Predicting Swiss Healthcare Costs using Machine Learning

Dino Collalti* Jonas Meier**

July 9, 2021

Abstract

Healthcare costs account for a substantial and steadily growing share of GDP. However, predicting the growth of healthcare costs amounts to be a difficult task due to complex time trends and dependencies on economic measures such as wages and spatial differences. Using physician data, this paper takes a machine learning approach towards predicting aggregate healthcare expenditures. The contribution is twofold. First, machine learning techniques are shown to yield improved forecasts. Compared to realized costs, the benchmark Random Forest model increases prediction accuracy by 30%. Second, using simplistic data avoids having data and causality issues. In a further exercise, we estimate the effect of the 2018 TARMED revision on aggregate cost via Causal Forest. We discover that the reform resulted in aggregate savings of 0.36 billion CHF, considerably less than the proposed savings target of 0.47 billion CHF.

Acknowledgements: We would like to thank SASIS AG for providing the data for the present analysis.

*Department of Economics, University of Bern, Schanzeneckstrasse 1, 3001 Bern, Switzerland. Mailto: dino.collalti@vwi.unibe.ch.

**Department of Economics, University of Bern, Schanzeneckstrasse 1, 3001 Bern, Switzerland. Mailto: jonas.meier@vwi.unibe.ch.

1 Introduction

Over the last 20 years, healthcare costs have risen substantially. In the US, aggregate expenditures have been growing by roughly 5.7% per year accounting for 17.7% of the Gross Domestic Product (GDP) in 2018 (CMS, 2020). For Switzerland, the same share rose from 12.4% in 2007 to 17.1% in 2017 (OECD, 2019). Given that those expenditures are at least partly publicly financed, governments incorporate growth trends into their budget decisions. Thus, sophisticated predictions are of great importance. A large body of existing studies tackles this issue by predictions based on patients data (e.g. DeSalvo et al., 2009; Morid et al., 2019; Stearns and Norton, 2004).

Considering machine learning techniques, there are several approaches with respect to predicting patients healthcare costs. A recent literature review can be found in Morid et al. (2017). Generally, Gradient Boosting, Artificial Neural Networks and Ridge regressions are shown to add valuable insights. Further, Yang et al. (2018) and Kim and Park (2019) conclude that machine learning methods are important to identify high-cost healthcare consumers. For Switzerland, Jödicke et al. (2019) showed that pharmacotherapy is an important driver of overall expenditures with a Boosted Tree model performing best. Another strand of the literature is concerned with predicting mortality by including healthcare expenditures as predictors. However, Einav et al. (2018) find that machine learning models have little predictive power, even when applied to evolved data. The authors conclude that mortality is fundamentally unpredictable.

In Switzerland, the Swiss Economic Institute (KOF) provides forecasts of aggregate healthcare costs. These models are based on aggregate data on population cohorts, physician density, female labor market participation, recent healthcare expenditures and predictions of wages as well as household incomes. Arguing that those variables are plausibly exogenous, the KOF estimates future costs based on first difference OLS regressions (Köthenbürger and Anderes, 2019) where the variables are chosen according to the Akaike criterion. Roughly every six months, the KOF publishes updated cost predictions. These publications will build our benchmark results to compare our predictions to. The approach of the present analysis differs from these existing predictions in two ways. First and foremost, using only a short panel of individual physician data and a small set of starting variables, the models are not jeopardized by reverse causality or subject to data processing and do not rely on multiple data sources. Second, the employed machine learning techniques flexibly account for complicated patterns without making strong parametric assumptions.

The contribution of this paper is threefold: (i) we propose a novel, tractable and flexible prediction approach for healthcare costs, (ii) our predictions increase the accuracy compared to the existing forecasts by 20% and (iii) we show how policy changes can be evaluated using our proposed method.

The paper is structured as follows. The empirical strategy is outlined in Section 3 where the employed prediction techniques are presented as well. Section 4 provides information on the data. The results are discussed in Section 5. Section 6 shows how machine learning methods may be used to evaluate policy analysis. Specifically, we analyse how a new tariff scheme, being in place from 1st January of 2018 on, changed overall costs. Finally, Section 7 concludes.

