The European Journal of Health Economics
https://doi.org/10.1007/s10198-021-01363-4

ORIGINAL PAPER

Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in France

Alexandre Vimont[1,2] · **Henri Leleu**[1] · **Isabelle Durand-Zaleski**[2]

Received: 7 February 2021 / Accepted: 29 July 2021

Abstract

Background Innovative provider payment methods that avoid adverse selection and reward performance require accurate prediction of healthcare costs based on individual risk adjustment. Our objective was to compare the performances of a simple neural network (NN) and random forest (RF) to a generalized linear model (GLM) for the prediction of medical cost at the individual level.

Methods A 1/97 representative sample of the French National Health Data Information System was used. Predictors selected were: demographic information; pre-existing conditions, Charlson comorbidity index; healthcare service use and costs. Predictive performances of each model were compared through individual-level (adjusted R-squared (adj-R^2), mean absolute error (MAE) and hit ratio (HiR)), and distribution-level metrics on different sets of covariates in the general population and by pre-existing morbid condition, using a quasi-Monte Carlo design.

Results We included 510,182 subjects alive on 31st December, 2015. Mean annual costs were 1894€ (standard deviation 9326€) (median 393€, IQ range 95€; 1480€), including zero-claim subjects. All models performed similarly after adjustment on demographics. RF model had better performances on other sets of covariates (pre-existing conditions, resource counts and past year costs). On full model, RF reached an adj-R^2 of 47.5%, a MAE of 1338€ and a HiR of 67%, while GLM and NN had an adj-R^2 of 34.7% and 31.6%, a MAE of 1635€ and 1660€, and a HiR of 58% and 55 M, respectively. RF model outperformed GLM and NN for most conditions and for high-cost subjects.

Conclusions RF should be preferred when the objective is to best predict medical costs. When the objective is to understand the contribution of predictors, GLM was well suited with demographics, conditions and base year cost.

Keywords Predictive analytics · Machine learning · Cost containment · Healthcare management · Healthcare costs · Random forest · Neural network

JEL Classification I11 · I13 · I15

Introduction

In countries with non-competitive health insurance, policy makers currently attempt to shift fee-for-services provider payment towards bundled or episode-based payments that are expected to contain costs while ensuring quality of care and performance [1]. Countries with competitive health plans tend to regulate benefits and service selection, and prohibit exclusion of people based on costs or pre-existing conditions [2].

Information about patients' pre-existing conditions and the type of medical services used is required to adjust for risk and avoid adverse selection of beneficiaries in competitive environments [3–5] or to ensure equity of resource allocation in non-competitive environments. Episode-based payments, bundled payments or pay-for-performance [6, 7], are often used in combination to improve the allocation of resources without risking exclusion of the most vulnerable individuals [4]. Payments can be disease-specific, episode-specific or

✉ Alexandre Vimont
alexandre.vimont@ph-expertise.com

[1] Public Health Expertise (PHE), Paris, France

[2] Assistance Publique Hôpitaux de Paris, URC-ECO, CRESS-UMR1153, Paris, France

Published online: 09 August 2021

adjusted to the characteristics of patients, as in risk adjusted capitation payment [8].

However, some beneficiaries remain unprofitable for health plans, because risk adjustment is imperfect [9]. In response, health plans may have an incentive to avoid enrolling or treating certain beneficiaries by reducing the provision of the healthcare services (service selection) that are most attractive to these people [10]. Another consequence is that health plans may have an incentive to reduce the minimum services covered on healthy enrolees, because it is used more heavily by high-risk people.

Reducing such risk selection can be managed by ensuring that a risk adjustment model can accurately predict healthcare expenditures at the individual level. Profile selection has been shown to occur more prevalently in high cost individuals, at the tail of the distribution of healthcare cost [9]. With a perfect risk adjustment, competing health plans would compete by offering services that maximize welfare. More realistically, with imperfect information, they distort the quantity of services in an inefficient way that is profitable for them [10]. To this end, they provide specific services to attract profitable profiles of beneficiaries and limit service for unprofitable profiles. To overcome this distortion, the concept of optimal risk adjustment was implemented by payers by compensating health plans with additional cost by high-risk type of patient [11].

Although traditional risk adjustment models have reduced risk selection and exclusion of some beneficiaries, the variance of medical costs still increases with the expected mean, and important discrepancies remain between costs and payments among beneficiaries with a "high risk" profile. Based on their complementary information from individuals, health plans of the Medicare Advantage Program (Medicare's risk adjustment of capitation payments to private health plans), were shown to enrol individuals who are low cost conditional on their health risk score [9].

The prediction of costs usually uses regression models [10]. In the USA, the standard model developed for Medicare and Medicaid, the CMS-HCC-type model (Centers for Medicare & Medicaid Services-Hierarchical Condition Categories), relies on linear regression and predefined hierarchical condition categories [12]. Risk adjustment models are designed to identify information that predicts the costs of individuals, but are commonly evaluated in terms of their ability to predict at both the individual and group level [13].

In France, the healthcare environment is a mix of public and private settings, with hospitals being mostly public and outpatient cares being mostly private. The national health insurance is the main payer, with private insurance accounting for the complementary part. Hospital revenue comes from an activity-based payment, the French-equivalent to the DRG system, and outpatient healthcare are paid trough tariffs fixed by the national health insurance. In that context, predictive models used conjointly with medical claim database may be used by the national health insurance, to improve care, improve resource allocation and optimize healthcare costs.

Predicting healthcare costs at the individual level is challenging due to the complexity of the underlying factors driving costs, in which correlation between variables causes multicollinearity. The distribution of costs increases the complexity of prediction, with its typical spike at zero and its skewness with a heavy right-hand tail [12]. The use of generalized linear models (GLM) is limited to linear relationships with a limited number of explanatory variables; innovative models that can handle large amounts of data with nonlinear relationships are being proposed [14].

Extended versions of GLM have been developed, which included flexible parametric models (generalized gamma and beta models), and semi-parametric models (extended estimating equations, finite mixture models and conditional density models), and were applied to healthcare costs prediction, but no single model dominated on all criteria (bias, accuracy and goodness-of-fit) over all possible sample sizes [15, 16].

Knowing that strong heterogeneity in healthcare cost exists between individuals of the same risk class or morbidity, and leads to poor model performance, we aimed at using innovative models designed to identify interactions between covariates. Random forest (RF) or neural network (NN) models that use machine learning have been recently applied to the individual prediction of healthcare costs with better results than regression models [13, 17, 18]. Interest in these models stems from the fact that they can handle a nonlinear relationship between costs and risk factors, they use supervised learning algorithms and they can manage a considerable volume of observations and covariates. Neural network and random forest have been shown to be remarkably efficient in many fields, but their use for the prediction of healthcare costs needs to be further explored [14, 19]. Recently, regression trees demonstrated their ability to detect complex interaction effects [20] and specially to improve risk adjustment regression for healthcare cost prediction [21]. Random forests are an extension of regression tree, and are more stable due to the use of re-sampling.

Our objective was to compare the performances of two simple models that use machine learning, a RF model and an NN model, to classic linear regression to predict healthcare costs of individuals.

