	12%
49 Fine arts	10%	82%	12%
50 Humanities	10%	76%	12%
51 Library, information, archive	9%	78%	12%
52 Travel, tourism and leisure	9%	77%	12%
53 Electricity and energy	9%	88%	11%
54 Veterinary	9%	75%	12%
55 Mother tongue	9%	74%	12%
56 Audio-visual techniques and media production	9%	83%	10%
57 Building and civil engineering	9%	86%	10%
58 Life science	9%	79%	11%
59 Crop and livestock production	9%	79%	11%
60 Mechanics and metal work	9%	85%	10%
61 Wholesale and retail sales	8%	79%	11%
62 Foreign languages	8%	74%	11%
63 Motor vehicles, ships and aircraft	8%	87%	10%
64 Training for teachers at basic levels	8%	75%	10%
65 Materials (wood, paper, plastic, glass)	8%	86%	9%
66 Sports	8%	83%	10%
67 Teacher training and education science	8%	74%	10%
68 Military and defence	7%	81%	9%
69 Transport services	7%	83%	9%
70 Food processing	7%	78%	9%
72 Natural environments and wildlife	6%	86%	7%
73 Hotel, restaurant and catering	6%	77%	8%
74 Basic / broad, general programmes	6%	72%	9%
75 Social work and counselling	6%	70%	8%
77 Personal skills	6%	68%	8%
78 Textiles, clothes, footwear, leather	5%	70%	7%
79 Horticulture	5%	80%	6%
80 General Programmes	5%	71%	7%
81 Nursing and caring	5%	66%	7%
82 Domestic services	5%	66%	7%
83 Secretarial and office work	5%	65%	7%
84 <b>Protection of persons and property</b>	4%	78%	6%
85 Child care and youth services	4%	66%	6%
86 Computer use	4%	65%	6%
87 <b>Hair and beauty services</b>	4%	65%	5%
88 Occupational health and safety	4%	75%	5%
89 Training for pre-school teachers	3%	62%	0%
90 Literacy and numeracy	2%	62%	4%

**Notes:** For each field of study, this table shows: in Column (1), the fraction of individuals who take-up the 500 EUR extra deductible, in Column (2), the fraction of individuals with a probability of low costs < 375 EUR, and in Column (3), the fraction of individuals who take-up the 500 EUR extra deductible, conditional on having predicted health costs < 375 EUR.

TABLE A.7: DEDUCTIBLE TAKE-UP AND PREDICTED HEALTH BY PROFESSIONAL SECTOR

Professional Sector	(1) Take-up of 500 Deductible	(2) Probability Low Costs	(3) Take-up of 500 Ded.   Being Predictably Healthy
1 <b>Business Services II</b>	13%	84%	16%
2 <b>Insurance and Health Insurance Firms</b>	12%	79%	15%
3 Business Services I	12%	82%	15%
4 Dairy Industry	12%	82%	14%
5 Banks	10%	81%	12%
6 Other Passenger Transport Land and Air	10%	79%	13%
7 Business Services III	10%	79%	13%
8 Agriculture	10%	85%	11%
9 Stoneware	9%	83%	11%
10 Publishers	9%	79%	11%
11 Cultural Institutions	9%	80%	11%
12 Telecommunications	9%	81%	12%
13 Government, Education and Science	9%	75%	12%
14 Food Industry	9%	80%	11%
15 Catering Industry I	9%	84%	10%
16 Tobacco Processing Industry	9%	76%	11%
17 Wholesale I	8%	82%	11%
18 Wholesale II	8%	81%	10%
20 Government, Police and Judiciary	8%	74%	11%
21 Wholesale of Wood	8%	82%	10%
22 Electronic Industry	8%	81%	13%
23 Carpentry	8%	83%	9%
24 Furniture and Organ Building	8%	83%	9%
25 Rail Construction	8%	78%	11%
26 NS Transport	8%	74%	10%
27 Sugar Processing Industry	7%	78%	10%
28 Chain Stores	7%	80%	9%
29 <b>Retail</b>	7%	79%	9%
30 Lending Industry	7%	81%	9%
31 Other Branches of Business	7%	79%	9%
32 Postal Transport	7%	72%	10%
33 Metal Industry	7%	80%	10%
34 <b>Construction</b>	7%	83%	9%
35 Merchant	7%	89%	8%
36 Mortar	7%	72%	9%
37 KLM Transport	7%	77%	9%
38 Bakeries	7%	79%	9%
39 Metal and Technical Industry	7%	82%	8%
40 Port Companies	7%	82%	9%
41 Chemical Industry	7%	79%	9%
42 General Industry	7%	81%	9%
43 Stone, Cement, Glass and Ceramic Industry	7%	77%	9%
44 Butchers Other	7%	80%	8%
45 Health, Mental and Social Industry	7%	71%	9%
46 Printing Industry	7%	80%	8%
47 Textiles Industry	7%	77%	9%
48 Inland Shipping	7%	83%	8%
49 Private Bus Transport	6%	70%	9%