Data and method

Data

We used a representative sample of the French National Health Data Information System (SNDS) which is gathering

all medical claims from the 70 million insured individuals and is centralized in one database. It also contains demographic information and diagnostic information and was described elsewhere [22]. For general population studies, a representative sample in terms of age, gender and location, at 1/97 of this database is available to researchers.

The SNDS database contains several data sets linked together via a unique patient ID (social security number). These data sets include demographic and administrative patient data (e.g., age, sex, and place of residence); healthcare visits and procedures reimbursed (e.g., medicines, medical procedures, medical devices, lab tests); and date of death. Data from hospitals and other healthcare facilities are also available which includes inpatient data, such as medical information, diagnosis (based on International Classification of Diseases, 10th Revision, Clinical Modification [ICD-10-CM] codes), medical procedures, imaging, external visits, external procedures performed, expensive medicines, and implantable devices.

Study population

All individuals with or without claims during 2015 and who were alive on December 31, 2015, were included. Exclusion criteria were based on health status for the year 2016. Pregnancy-related care was excluded, because associated costs were considered predictable and highly correlated with age and sex. Subjects with psychiatric disorders were excluded, because the database did not include hospital stays in public psychiatric wards which represent around 60% of costs in this population [23]. Individuals living in extra-continental France during the year 2016 were excluded, because they have a different tariff policy, with higher tariffs than continental France due to a higher cost of living which would systematically bias the results towards higher costs for a particular subgroup of the population. Subjects living abroad were also excluded.

Outcome and costs measurement

The models aimed at predicting total individual healthcare costs including inpatient and outpatient costs directly attributable to care provided in 2016. Costs were not limited to the part reimbursed by the National Health Insurance, but included the entire bill, including complementary health insurance and out-of-pocket costs. Costs were aggregated at the individual level for the base year (2015) and the predicted year (2016), and summarized into the following service types [24]: hospital stays, pharmacy, general practitioner (GP) and specialist visits, supportive care, biology, medical devices, transportation and allowance. Psychiatric stays in public hospitals and rehabilitation care which were not available in the database were not included in the study.

Hospital stay costs, both public and private, included Disease-Related Group (DRG) tariffs, additional lump-sums related to the DRG and cost of drugs which were not covered by the DRG tariff. Pharmacy costs included the price of the drug delivered in a retail pharmacy or at the hospital. Specialist and GP visits costs covered consultations, medical acts and other applicable fees. Supportive care included nurse care, physiotherapy, and dental care. Transportation costs for care providers were included in the fee. Transportation included only travel expenses for patients (not their dependents). Allowances covered compensation for sick leave and disability.

Predictors

Six different sets of predictors from the base year (2015) were considered:

- Demographic information: sex, age, means-tested state-sponsored health insurance, and deprivation index [25];
- Pre-existing conditions, Charlson comorbidity index [26] and full compensation for long-term illness;
- Counts of resource use for base year;
- Pharmacy delivery for base year;
- Total healthcare costs for base year, continuous and by cost ranges;
- Costs by service type for base year;

Fifty-six pre-existing morbid conditions were identified by medical algorithms combining diagnoses, procedures and pharmacy [24]. For chronic diseases or a continuous therapy of a minimum 6-month, healthcare costs are fully covered by the national health insurance and full coverage is recorded in the claims database with the mention 'long term illness'. The presence of a long-term illness and its duration were used as predictors with the Charlson comorbidity index previously validated in this database [26]. The presence of a condition was identified when 1/ individuals had ongoing 'long-term illness' allowance related to the condition, 2/ individuals were hospitalized within the past 2 years (2015 or 2014) with a primary or secondary diagnosis related to the condition, or 3/ individuals having several deliveries of condition-specific treatments.

All predictors may not be relevant as risk adjusters, because they can potentially be subject to ex post moral hazard. Health plans or beneficiaries have incentives to manipulate use and affect payment. Beyond that, we aimed at showing in which proportion past resources use and past costs improve model performance and discuss their implication. Resource use included yearly counts of each service type. The pharmacy yearly count was measured by the yearly number of deliveries for the top 20 of the most delivered 2nd level class of the drug classification system (Anatomical

Therapeutic Chemical classification, ATC, e.g., J01 Antimicrobials for systemic use) which could be either pharmacological or therapeutic groups.

Models

Performances of random forest and neural networks were compared to standard regression on their ability to predict the level of healthcare costs for individuals. Analysis and interpretation of the healthcare demand was not an objective in this study, so a two-part model was not considered. Raw-scale outcome rather than the log-transformed scale was modelled after examining heteroscedasticity according to the framework proposed by Manning and Mullahy [27]. Subjects with zero claim for the predicted year (2016) were imputed a GP visit cost not to exclude these subjects and because 60% of them had at least one GP visit within the three previous years [28].

The NN had a multilayer perceptron architecture, as described by Bishop, with several hidden layers, and several neurons in each layers [29]. The NN was composed of successive sets of neurons (layers) with or without connections between them. Parameters and weights were estimated by minimizing the loss function using the SAS procedure HPNEURAL. Because the number of layers and neurons in each layer is dependent on the data set and the NN architecture, there is no consensus for a method to determine these numbers. Consequently, they were fixed manually and step-by-step, by minimizing the average absolute error. An architecture of 2 layers was chosen, with k neurons (being the numbers of covariates, different in each set) for the first layer, and $k/4$ for the second layer.

The RF model relied on binary prediction trees which divide the input space into hyper-rectangular regions, making a constant prediction within each region. All individuals belonging to the same region are assessed as equal risks. The splitting of the input space can be represented by a decision tree with binary splits. Variable selection is mechanically carried out, because each split uses only a single covariate, and approximates variable interaction through the hierarchical structure of the node splits. RF has been developed based on this methodology, corresponding to a collection of bootstrap samples that aim to reduce the variance of the prediction [30]. For this study, RF used the SAS procedure HPFOREST, parametrized with a number of 50 maximum trees and a number of 50 maximum nodes.

Both models were compared to a standard generalized linear model (GLM). The GLM framework uses the underlying least squares approach [29]. Pairwise interaction terms from each set of covariates were introduced in the model for variable selection carried out by stepwise regression. Interaction terms were assumed theoretically to be modelled if relevant by NN and RF. GLM used a log link function and a gamma response probability distribution to obtain unbiased estimates. Such models have been previously used to predict healthcare costs in this database [31].

Quasi-Monte Carlo design

A quasi-Monte Carlo design was used, where the total population was randomly divided into two equally sized subpopulations: an estimation set and a validation set. From within the estimation set, 100 samples were randomly draw with replacement. The models were estimated on the samples and performance was evaluated on the full validation set. Using the estimated models, healthcare costs for the second year was calculated in the validation set based on their predictors. This process was repeated 100 times and the average for each performance metric was calculated over these 100 iterations.

The quasi-Monte Carlo design enables valid comparison of prediction performance across estimators [15]. The design uses an out-of-sample prediction technique and ensures the results are not driven either by overfitting or traditional Monte Carlo assumptions.