TABLE A.7: DEDUCTIBLE TAKE-UP AND PREDICTED HEALTH BY PROFESSIONAL SECTOR (CONT'D)

50 Government, Local Government	6%	70%	9%
51 Butchers	6%	79%	8%
52 Wood, Brush and Packaging Industry	6%	82%	8%
53 Other Goods Transport Land and Air	6%	80%	8%
54 Government, Defense	6%	82%	11%
55 <b>Government, Public Utilities</b>	6%	77%	7%
56 Public Transport	5%	65%	8%
57 Security	5%	75%	7%
58 Plastering	5%	85%	6%
59 Taxi and Ambulance	5%	65%	8%
60 Catering Industry II	5%	70%	7%
61 Painting Industry	5%	81%	6%
62 Port Classifiers	5%	79%	6%
63 Fishing	4%	81%	6%
64 Work and Integration	4%	64%	6%
65 Dredging Industry	4%	85%	9%
66 Government, Other Institutions	4%	60%	7%
67 Roofing	4%	82%	5%
68 <b>Cleaning</b>	3%	70%	5%

**Notes:** For each professional sector, this table shows: in Column (1), the fraction of individuals who take-up the 500 EUR extra deductible, in Column (2), the fraction of individuals with a probability of low costs < 375 EUR, and in Column (3), the fraction of individuals who take-up the 500 EUR extra deductible, conditional on having predicted health costs < 375 EUR.

## A.6 Model Assumptions and Structural Choice Foundations

This section discusses some of the simplifying model assumptions and presents empirical evidence that suggests that relaxing these assumptions would not qualitatively change any of the conclusions. We then assess what kinds of micro-foundations can in principle rationalize the decision-making patterns that we document. This could also allow for a further refinement of any welfare analysis and policy recommendations.

### A.6.1 Simplified Model

Each individual is subject to a compulsory deductible of 375 and can choose an extra deductible  $d$  at corresponding premium  $p$  from menu  $\Omega = \{(d, p_d)\}$ . An individual draws health cost  $x$  from an individual-specific distribution  $F_i(x)$ . Depending on her deductible choice  $d$ , health cost translates into an out-of-pocket expense  $s = \min\{d, x\}$ . We can denote by  $G_{i,d}(s)$  the distribution of out-of-pocket spending, derived from  $F_i(x)$  and the deductible choice  $d$ . Expected utility in this environment is defined as:

$$(3) \quad U_{i,d} = \int u_i(W_i - p_d - s)G_{i,d}(s)ds.$$

Using this definition of expected utility, we can define an individual's certainty equivalent from choosing one contract as  $CE_{i,d}$ , where  $U_{i,d} = u_i(W_i - CE_{i,d})$ .

In practice, there are 6 different deductible choices possible: 375, 475, 575, 675, 775 and 875 EUR. We simplified the deductible choice problem into a binary choice between selecting a 875 EUR deductible, or the mandatory 375 EUR deductible. The contract space thus simplifies to:

$$\Omega = \{(0, 0), (500, -250)\}.$$

Under the simplified environment proposed in Section II, we approximate expected utility by:

$$(4) \quad U_{i,d} \approx \pi_i u_i(W_i - p_d) + (1 - \pi_i) u_i(W_i - p_d - d),$$

where  $\pi_i$  denotes the individual's probability to draw health cost  $x$  below 375 EUR.

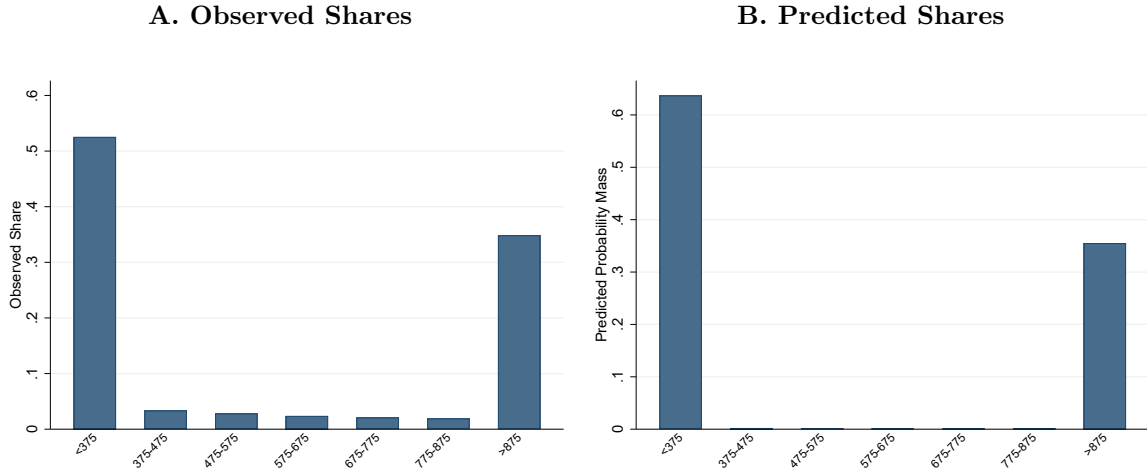
### A.6.2 Binary Choice and Risk

Several pieces of evidence suggest that reducing the problem to a binary choice with binary risk is indeed appropriate for our context.

First, while in theory the optimal decision depends on the probability distribution of expenditures between 375 EUR and 875 EUR too, the share of expenditures that fall in this range is small. Indeed, Panel A of Figure A.9 shows that the ex-post observed shares within each intermediary deductible bracket are small. This means that only a small fraction of individuals fall into the intermediary deductible ranges, which decreases the likelihood that the intermediary deductibles are optimal choices. Second, we find that when using a machine learning classifier to predict which individuals are going to fall into the intermediary brackets, the predicted mass in these intermediary brackets is small. Panel B of Figure A.9 shows that ex-ante, a random forest model trained on an unbalanced sample will give less than 1% probability mass to the intermediate categories. This is largely due to the unbalanced classes, where the majority of individuals fall into the lowest or highest bracket. However, insofar as we cannot expect individuals to predict their future costs more accurately, the low probability with which most individuals are predicted to be in the intermediary deductible brackets further strengthens the case

for a binary decision rule.

FIGURE A.9: COST PREDICTIONS WITH MULTIPLE DEDUCTIBLE CATEGORIES



**Notes:** Panel A plots the observed share of individuals with health costs in all the deductible health cost brackets in 2015. Panel B plots the predicted shares of individuals in all deductible health cost brackets, where the prediction is from a random forest with the same predictors as described in Section ??.

The large majority of individuals (more than two thirds) who choose a higher deductible opt for the highest deductible. In our baseline analysis, we define take-up as choosing the 500 deductible, as opposed to choosing any other deductible. We analyze the sensitivity of our results to alternative definitions of take-up of the high deductible: we can keep only choices that are the 500 or the 0 deductible, and drop intermediate choices. Or, we can instead define take-up as choosing any deductible strictly greater than 0. The regression estimates remain very similar when using these alternative definitions, as shown in Appendix Table A.5.

### A.6.3 Asymmetric Information and Moral Hazard

Moral hazard could cause consumers to reduce care consumption in response to greater cost sharing (e.g., Newhouse (1993), Einav, Finkelstein and Schrimpf (2015), Brot-Goldberg et al. (2017)). Under a classical model of moral hazard, our framework under-predicts value from the high deductible plan since it rules out reductions in care that are lower in value than the associated cost savings. Since our empirical results focus on significant under-adoption of higher deductibles, having the lower bound interpretation does not impact the main import of our results. That is, to the extent that consumers spend less under a high deductible plan because of classical moral hazard, our model threshold for choosing the high deductible ( $\pi = 0.5$  for risk neutral,  $\pi = 0.56$  for very risk averse) is slightly high (i.e. more people should choose the high deductible) and the normative benefits from doing so in Section III.C are too low, working against our main results.

We can also shed empirical light on the potential role of moral hazard in our context. Panel A of Figure A.7 addresses the issue whether enrolling in a higher deductible impact ex post cost realization relative to our prediction model? This combines the classic issues of selection on private information (not included in our prediction model) and moral hazard (ex post utilization changes due to different prices). See Chiappori and Salanie (2000) for prior research describing a correlation test to jointly detect the presence of adverse selection on private information and moral hazard. Panel A shows that individuals who choose a 500 EUR deductible are more likely to have low costs than individuals who choose no extra deductible, conditional on the prediction of

our model. This is consistent with some combination of additional selection on private information and moral hazard. However, the difference in the *ex post* realized low cost fraction relative to the predicted fraction is quite small, leading us to conclude that the private information and moral hazard, conditional on our predictors, is small. More specifically, the average gap across probability bins between individuals who choose and who do not choose an extra deductible is 6.667%. Taking into account that across probability bins, the average share with low costs among people without an extra deductible is 51.215%, we find that individuals who take a deductible are on average are 13.017% more likely to have low costs than our model predicts. This also corroborates earlier evidence in the Dutch context (Remmerswaal, Boone and Douven (2023)).

Finally, Table A.8 below sheds light on potential behavioral hazard (Baicker, Mullainathan and Schwartzstein (2015)) and suggests that up front rational avoidance of ex post behavioral hazard is not a major concern. Specifically, conditional on one’s predicted low cost bin, there are minimal impacts on key high-value care categories like preventive care, maternity care and drugs. There is a more significant impact on mental health spending, indicating some potential for moral hazard and or selection on private information in that domain.