Comparative performance

Predictive performances were evaluated through different individual-level metrics presented globally for each set of predictors and for different subpopulation. Individual-level prediction accuracy in the validation set was evaluated by the adjusted R-squared (adj-R^2), Mean Absolute Error (MAE) and Hit Ratio (HiR). Adj-R^2 represents the part of the total variance explained by the model to ascertain how a model fits to the outcome. In our study, a model with an adj-R^2 below 0.3 will poorly explain the data, whereas a model with an adj-R^2 above 0.5 will be acceptable. However, adj-R^2 may not be appropriate for machine learning models because of the underlying linear assumption that total variance is the sum of regression variance plus error variance which may not be verified. For this reason, the average of the absolute errors between the prediction and the true value (MAE) was also computed and was considered acceptable when inferior to the average total cost of the overall sample. HiR was calculated for each model and cost ranges were based on the following four cost thresholds were used: 100€, 1500€, 5000€, and 15,000€, corresponding to approximately the 25th, 75th, 95th, and 99th percentiles of the healthcare cost distribution. The HiR measures the percentage of individuals for whom a model predicts the correct cost range. A HiR over 50% means that the model predicts the correct cost range once out of two times and was considered acceptable.

These metrics measure prediction accuracy at the aggregate level and do not inform of inaccuracy at the tail distributions, where risk selection occurs most prevalently [9]. To

evaluate prediction accuracy at the individual level, confusion matrices were estimated for each model adjusted on the set D of covariates (excluding past total cost and past cost by services). Each row represents the instances in a predicted class and each column represents the instances in an actual class, so that all correct predictions are located in the diagonal. Each cell is the probability of a correct range classification. The following four cost thresholds were used: 100€, 1500€, 5000€, and 15,000€, corresponding to approximately the 25th, 75th, 95th, and 99th percentiles of the observed healthcare cost distribution. A graphic was also constructed to visualize prediction accuracy over the observed and the expected distributions. Observed versus predicted costs were averaged to the nearest 1000€ with their 95% confidence interval.

Individual-level prediction accuracy in the validation set was also evaluated through the probability of residuals between observed and predicted cost being above 10%, 50% and 100% of the observed cost. A probability of residuals being above 100% estimated to 0.5 means that error of prediction of 100% or more was expected for 50% of patients. These metrics were presented graphically by cost ranges of 2000€ and were called distribution-specific individual-level metrics.

Interpretation of effect size for covariates relied on Chi-square statistics for GLM, and the reduction of the absolute error for RF from the validation set. Contrary to these models, NN do not provide straightforward interpretation of effect size and would require complex analysis of hidden layers' weights. As no consensus on a particular method has been reached, effect size was not explored for NN.

Statistical analysis

The population was described for demographics, pre-existing conditions in 2015 and cost by service type for 2016. To evaluate the importance of past total costs, costs shift between 2015 and 2016 were described. The main analysis compared the predictive performances of models in the general population for six set of predictors as described above. Sets of predictors were introduced one by one to evaluate their predictive power. First was the set of demographics (set A), then was added the set of pre-existing conditions (set B), the set of resource counts (set C), the set of drugs counts (set D), the base year cost (set E) and the set of costs by service type (set F). Results were computed for the general population and for patients with identified diseases to evaluate how models performed in specific sub-populations. Diseases included diabetes (type I and type II), cancer (occurring within the five previous years), cardio neurovascular disease, chronic obstructive pulmonary disease (COPD) inflammatory or rare disease, end-stage renal disease and neurodegenerative disease.

In a secondary analysis, models were fitted on the same conditions only to evaluate if the models performed less accurately using a general population model than condition-specific models. Predictions by disease were compared to predictions from the general population model. All analyses used SAS version 9.4 on deidentified data with the approval of the French data protection authority (*Ref: MMS/MFI/AR1811775)* (SAS Institute Inc., Cary, NC, USA).

Results

Description of the population

We excluded 12.2% of the total study population with a psychiatric condition or a regular psychotropic treatment, 6.3% living in non-continental France and 1.1% for pregnancy-related care. We included 510,182 subjects alive on 31st December, 2015 (Figure S1). Mean age was 42 years, 48% were women and 20% had a long-term illness with full coverage. Of these subjects, 23% had at least one pre-existing condition in 2015, the most prevalent were vascular risk (10.3%), cardio-neurovascular disease (4.6%) and chronic respiratory disease (4.2%) (Table 1). The Charlson index was superior or equal to 3 for 5.1% of the population.

Description of costs

Mean annual costs were 1769€ (standard deviation 9158€) (median 375€, IQ range 89€; 1415€) for base year, including zero-claim subjects. Hospital stays, pharmacy and medical visits accounted for 70% of the 2016 costs (Table 1). Costs were highly variable between individuals, especially for pharmacy and hospital stays with standard deviations being two- to sixfold the mean. Figure 1 presents the variation in total costs between 2015 and 2016, with 60% of subjects remaining within the same cost range, 20% shifting upwards and 20% downwards. Zero-claim subjects represented 16% in 2016, and 65% of them also had null consumption the previous year. High-cost subjects (≥ 15,000€) who represented 2% of the sample accounted for 45% of overall costs of the included population. Total individual cost for the year 2016 was highly correlated to cost of the previous year 2015 (Pearson correlation $\rho = 0.60$, $p < 0.0001$).

Predictive performance

Individual-level metrics for predictive performances of models in the general population are presented in Table 2. Performances were similar when covariates included demographic information only (Set A) with an adj-R^2 of 2%, a MAE of 2380€ and a higher HiR for the RF model (42.5%). Adding information on pre-existing conditions (Set B) increased the

Table 1 Description of the population of analysis for base year (2015)

Variable	Included population
Patients (*n*)	510,182
Sex (female, %)	50.5%
Age (year, mean)	41.9 (24)
Medical	
Long term illness (LTI) (%, y/n)	19.9%
Years with LTI (year, mean)	1.05 (13)
^≤5	68.5%
^5–10	8.5%
^>10	23.0%
Charlson index (%)	
0	77.2%
1–2	17.7%
≥3	5.1%
Conditions	
Treatment for vascular risk	10.3%
Cardio neurovascular	4.6%
Chronic obstructive pulmonary disease	4.2%
Diabetes	4.1%
Cancers	3.1%
Inflammatory or rare	1.4%
Neurologic or degenerative	1.1%
Kidney disease terminal phase	0.1%
Liver or pancreas	0.5%
other disease	1.6%
Costs by service type (€, mean, sd)	
Hospital stays	876 (1432)
Pharmacy	311 (1845)
GP visit	81 (107)
Specialist visit	125 (428)
Care support†	239 (458)
Biology	52 (149)
Medical device	81 (431)
Transport	46 (512)
Allowance	253 (726)
Total costs	1769 (9158)
Repartition of total costs	
0	15.6%
<100€	11.1%
<1500€	52.5%
<5000€	13.7%
<15,000€	5.1%
≥15,000€	2.1%

†Nurse care, physiotherapist care, dental care, non-pharmaceutical supply

adj-R^2 for all models and explained around 10% of total variance for GLM and NN models, and over 15% for RF. RF had a better accuracy with a lower MAE (1928€) and a higher HiR (55.3%).