TABLE A.8: EX POST HEALTH EXPENSES, BY SUBGROUPS

	P(Low Costs)	Low Deductible	Any Incremental Deductible
N (Sample Size)			
0.6-0.7		1,156,446	91,263
0.7-0.8		1,514,402	171,016
0.8-0.9		1,850,417	298,369
0.9-1		471,746	96,877
Preventative Care (Always Insured)			
0.6-0.7		184.6	171.7
0.7-0.8		154.3	142.3
0.8-0.9		122.9	113.5
0.9-1		97.3	90.7
Drugs			
0.6-0.7		68.7	55.5
0.7-0.8		45.6	35.7
0.8-0.9		25.6	19.1
0.9-1		13.0	9.6
Maternity Care			
0.6-0.7		41.8	42.1
0.7-0.8		27.8	26.0
0.8-0.9		14.4	11.2
0.9-1		4.6	2.7
Mental Health			
0.6-0.7		234.3	173.2
0.7-0.8		155.5	117.0
0.8-0.9		98.0	66.1
0.9-1		64.9	38.0

**Notes:** This table presents statistics related to actual ex post spending on certain types of health care as a function of our ex ante prediction of the probability an individual has low costs. The top section gives the sample size for each group and subsequent sections give the mean EUR spent on each kind of care by individuals in each group. This table supports the discussion of behavioral hazard in Section II, suggesting that up front rational avoidance of ex post behavioral hazard is not a major concern.

#### A.6.4 Models of Choice Barriers

We first consider a model with default effects. Switching costs occur when consumers with a default plan option must pay some cost  $c_s$  to switch plans. This could be, e.g., a paperwork / transaction cost or reflect some reduced form of a multi-stage model with search and search costs. See a discussion of potential inputs into switching costs in [Handel \(2013\)](#). Specifically, setting the low deductible as the default plan option, a consumer chooses the high deductible if:

$$(5) \quad 250 - (1 - \pi)500 - c_s > 0$$

This assumes the model premium reduction of 250 EUR when taking the 500 EUR deductible. We consider heterogeneous population switching costs  $c_s \sim U(0, 2 \times \bar{c}_s)$  for different average switching costs  $\bar{c}_s$ . [Brot-Goldberg et al. \(2023\)](#) find strong default effects in Medicare Part D and show this is primarily due to inattention rather than switching costs. Note that we could alternatively model the default effects by for example allowing for a heterogeneous probability  $\mu$  with which an individual is attentive and optimizes her deductible choice. Otherwise, she sticks to the default low deductible. The predicted choice patterns would be very similar.

Loss aversion occurs when losses loom larger than gains. In contrast with standard risk aversion, loss aversion can reduce the take-up of a deductible even when financial stakes are small. See [Sydnor \(2010\)](#) for a discussion of loss aversion as a potential driver of the over-insurance of modest risks. Following [Kőszegi and Rabin \(2007\)](#), we assume that realized payoffs are evaluated relative to expected payoffs, conditional on the deductible choice made, and losses receive a relative weight  $\lambda$ . In our setup, agents will then choose the high deductible if:

$$(6) \quad 250 - (1 - \pi)500 - (\lambda - 1)\pi(1 - \pi)500 > 0.$$

Decisions could be made based on imperfect information. In our context, imperfect information enters by allowing consumers to receive a noisy signal  $\hat{\pi}$  about their health, where  $\hat{\pi} = \pi + \epsilon$  and  $\epsilon \sim N(0, \sigma_\epsilon)$ . They make a decision based on that noisy signal and choose the high deductible (for the model premium reduction of 250) if and only if

$$(7) \quad 250 - (1 - \hat{\pi})500 > 0.$$

where the signal-to-noise ratio equals  $\sigma_\pi / \sigma_\epsilon$ .

Alternatively, individuals may decide rationally whether to pay attention and acquire information. In our context, rational inattention means that consumers, again, receive a noisy signal about their health, but then decide whether or not to pay a cost  $c_r$  to learn the true value of his/her health risk. Upon receiving the signal, agents face an expected choice value that integrates over the probability distribution of their potential true health statuses.<sup>12</sup> The value of acquiring the accurate information depends on whether the information would change her deductible choice and thus on the condition density  $f(\pi|\hat{\pi})$  for  $\pi > .5$  and  $\hat{\pi} < .5$  and vice versa.<sup>13</sup> The result

<sup>12</sup>Our model is similar in spirit to that laid out in [Ho, Hogan and Scott Morton \(2017\)](#), though there consumers obtain signals about plan characteristics while here they about signals about their own health status. We could recast our model as related to uncertainty about plan characteristics, likely with similar results.

<sup>13</sup>We simulate the conditional density by taking random draws from the empirical distribution of  $\pi$  and the normal distribution of  $\epsilon$ . We then group the resulting  $\pi$  and  $\hat{\pi}$  in ten bins of length 0.1, indexing them from 1 to 10. Then for each bin  $j$  of  $\hat{\pi}$ , we approximate the conditional density using:

$$p(\pi \in \pi_k | \hat{\pi} \in \hat{\pi}_j) = \frac{\#\text{individuals} \in \{\hat{\pi}_j \cap \pi_k\}}{\#\text{individuals} \in \hat{\pi}_j}$$

of our rational inattention setup is that, if a consumer starts with the low deductible, they will choose the high deductible if and only if one of the following conditions holds:

$$(8) \quad \hat{\pi} > 0.5 \text{ and } \int_0^{0.5} [-250 + (1 - \pi)500]f(\pi|\hat{\pi}) \, d\pi < c_r$$

$$(9) \quad \hat{\pi} > 0.5 \text{ and } \int_0^{0.5} [-250 + (1 - \pi)500]f(\pi|\hat{\pi}) \, d\pi > c_r \text{ and } \pi > 0.5$$

$$(10) \quad \hat{\pi} \leq 0.5 \text{ and } \int_{0.5}^z 1[250 - (1 - \pi)500]f(\pi|\hat{\pi}) \, d\pi > c_r \text{ and } \pi > 0.5$$

The first condition results when consumers are so confident they are low that they don't find it worthwhile to pay the cost of precisely determining their health status, instead just electing to choose the high deductible right away. The second and third conditions occur when consumers decide to pay the cost to obtain a more precise signal, and are differentiated only by whether the initial signal value is bigger or smaller than the risk-neutral threshold of  $\pi = 0.5$  for high deductible choice under the modal premium reduction.

Finally, consumers may simply make mistakes. In our model, we assume a share  $1 - \alpha$  of agents make rational, frictionless choices, while share  $\alpha$  of agents make random choices.

#### A.6.5 Simulations

Figure A.10 presents simulations of the deductible take-up rate as a function of health risk for the alternative decision models. For comparison, each panel plots the observed take-up rates and the deductible choice for the case where consumers are rational, frictionless, and risk-neutral, as in Figure 2. As discussed before, in a frictionless world, all consumers below a 50% probability of clearing the low deductible will elect the high-deductible, which looks starkly different from the observed low take-up rates. Risk-aversion only slightly alters this threshold, moving it to a marginally higher probability of low spending for the case where consumers are risk-averse with CARA coefficient of  $1 * 10^{-4}$  (Panel A).

We then turn to the simulations for a decision models with switching costs. Note that with a homogeneous switching cost of 119 EUR, about 10 percent of the population would take up the high deductible, which corresponds to the observed take up rate. However, with heterogeneous switching costs uniformly distributed around the same mean of 119 EUR, we still predict meaningfully more high deductible purchases than we observe in the data, especially as consumers become predictably healthier and healthier. Heterogeneous switching costs with a higher mean of 650 EUR (panel B) look much more similar to observed purchases as a function of health status. But this specification still predicts no purchasing of a high deductible for consumers with higher predicted probabilities of higher health spending. However, when we combine our model of high switching costs with our model of imperfect information about health status (with an assumed signal-to-noise ratio of 1), the simulated choices as a function of health status map very closely to observed choices (panel C).

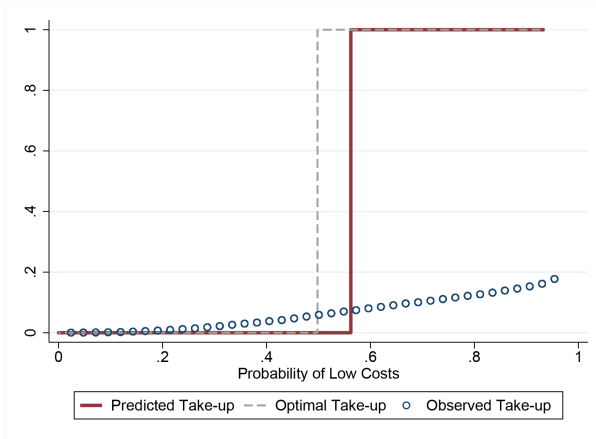
Like switching costs for taking up the high deductible, loss-aversion helps to reduce the take-up rate of individuals around the 50% threshold. But similarly as for the case of risk aversion, the simulated take-up rates remain too high for reasonable loss-aversion parameters. Panel D simulates the deductible choices for a loss-aversion parameter of  $\lambda = 2.25$  (i.e., when choosing the high deductible the payoff is reduced by  $(2.25 - 1)\pi(1 - \pi)500$ ). Even with such strong loss aversion, individuals in very good health are predicted to always take up the deductible as the variance in financial payoffs they would get exposed to converges to zero.

---

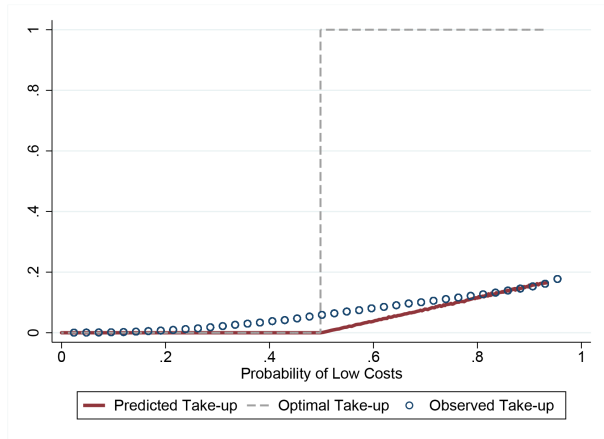
where  $\pi_k$  is bin  $k$  of  $\pi$ , and  $\hat{\pi}_j$  is bin  $j$  of  $\hat{\pi}$ . To calculate the expected payoff, we use the middle value of each bin  $k$  of  $\pi$ .