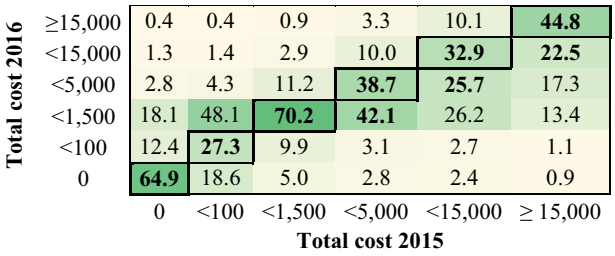

Fig. 1 Total costs shift from 2015 to 2016 (% of column)

Introduction of resource counts (Set C) improved the adj-R^2 for all models, but more importantly for RF. In both models, the total cost for base year explained around 10% of the total variance, while it explained 20% in the RF model. GLM and NN models performed similarly on Set C, with an adj-R^2 ranging from 22 to 25% and a MAE estimated to 1890€. The HiR was higher for RF at 60% versus 50% for both GLM and NN.

Adding drug counts (set E) had very little improvement. However, adding total cost of 2015 improved considerably the individual-level accuracy of all models. RF had adj-R^2 estimated to 46%, the MAE to 1366€ t and the HiR reaching a maximum of 66%. For GLM and NN, adj-R^2 increased by 8% and MAE decreased by 200€. Adding 2015 costs of service type (set F) improve by little performance of models.

Figure 2 presents a functional evaluation of performance with prediction accuracy for the distribution of healthcare costs. Observed versus predicted costs for 2016 adjusted on the set D of covariates, rounded to the nearest 1000€ are presented for each model with their 95% confidence interval. In subjects with total costs < 5000€, all models performed similarly with slight overprediction for GLM and NN models. In patients over 5000€, RF fitted best and GLM had the poorest accuracy with narrower confidence interval.

Prediction accuracy for individual-level healthcare costs is presented with confusion matrices in Fig. 3. Metrics show that all models overestimated costs below 100€. For total costs between 1500€ and 5000€, accuracy was similar between models. Over 5000€, predictions were 10–20% better with RF than GLM or NN models. For example, GLM and NN misclassified 33% and 28% of subjects over 15,000€ (Top 1%) to the range 1500€–5000€ (75th–95th percentile), while RF misclassified only 7% of them to the same range. Predictions at the tail of the distribution (≥15,000€, Top 1%) were 20% more accurate for RF (50.6%) than for GLM (30.4%) or NN (35.5%). Figures 2 and 3 demonstrate that RF had much better performance accuracy on the tail distribution.

Performance accuracy was also evaluated with distribution-specific individual-level metrics, presented graphically in Fig. 4. Probabilities of residuals between observed and predicted cost being above 10%, 50% and 100%,

Table 2 Evaluation of performance with individual-level accuracy of prediction of healthcare costs for the year 2016

Covariates	Set	DF	Fit	GLM	Neural network	Random forest
Demographic	**[A]**	**5**	**adj-R^2**	**0.016**	**0.015**	**0.020**
			MAE	2256	2310	2380
			HiR	37.6	31.5	42.5
Demo + conditions	**[B]**	**17**	**adj-R^2**	**0.114**	**0.105**	**0.186**
			MAE	2103	2056	1928
			HiR	41.2	40.1	55.3
Demo + conditions + resource counts	**[C]**	**27**	**adj-R^2**	**0.250**	**0.212**	**0.381**
			MAE	1887	1895	1547
			HiR	50.7	49.5	60.4
Demo + conditions + resource counts + drug counts	**[D]**	**48**	**adj-R^2**	**0.264**	**0.226**	**0.398**
			MAE	1862	1871	1495
			HiR	52.3	50.8	61.7
Demo + conditions + resource counts + drug counts + total cost 2015	**[E]**	**49**	**adj-R^2**	**0.336**	**0.308**	**0.462**
			MAE	1670	1689	1366
			HiR	57.1	54.6	66.3
Demo + conditions + resource counts + drug counts + total cost 2015 + costs by service type	**[F]**	**59**	**adj-R^2**	**0.347**	**0.316**	**0.475**
			MAE	1635	1660	1338
			HiR	58.4	55.2	67.5

Models were estimated separately for each set of covariates derived from 2015 and their associated performances are presented in this table. D.F stands for degrees of freedom for the full model (before stepwise selection of covariates)

respectively, were computed for observed cost intervals of 2000€. Below 6000€, RF model had better prediction whatever threshold is allowed, while between 10,000€ and 20,000€, GLM showed slight better accuracy.

Influence of covariates

Influence of the covariates was compared between RF and GLM for the set D of covariates and is presented in Supplementary material. Cumulative length of stay and number of stays contributed the most in the predictions with the RF model, along pharmacy deliveries, LTI and age. For the GLM, cumulative length of stay alone was the best predictor as well, followed by its interaction terms with universal allowance, long-term illness, gender, deprivation index and chronic end-stage renal disease. However, age and number of pharmacy deliveries had higher predictive importance in RF than in GLM.

In the GLM, pre-existing conditions including end-stage renal failure, cardio neurovascular diseases and diabetes were not found to be statistically significant alone in predicting costs, but interaction with cumulative length of stay was. Because variability of cumulative length of stay was high in chronic conditions, interaction terms were necessary so that the weight of the disease in the model was different regarding the number of hospital days during the base year. For example, for end-stage renal disease, the model applied a greater weight for patients undergoing haemodialysis or kidney transplant (greater cumulative length of stay) than for patients with post-transplant follow-up.

Performances in specific pre-existing conditions

Condition-specific models were fitted separately, on subjects with specific pre-existing conditions only (Table 3), to explore whether they performed better than the general population model in these individuals.

For the GLM, condition-specific models performed slightly better than the general population model for inflammatory and rare diseases, end-stage renal disease and neurodegenerative diseases. For NN and RF, condition-specific models had no benefit compared to the general population model, except for end-stage renal disease.

RF performed better than GLM and NN in diabetes, cardio neurovascular diseases, COPD and long-term illness, whether with the general population model or with condition-specific models.

Pre-existing conditions, such as inflammatory and rare diseases, end-stage renal disease and neurodegenerative diseases with high correlation between base year and current costs, which made the prediction mostly linear, were the conditions for which the three models had the best predictive performances.