FIGURE A.10: DEDUCTIBLE TAKE-UP FOR DIFFERENT BEHAVIORAL MODELS

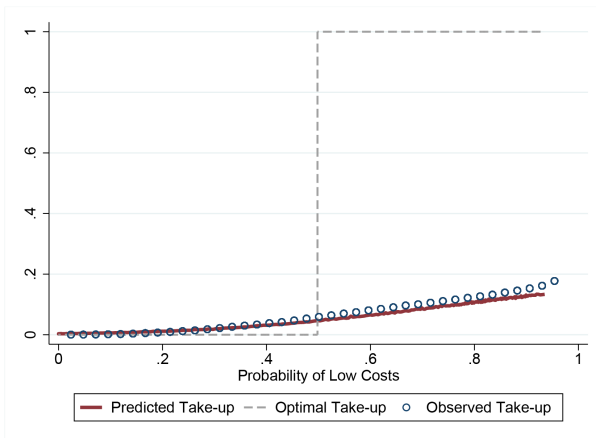
**A. Optimal Choice**



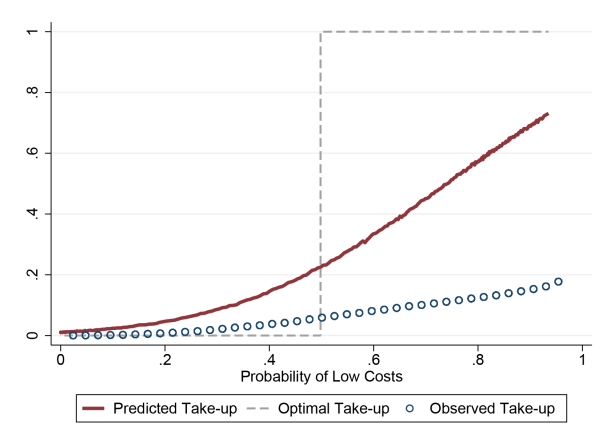
**B. Heterogeneous Switching Costs**



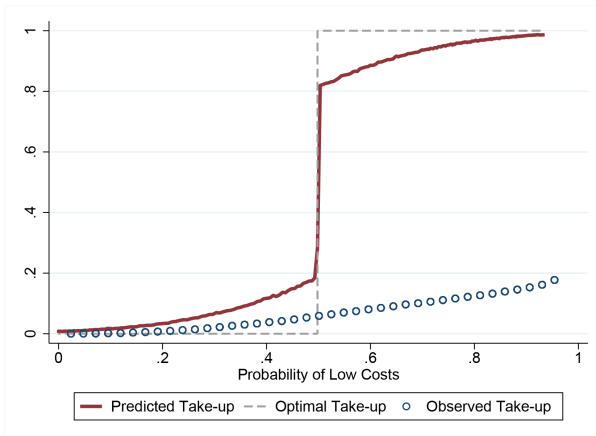
**C. Hetero. Switching Costs and Imperfect Info**



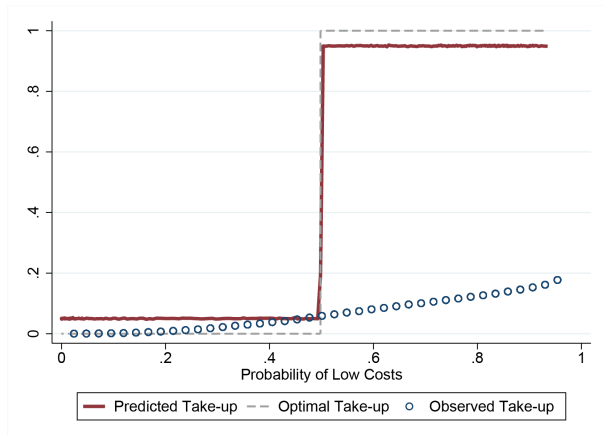
**D. Loss Aversion and Imperfect Info**



**E. Rational Inattention**



**F. Mistakes**



**Notes:** This figure presents the results from decision-making simulations for the various models discussed in detail in the text. For each model, we contrast the predicted take-up rate with both the observed take-up rate and the take-up rate by a rational consumer in a frictionless world.

Figure A.10 also presents results for the rational inattention model (panel E) and the random mistakes model (panel F). The simulations for the rational inattention model use an information acquisition cost of  $c_r = 25$  (for much higher values, no one pays this cost to learn about their true health status, making the model's predictions the same as the imperfect information model). We see that the take-up rate becomes more responsive to health risk around the threshold value, since individuals have to have probabilistic signals close to the marginal thresholds to acquire information, even with a reasonably small cost of 25 EUR. Furthermore, consumers with larger probabilities of being healthy are predicted to purchase the higher deductible much more than they actually do in practice. So we would need to combine the model of rational inattention with high switching costs to obtain predictions that are closer to observed choices. The simulations for the random mistakes model assume that a random 10% of consumers make mistakes. Clearly, the overall take-up rate is too high, so we again need an extra force to lower the take-up rate. Moreover, in the random mistakes model, the take-up rate is now also too high for individuals who are predicted to have high costs. This would not be resolved by combining the mistakes model with the imperfect information model.