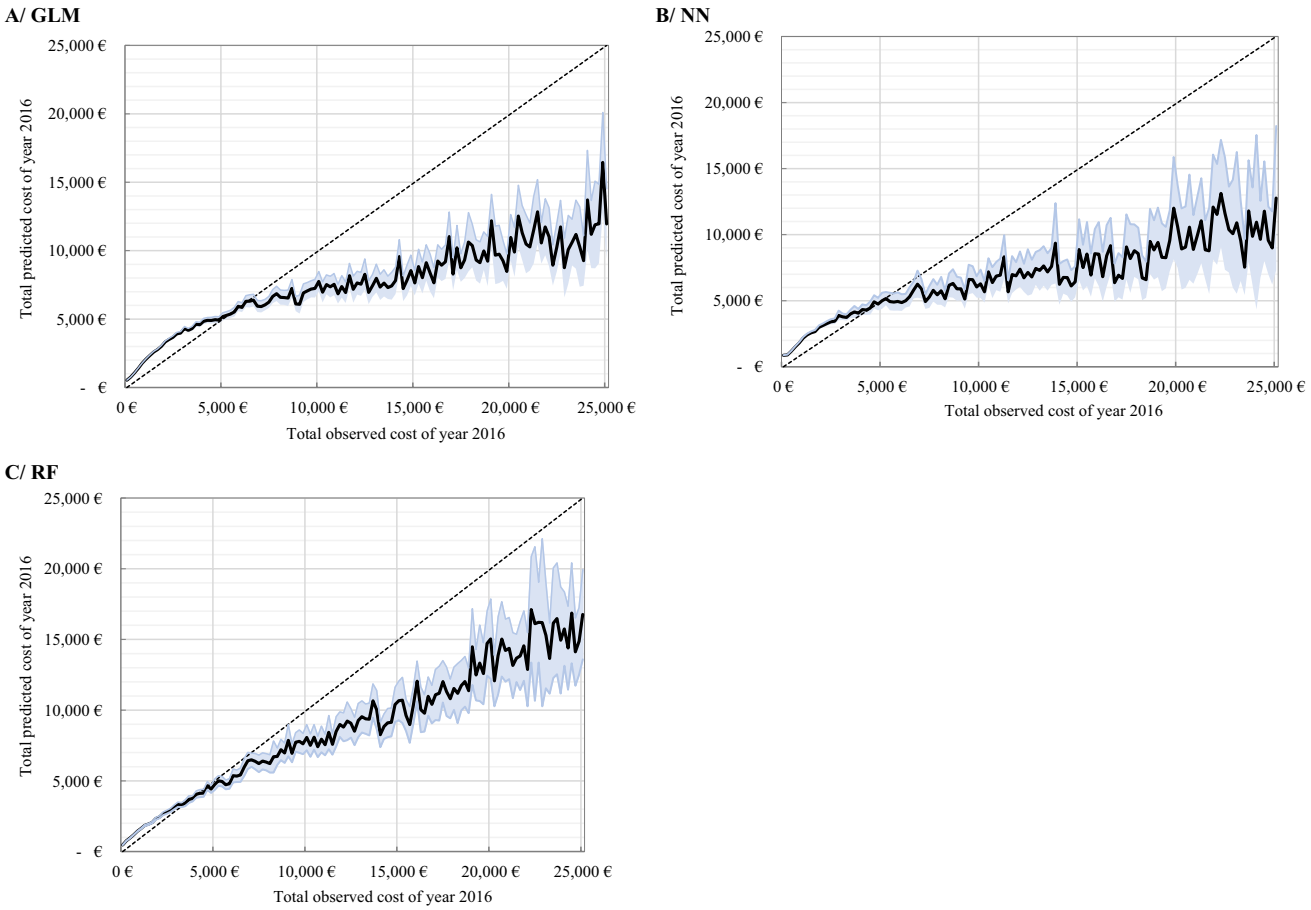

Fig. 2 Evaluation of performance (Set D of covariates) with prediction accuracy for the distribution of healthcare costs (observed versus predicted cost for 2016, rounded to the nearest 1000€, and 95% confidence interval, for GLM (1), NN (2) and RF (3) models)

Fig. 3 Evaluation of performance (Set D of covariates) with prediction accuracy for tail distribution of healthcare costs (confusion matrices for GLM (1), NN (2) and RF (3) models (set D of covariates)

1/ GLM

Predicted value \ Observed value	0	<100	<1,500	<5,000	<15,000	≥ 15,000
≥15,000	0	0.1	0.2	1.1	5.2	**32.5**
<15,000	0.7	0.6	2.8	18.2	**40.6**	**32.8**
<5,000	13.4	10.6	23.6	**48.3**	**33.6**	26.8
<1,500	76.7	81.9	**70.5**	**30.9**	20.2	7.7
<100	9.2	**6.8**	2.9	1.5	0.4	0.2
0	**0.0**	0.0	0.0	0.0	0.0	0.0

2/ NN

Predicted value \ Observed value	0	<100	<1,500	<5,000	<15,000	≥ 15,000
≥15,000	0.2	0.1	0.3	1.6	7.4	**36.8**
<15,000	0.8	0.9	2.5	13.9	**35.6**	**28.4**
<5,000	10.8	10.7	22.8	**48.5**	**36.1**	26.2
<1,500	49.7	56.6	**59.6**	**32.4**	19.8	7.6
<100	38.5	**31.7**	14.8	3.6	1.1	1
0	**0.0**	0.0	0.0	0.0	0.0	0.0

3/ RF

Predicted value \ Observed value	0	<100	<1,500	<5,000	<15,000	≥ 15,000
≥15,000	0	0.1	0.1	0.9	5.1	**51.4**
<15,000	0.6	0.2	0.5	11.4	**52.3**	**41.1**
<5,000	4.6	6.2	18.6	**53.9**	**36.4**	6.5
<1,500	94.8	93.5	**80.8**	**33.8**	6.2	1.0
<100	0.0	**0.0**	0.0	0.0	0.0	0.0
0	**0.0**	0.0	0.0	0.0	0.0	0.0

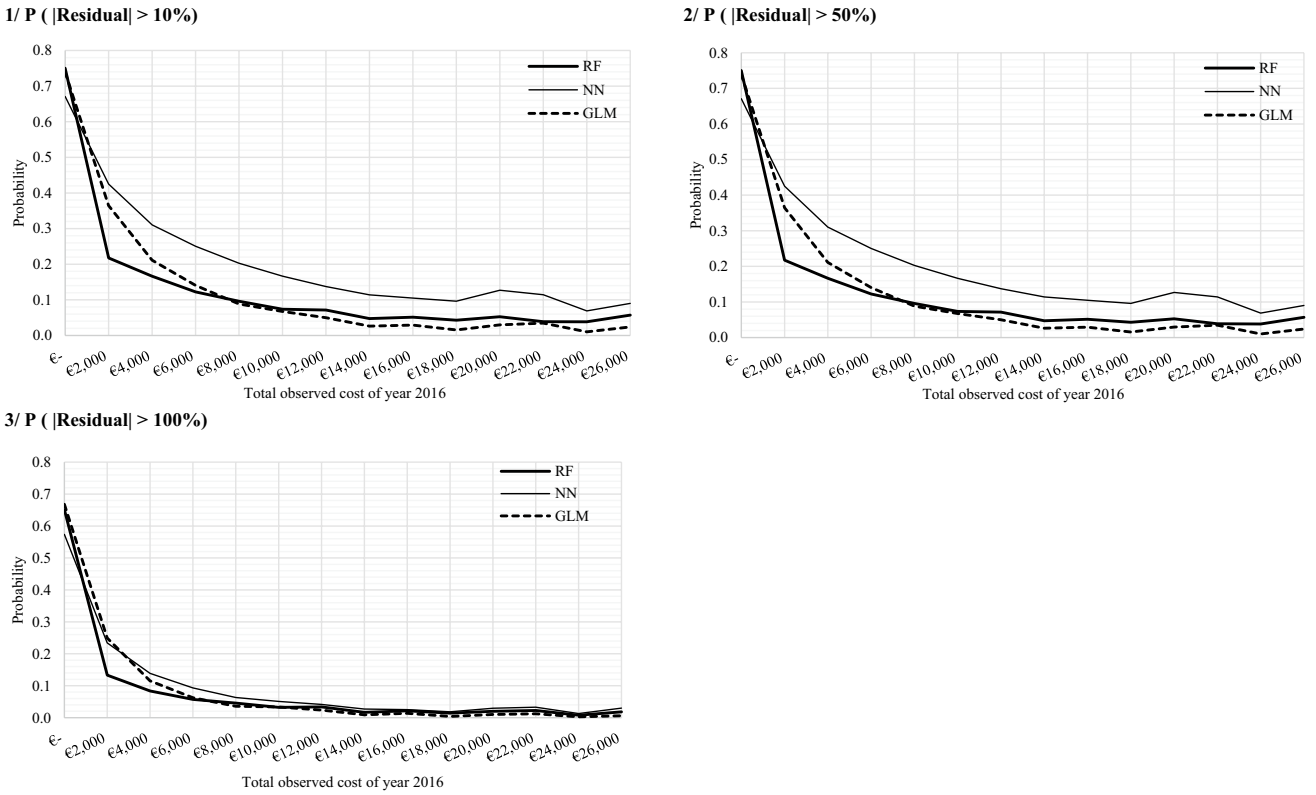

Fig. 4 Evaluation of performance (Set D of covariates) with distribution-specific individual-level prediction of healthcare costs (probability of residuals being above 10% (1), 50% (2) and 100% (3)

Discussion

Findings and related literature

We compared the performance of two supervised learning models against a standard regression model in predicting healthcare costs at the individual level. Main findings were that RF model performed better than other models, particularly on high-cost individuals, and that fitting disease-specific models in subjects with a specific pre-existing condition did not systematically improve the predictive performance.

Predictive accuracy was evaluated through different metrics: individual-level metrics and distribution-level metrics presented globally and by sub-populations. Resource counts and total past costs were the main predictors of total costs, with 60% of subjects remaining in the same cost range from 1 year to the next, followed by hospital stays (number and cumulative length of stay), and the presence of a long-term illness.

The main finding of our study was that RF outperformed NN and GLM for our population. At the individual-level, when using a large number of covariates RF explained 13% more of the total variance than GLM and NN, reaching an adj-R^2 of 47.5%. RF correctly classified 67.5% of subjects, while NN and GLM classified less than 60% of subjects in the exact range of costs. MAE was estimated as 1338€ for RF which was comparable to other studies that used semi-parametric models [15].

Individual-level metrics with confusion matrices and functional graphs showed that RF was the most accurate model in subjects with total costs between 5000€ and 15,000€ and also at the tail of the distribution (> 15,000€, Top 1%), while the performances of GLM and NN largely decreased with growing total costs. The excellent performances of RF lie in its statistical framework that allow identification of high-level interactions in a non-linear form between covariates. In addition, RF builds bootstrap samples of random trees and makes predictions by averaging the individual tree predictions leading to a greater stability of results. The fact that RF splits input spaces makes the model much less sensitive to outliers and extreme values than GLM and NN [32]. RF models for provider payment schemes are easily handled, since they are fully nonparametric, requiring no restrictive assumption, and deal mechanically with nonlinear relationship and high-level interactions among covariates.

Results based on demographics and pre-existing conditions presented an estimated adj-R^2 of approximately 10% for GLM and 20% for RF which is consistent with the existing literature in the general population [16]. Although using

Table 3 Individual-level accuracy for sub-models fitted on sub-populations (set D of covariates)

Sub-populations	Fit	GLM		Neural network		Random forest	
		General population	Disease-specific	General population	Disease-specific	General population	Disease-specific
Diabetic	**adj-R^2**	**0.318**	**0.306**	**0.274**	**0.323**	**0.468**	**0.467**
	MAE	5385	5238	5025	5169	4428	4496
	HiR	35.2	38.6	40.8	37.1	47.1	47.5
Cancer	**adj-R^2**	**0.196**	**0.164**	**0.197**	**0.090**	**0.271**	**0.221**
	MAE	8208	8107	8152	8528	7211	7634
	HiR	21.1	26.2	32.1	28.2	42.8	37.3
Cardio neurovascular	**adj-R^2**	**0.278**	**0.293**	**0.201**	**0.305**	**0.466**	**0.439**
	MAE	6641	6648	6698	6555	5755	5829
	HiR	32.8	32.3	37.5	31.5	46.0	40.2
Chronic obstructive pulmonary disease (COPD)	**adj-R^2**	**0.255**	**0.183**	**0.271**	**0.285**	**0.417**	**0.374**
	MAE	3943	3827	4150	4351	3384	3571
	HiR	43.0	43.3	46.6	35.8	63.3	54.9
Inflammatory/rare	**adj-R^2**	**0.497**	**0.432**	**0.492**	**0.486**	**0.573**	**0.434**
	MAE	6213	5816	5992	6097	5323	6009
	HiR	32.2	39.3	37.8	36.5	48.1	39.9
End-stage renal disease	**adj-R^2**	**0.469**	**0.521**	**0.346**	**0.593**	**0.566**	**0.603**
	MAE	611	439	579	463	595	585
	HiR	27.1	29.1	17.1	28.0	35.4	28.0
Neurodegenerative	**adj-R^2**	**0.362**	**0.379**	**0.445**	**0.451**	**0.415**	**0.351**
	MAE	6617	6030	6053	6072	5926	6244
	HiR	29.5	37.6	38.0	34.6	51.2	42.5
Long-term illness	**adj-R^2**	**0.250**	**0.224**	**0.226**	**0.201**	**0.423**	**0.377**
	MAE	5545	5376	5441	5391	4458	4514
	HiR	22.9	25.8	27.9	28.5	48.6	43.0

Models were estimated separately for each subpopulation adjusted on set D of covariates. Their associated performances are presented in this table and are compared to performances from the general population model in each sub-population

past costs or past resource counts for prospective payment may not fulfil the objectives of maintaining incentives for healthcare cost control [11, 33], including such information should be reconsidered to avoid selection problems. Introduction of past resource counts and past total costs led to a considerable benefit in individual-level metrics in our study. It has been shown that due to important heterogeneity of healthcare costs within groups of patients with the same comorbid conditions, models based on demographics and pre-existing conditions only, have a poor level of accuracy [16]. Our work suggests that including past resource counts or past total costs could help to refine cost predictions or other risk scores generated by risk-adjustment models.

In practice, the use of RF model in an environment, where health plans are competitive may lead to a substantial decrease in profile or service selection. Better prediction accuracy on all cost ranges would reduce discrepancy between fixed premium and actual cost and thus reduce private health plan incentives for adverse selection. Furthermore, the fact that the RF model was more accurate at the tail of distribution may reduce another selection bias as highlighted by Brown et al. [9]. In their evaluation of risk adjustment use for the private Medicare Advantage program, authors showed that health plans had incentives to enrol higher risk scores and “over-priced" individuals with actual costs significantly below the formula’s prediction, resulting in global increase of overpayments by Medicare [9].