This section illustrates how simulations based on different choice models compare with our data. Though there are a plethora of models one could write down that could help rationalizing the data (e.g., inertia, limited attention), a model of high switching costs combined with imperfect information fits the data very well. Importantly, high switching costs would further decrease the welfare gains from offering deductible choice. While we don't structurally estimate these models in our current context, these simulations give a sense of what models might make sense to estimate, and potentially test formally vs. one another, to implement a more detailed investigation of the mechanisms underlying the choice patterns we have documented.

## A.7 Choice Quality Ranking

This Appendix Section provides further details underlying our analysis of choice quality rankings in Section III.C.

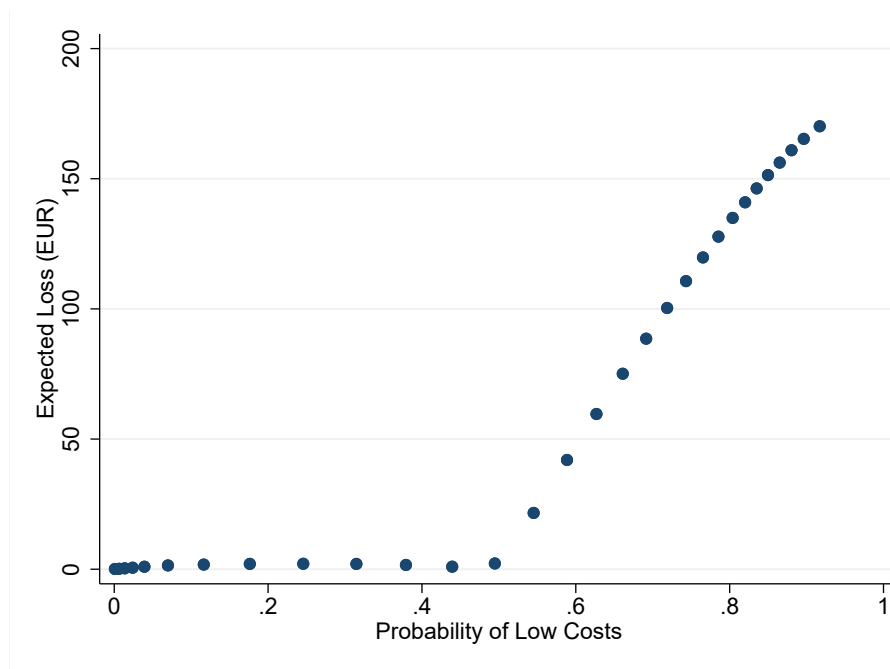
### A.7.1 Consumer Welfare

To evaluate choice quality, we calculate the potential cost savings for each individual by comparing her deductible choice  $C_i$  to the choice that minimizes her expected out-of-pocket expenditures. These cost savings simply correspond to the difference in certainty equivalents for risk-neutral preferences:

$$\Delta w_i^{*,\sigma=0} = CE_i^{*,\sigma=0} - CE_i^{\sigma=0}.$$

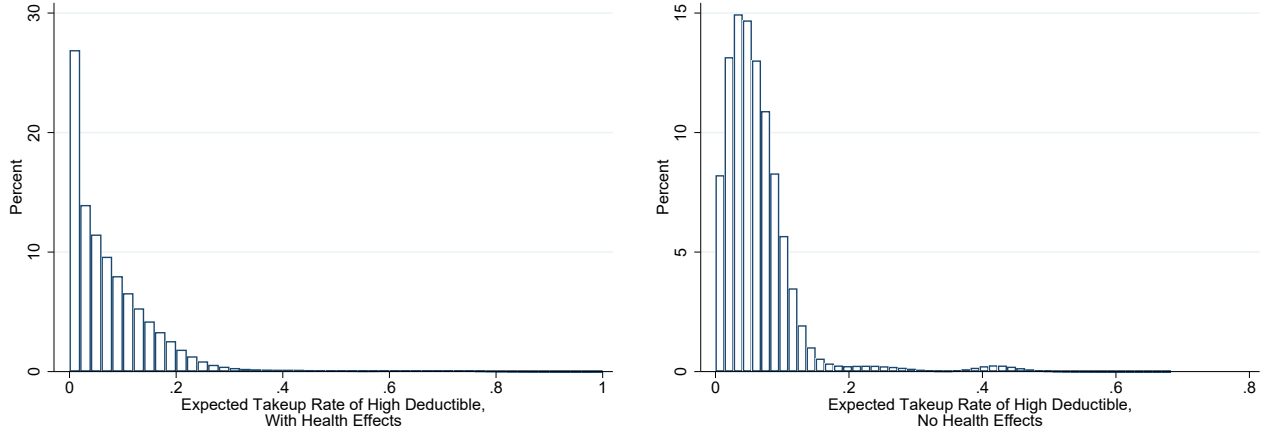
In our simplified, it is optimal (in expected payoff terms) for an individual to choose the extra 500 EUR deductible if the predicted probability to make lower costs is higher than 50 percent, leaving no potential costs savings on the table. Figure A.11 shows the expected loss due to over- or under-insurance as a function of the predicted costs. For individuals with a predicted probability of low costs below 0.5, the expected losses due to under-insurance are very small (on average close to zero), as a very low fraction of people under-insures by taking the 500 EUR extra deductible. For individuals with a predicted probability of low costs above 0.5, expected losses due to over-insurance increase with this probability, and reach almost 170 EUR for people with a very high chance (0.9+) of low costs, as most people leave money on the table by over-insuring for costs that happen with a very low probability.

FIGURE A.11: EXPECTED LOSS AND HEALTH COST PROBABILITY



**Notes:** This figure is a binned scatterplot of the relationship between the predicted probability of health costs below 375 EUR and the expected loss due to over- or under-insurance. For individuals with a predicted probability of low costs below 0.5, the expected losses due to under-insurance are very small (on average close to zero), as a very low fraction of people under-insures by taking the 500 EUR extra deductible. For individuals with a predicted probability of low costs above 0.5, expected losses due to over-insurance increase with this probability, and reach almost 170 EUR for people with a very high chance (0.9+) of low costs, as most people leave money on the table by over-insuring for costs that happen with a very low probability.

FIGURE A.12: PREDICTED DEDUCTIBLE CHOICE



**Notes:** This figure shows the distribution of predicted 500 EUR extra deductible take-up rate. Panel A shows the predicted 500 EUR deductible take-up with health effects, while Panel B shows the take-up without the health effects.

### A.7.2 Predicted Choice Model

We predict the deductible take-up rate  $d(X_{it}, \pi_{it})$  as a function of their predicted health  $\pi_{it}$ , observables  $X_{it}$  and their interaction by running the regression:

$$Y = \alpha + \sum \beta_{\delta} 1[\pi = \delta] + \gamma X + \sum \nu_{\delta} 1[\pi = \delta] X + \epsilon$$

Here,  $Y$  is a binary variable that is 1 when an individual takes the extra 500 deductible and  $X$  is a rich set of controls, including demographics (gender, age, having children, living with a partner), financial variables (household gross income in deciles, net worth in quartiles, a dummy for having savings > 2000 EUR, for having a mortgage debt, for having another type of debt), education level and field and professional sector.

We then define

$$d_{\pi_{pop}}(X_{it}) = \sum_{\delta} d(X_{it}, \delta) dF_{\delta},$$

which gives us the predicted deductible take-up rate for each observed  $X_{it}$  combination but as if there were a population of individuals with that  $X_{it}$  with the same health distribution as the overall population. In the same way, we predict the choice quality for individuals with demographic vector  $X_{it}$ , as captured by the probability to choose the contract that minimizes expected expenditures,  $d_{\pi_{pop}}^*(X_{it})$ , and the corresponding average financial loss  $\Delta w_{\pi_{pop}}^{*,\sigma}(X_{it})$ . That is,<sup>14</sup>

$$d_{\pi_{pop}}^{*,\sigma}(X_i) = \sum_{\delta} \{1[\pi_{\delta} \leq .5] [1 - d(X_{it}, \delta)] + 1[\pi_{\delta} > .5] d(X_{it}, \delta)\} dF_{\delta},$$

$$\Delta w_{\pi_{pop}}^*(X_{it}) = \sum \{1[\pi_{\delta} \leq .5] d(X_{it}, \delta) [CE_{\pi_{\delta},0}^{\sigma} - CE_{\pi_{\delta},500}^{\sigma}] + 1[\pi_{\delta} > .5] [1 - d(X_{it}, \delta)] [CE_{\pi_{\delta},500}^{\sigma} - CE_{\pi_{\delta},0}^{\sigma}]\} dF_{\delta}.$$

The choice quality varies through the deductible choice predicted by the set of demographics  $X_i$  for different health risks, but again reflects the population distribution of health risks.

Figure A.12 compares the distribution of predicted deductible choice, with and without the effect of healthcare

<sup>14</sup>Note that we use the average predicted risk for the different health deciles to calculate the certainty equivalents and to determine whether one should take up the deductible or not.

cost risk. These are denoted in previous equations as  $d(X_{it}, \pi_{it})$  and  $d_{\pi_{pop}}(X_{it})$  respectively. As shown before, health has a meaningful impact on deductible choice, but there is substantial heterogeneity in likelihood of choosing a deductible just as a function of  $X_{it}$ , netting out health effects. While losses range up to 200 EUR when factoring health risk into choices, when assuming the population distribution of health for a given  $X_i$  the expected loss range only between 50 and 80 as a function of  $X_i$ .