In a non-competitive environment, the use of RF model may help in maximizing efficiency of resource use by health providers or regions while taking into account prospective payment equity and individuals’ welfare [11]. Geographic areas may have some differences in terms of population medical conditions and hence healthcare costs. Historical reasons, such as economic activity or unemployment, may explain some of these differences that are not captured by simple risk adjustment models for prospective payment [10].

Another contribution of this study was that fitting disease-specific models in subjects with a specific pre-existing condition did not improve the predictive performance except

for GLM in inflammatory and rare diseases, end-stage renal disease or neurodegenerative diseases. The choice of a general population model versus condition-specific models thus depends on the characteristics of the target populations and the statistical model to be used. Our study confirmed that the GLM struggles to predict values far from the observed mean. With a limited number of covariates and interaction terms, fitting a GLM is relatively fast, but becomes cumbersome with the growing number of covariates. Because of their machine learning properties, RF and NN converge rapidly with high numbers of observations and covariates.

No common measure exists for variable importance comparison between models. However, similarities and differences can be observed by comparing orders of variable importance between GLM and RF, with cumulative length of stay being the most important predictor for both. If GLM remains the best model for understanding the contribution of predictors, since effect size of each predictor can be interpreted incrementally with the outcome, such interpretation is not possible with RF, which leads to categorize RF as a predictive model only, contrary to GLM which can be seen also as an explicative model.

The three models used in this study are reproductible across country and databases, since (1) they have simple architectures and simple specifications (detailed in method section), and (2) they were developed with pre-programmed procedures in SAS software. Running time was similar between GLM and machine learning models, mainly because data were structured and the number of patients was reasonable. Most of health administrative database are structured but contains millions of individuals. If considering a routine use of predictive models over millions of patients, machine leaning models may be more computationally burdensome than GLM because of their non-linear properties.

Other authors have demonstrated the high performances of RF to predict claims costs [19, 34]. Duncan et al. also showed better performance of RF against regression models to predict healthcare costs from an insurance claims database but warned that RF could provide unstable results without any statistical tool for clear interpretation of covariates' effect [13]. Contrary to Duncan et al., our study presented stable results for RF across sub-populations. Duncan et al. also showed that RF successfully predict patients with high-risk profile which is consistent with our results.

Identifying subjects with very low costs or zero claims was not an objective of this study. As a result, subjects with zero claims were imputed a GP visit cost for 2016 in order not to exclude them from GLM log-link modelling [35]. Reasons for being zero-cost are diverse. Individuals can be healthy or not, being inclined to visit a doctor or not, and others can be out of the healthcare system [36]. Results showed that all models overestimated costs below 100€.

Limitations

Exclusion of patients with psychiatric disorders, who represent a non-negligible part of the population, was the main limitation of this study. These disorders are some of the most expensive conditions, with around 15% of national healthcare spending [23] and may limit in that extent the reproducibility of this work in general population. This population will be explored and included in further work. For patients with pre-existing conditions, such as end-stage renal disease and to a large extent neurodegenerative disease, predictive models are nearly useless as expenditures are typically recurrent.

Sample design study are often suffering from potential lack of representativeness, especially when working on general population. Depending on how the sample database is constructed, the number of included patients with rare diseases may not be representative enough and may be inaccurate when models are applied in practice. However, our study showed that performances were similar whether models were estimated from general population or from specific diseases.

Only NN models with simple architecture were applied in this study. More complex models, such as the recurrent neural network (RNN) or the gradient boosting model (GBM) may perform better [17]. The fact that NN did not perform as well as in the literature resulted from the computational limitations restricting the model to a simple structure of multilayer perceptron unable to capture complex and localized relationship between covariates [17].

Traditionally, adj-R^2 has been used to evaluate prediction models assuming Gaussian errors and summation of variances and may not be suitable to compare performance with NN and RF [37]. While R^2 does measure how well a model fits under linear and homoscedasticity assumptions, it is biased when variance of error terms depends on covariates. However, normality of error terms was checked and was validated for every model.

To counteract potential estimation bias for performance accuracy, other metrics were employed at distribution level and individual level. Such metrics previously demonstrated their ability to assess accuracy of prediction [16]. Additionally, the quasi-Monte Carlo design allowed to limit other sources of overfitting in our study [15].

Using health information and expenditures from claims data was also a limitation. Costs were aggregated yearly, losing temporality effect and intensity of consumption, and they were mostly restricted to service type and resource counts. The use of quarterly information over several years may help to overcome long-term costs dependencies of chronic illness, irregular timing of episodes or resource use by accounting for personal history. Additional data information, such as electronic health records, disease severity measures and

social determinants of health may help detecting signals and accounting for confounding interactions, so that models can predict more accurately healthcare trajectories and costs.

We included past annual costs, because they are strong predictors of future costs due in part to supply-induced demand [14]. In non-competitive settings, where the payer also regulates equity and supply, it is not possible to use past costs in the risk adjustment models. Variations in supply may not only arise from the nature of the care provided but also from contextual reasons, such as geographic location, unemployment or organisation of care on a territory, and future predictive models for payment purposes will need to account for these factors.

Conclusion

While risk adjustment using regression is still the standard tool to predict healthcare costs, alternative models based on machine learning showed promising results. Models perform differently depending upon the type of data and the population, thus the type of model to adopt depends on the objective. When the objective is to understand the contribution of predictors, a regression model, such as GLM, should be chosen. When the objective is to best predict healthcare costs, our study supports the use of supervised learning model such as RF or an NN with more complex architecture than a multilayer perceptron. Improving model accuracy may reduce profile and service selection in a competitive environment and ensure equity of resource allocation in others. In a context of massive available healthcare data, which offers much more medical and sociodemographic information than a classic claims database, innovative models with structures adapted to the complexity of healthcare data should be adopted.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10198-021-01363-4.

Acknowledgements We are grateful to Meryl Darlington for helpful reviews and comments.

Funding This work was supported by PHE.

Declarations

Conflict of interest AV and HL are employees of PHE.

Data availability Data and materials are available for review if needed.

Code availability Code is available for review if needed.

References

1. WHO: Health Systems Financing: The Path to Universal Coverage. WHO, Geneva (2010)
2. Ellis, R.P., Martins, B., Zhu, W.: Demand elasticities and service selection incentives among competing private health plans. J. Health Econ. **56**, 352–367 (2017). https://doi.org/10.1016/j.jhealeco.2017.09.006
3. OECD: Fiscal Sustainability of Health Systems; Bridging Health and Finance Perspectives. OECD Publishing, Paris (2015)
4. OECD: Better Ways to Pay for Health Care. OECD Health Policy Studies. OECD Publishing, Paris (2016)
5. Newhart, P.K.I.F.S.: Evaluation of the CMS-HCC Risk Adjustment Model. Technical report, RTI International and the Centers for Medicare & Medicaid Services (2011)
6. Cashin, C.C.Y., Smith, P., Borowitz, M., Thomson, S.: Paying for Performance in Health Care: Implications for Health System Performance and Accountability. McGraw-Hill Education, London (2014)
7. McClellan, M.: Reforming payments to healthcare providers: The key to slowing healthcare cost growth while improving quality? J. Econ. Perspect. **25**(2), 69–92 (2011)
8. Lamers, L.M., Van Vliet, R.C., Van De Ven, W.P.: Risk-adjusted capitation payment systems for health insurance plans in a competitive market. Expert Rev. Pharmacoecon. Outcomes Res. **3**(5), 541–549 (2003). https://doi.org/10.1586/14737167.3.5.541
9. Brown, J., Duggan, M., Kuziemko, I., Woolston, W.: How does risk selection respond to risk adjustment? New Evidence from the medicare advantage program. Am Econ Rev **104**(10), 3335–3364 (2014). https://doi.org/10.1257/aer.104.10.3335
10. Ellis, R.P.: Risk Adjustment in Health Care Markets: Concepts and Applications. Financing Health Care: New Ideas for a Changing Society. Wiley-VCH publishers, Weinheim (2007)
11. Glazer, J., McGuire, T.G.: Optimal quality reporting in markets for health plans. J. Health Econ. **25**(2), 295–310 (2006). https://doi.org/10.1016/j.jhealeco.2005.10.002
12. Jones, A.: Models for Health Care. University of York, Centre for Health Economics, York (2010)
13. Duncan, I., Loginov, M., Ludkovski, M.: Testing alternative regression frameworks for predictive modeling of health care costs. North Am. Actuarial J. **20**(1), 65–87 (2016)
14. Morid, M., Kawamoto, K., Ault, T., Dorius, J., Abdelrahman, S.: Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation. AMIA Annu. Symp. Proc. **2018**(2017), 1312–1321 (2018)
15. Jones, A.M., Lomas, J., Moore, P.T., Rice, N.: A quasi-Monte-Carlo comparison of parametric and semiparametric regression methods for heavy-tailed and non-normal data: an application to healthcare costs. J. R. Stat. Soc. Ser. A Stat. Soc. **179**(4), 951–974 (2016). https://doi.org/10.1111/rssa.12141
16. Park, S., Basu, A.: Alternative evaluation metrics for risk adjustment methods. Health Econ. **27**(6), 984–1010 (2018). https://doi.org/10.1002/hec.3657
17. Yang, C.D., Shenkman, E., Ranka, S.: Machine learning approaches for predicting high cost high need patient expenditures in health care. Biomed. Eng. Online **17**(Suppl 1), 131 (2018). https://doi.org/10.1186/s12938-018-0568-3
18. Rose, S.: A machine learning framework for plan payment risk adjustment. Health Serv. Res. **51**(6), 2358–2374 (2016). https://doi.org/10.1111/1475-6773.12464
19. Sushmita, S., Newman, S., Marquardt, J., Ram, P., Prasad, V., Cock, M.D.: Population cost prediction on public healthcare datasets. In: Proceedings of the 5th International Conference on Digital Health 2015 (2015)

20. van Veen, S., van Kleef, R.C., van de Ven, W., van Vliet, R.: Exploring the predictive power of interaction terms in a sophisticated risk equalization model using regression trees. Health Econ. **27**(2), e1–e12 (2018). https://doi.org/10.1002/hec.3523
21. Buchner, F., Wasem, J., Schillo, S.: Regression trees identify relevant interactions: can this improve the predictive performance of risk adjustment? Health Econ. **26**(1), 74–85 (2017). https://doi.org/10.1002/hec.3277
22. Tuppin, P.R.J., Constantinou, P., Gastaldi-Ménager, C., Rachas, A., de Roquefeuil, L., Maura, G., Caillol, H., Tajahmady, A., Coste, J., Gissot, C., Weill, A., Fagot-Campagna, A.: Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. Rev. Epidemiol. Sante Publique. **65**(Suppl 4), S149–S167 (2017)
23. Cnam: Améliorer la qualité du système de santé et maîtriser les dépenses - Propositions de l'Assurance Maladie pour 2020. (2019).
24. Cnam: Méthode de repérage des pathologies et d'affectation des dépenses aux pathologies. http://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/etudes-en-sante-publique/cartographie-des-pathologies-etdes-depenses/methodologie.php. (2015)
25. Pornet, C., Delpierre, C., Dejardin, O., Grosclaude, P., Launay, L., Guittet, L., Lang, T., Launoy, G.: Construction of an adaptable European transnational ecological deprivation index: the French version. J. Epidemiol. Community Health **66**(11), 982–989 (2012). https://doi.org/10.1136/jech-2011-200311
26. Charlson, M.E., Charlson, R.E., Peterson, J.C., Marinopoulos, S.S., Briggs, W.M., Hollenberg, J.P.: The Charlson comorbidity index is adapted to predict costs of chronic disease in primary care patients. J. Clin. Epidemiol. **61**(12), 1234–1240 (2008). https://doi.org/10.1016/j.jclinepi.2008.01.006
27. Manning, W.G., Mullahy, J.: Estimating log models: to transform or not to transform? J. Health Econ. **20**(4), 461–494 (2001). https://doi.org/10.1016/s0167-6296(01)00086-8
28. Mihaylova, B., Briggs, A., O'Hagan, A., Thompson, S.G.: Review of statistical methods for analysing healthcare resources and costs. Health Econ. **20**(8), 897–916 (2011). https://doi.org/10.1002/hec.1653
29. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)
30. Breiman: Setting up, using, and understanding random forests. (2003)
31. de Lagasnerie, G., Aguadé, A.S., Denis, P., Fagot-Campagna, A., Gastaldi-Menager, C.: The economic burden of diabetes to French national health insurance: a new cost-of-illness method based on a combined medicalized and incremental approach. Eur. J. Health Econ. **19**(2), 189–201 (2018). https://doi.org/10.1007/s10198-017-0873-y
32. Riddle, D.L., Kong, X., Jiranek, W.A.: Two-year incidence and predictors of future knee arthroplasty in persons with symptomatic knee osteoarthritis: preliminary analysis of longitudinal data from the osteoarthritis initiative. Knee **16**(6), 494–500 (2009). https://doi.org/10.1016/j.knee.2009.04.002
33. Frank, R.G., Glazer, J., McGuire, T.G.: Measuring adverse selection in managed health care. J. Health Econ. **19**(6), 829–854 (2000). https://doi.org/10.1016/s0167-6296(00)00059-x
34. Kim, Y.J., Park, H.: Improving prediction of high-cost health care users with medical check-up data. Big Data **7**(3), 163–175 (2019). https://doi.org/10.1089/big.2018.0096
35. Kleef, M.V.: Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets: Theory and Practice. Academic Press Inc, New York (2018)
36. OECD: Income-Related Inequality in the Use of Medical Care in 21 OECD Countries—Health Equity Research Group Members. OECD HEALTH WORKING PAPERS (2004).
37. Cameron, A.C., Windmeijer, F.A.: An R-squared measure of goodness of fit for some common nonlinear regression models. J. Econom. **77**, 329 (1997)