2 Motivation

The quality of predictions is subject to available information aswell as the correct specification of the forecast mechanism. If the information set is increased, then a correctly specified model will yield more accurate predictions. Consider the case of observing $p+1$ random variables such that the information set is given by $\mathcal{F} = \sigma(\boldsymbol{x_1}, ..., \boldsymbol{x_p})$. Consider a regression framework where Y is the (dependent) random variable given by a vector $\boldsymbol{y}$ while the $(X_1, ..., X_p)$ are explanatory variables. The canonical regression estimate is then $\tilde{\boldsymbol{y}}_1 = f(X_1, ..., X_p) = E(Y|X_1 = x_1, ..., X_p = x_p)$. We could decide to use only a subset $\mathcal{G} \subset \mathcal{F}$. If, for instance, $\mathcal{G} = \mathcal{F} \setminus X_p$, then an alternative estimate is given by

$\tilde{\boldsymbol{y}}_2 = h(X_1, ..., X_{p-1}) = E(Y|X_1 = x_1, ..., X_p = x_p)$. If $E(Y^2) < \infty$ it follows that almost surely

$$\begin{aligned} E((Y - f(X_1, ..., X_p))^2|X_1, ..., X_{p-1}) &= E(Y^2|\mathcal{G}) - E((E(Y|\mathcal{F})^2|\mathcal{G}) && (1) \\ &\leq E(Y^2|\mathcal{G}) - E(Y|\mathcal{F})^2 && (2) \\ &= E((Y - h(X_1, ..., X_{p-1}))^2|X_1, ...X_{p-1}) && (3) \end{aligned}$$

which is due to the conditional Jensen inequality. It directly follows that the larger information set also yields weakly smaller unconditional squared error:

$$E((Y - f(X_1, ..., X_p))^2) \leq E((Y - f(X_1, ..., X_{p-1}))^2). \tag{4}$$

Thus, using the larger information set when estimating a regression is beneficial for predictions. Besides regressions in first differences, time series models are often employed in order to forecast economic variables. A variation of the argument above holds also for predictions based on time series models like ARMA(X). So far, we abstracted from two important restrictions. i) we have to correctly specify a model and ii) the above result is only valid in asymptotics (if $n \to \infty$ or $n =$ population) as there is no sample uncertainty. While the two issues are closely interlinked in practice, their occurence should be analyzed separately. i) and ii) are both present in the case of overfitting, where information exhibiting spurios correlation for a small sample is incorporated into the model.[1] i) alone occurs if new information includes variables that are temselves dependent on the same set of independent variables as the dependent variable Y.[2] Such simultaneity induces additional bias to the prediction through model misspecification. ii) alone happens when the sample is sufficiently small such that even when a model is not misspecified (the functional form might be correct) generalization of the estimated effects

[1] Spurios correlation in small samples even lead to non-mainstream research areas. Sunspots for instance have been linked to cycles in economic variables since Herschel (1801) observed a relationship between the price of wheat and solar flare activity.

[2] Formally, this translates to a violation of strict exogeneity in the Gauss-Markov Theorem, $E(\varepsilon_i|\mathcal{F}) \neq 0$. Under some circumstances the issue can be resolved with instrumental variable techniques.

is poor. In summary, gathering and using more information about the process one is interested in forecasting is beneficial even if it requires greater care in modeling and a paradigm shift towards non-classical information sources. For practical applications, the issue of sample uncertainty diminishes with "big datasets" that even equal the population.[3] The challenge of correctly specifying a forecasting model increases with the available information set. Examples include Medeiros et al. (2019) exploiting large macroeconomics datasets via machine learning to to predict US CPI-inflation, Henderson et al. (2012) use satellite derived nightlight activity to forecast regional GDP growth in Africa and Varian (2014) examining the effect of an applicants characteristics on her probability of getting a mortgage.

3 Empirical Strategy

We will decompose our prediction algorithm into two separate parts: (i) a prediction on how many physicians will be active in the future and (ii) on the costs of a single physician. As we will employ different techniques for either channel, we present the according strategies in the following.

3.1 Dynamic Panel Modelling

Conceptually, we will predict the number of practising physicians based on two opposed assumptions. On one hand, we will assume that the number of physicians stays constant. On the other hand, we will assume that the number of physicians follows a linear time trend. Formally, the first assumptions is represented in the following equation.

$$\tilde{N}_{A1,s,c,t} = N_{s,c,t_0}, \quad \text{where} \quad t > t_0 \tag{5}$$

[3]Even in such cases, high dimensionality in the number of regressors p can be problematic since the sample density is given by $N^{1/p}$

and N represents the actual and $\tilde{N}$ the predicted number of physicians in specialty s, canton c in a given month t. Further, t_0 represents the month when the prediction is formed and $A1$ indicates that this is the predicted number of physicians under assumption 1. The first assumption may serve as a good starting point but is likely to be inaccurate, especially for long run prediction. Thus, we add a dynamic part to our prediction by forming a linear trend on the national wide specialty level. Accordingly, the second assumption can be written as

$$\tilde{N}_{A2,s,c,t} = f_{s,c,t_0}(\hat{\alpha}_{s,t<t_0} + \hat{\beta}_{s,t<t_0} \cdot t) \tag{6}$$

where f_{s,c,t_0} is the fraction physicians in specialty s and canton c compared to all physicians of that specialty in Switzerland. Further, note that $\alpha_{s,t<t_0}$ and $\beta_{s,t<t_0}$ are the regression coefficients of $N_{s,t<t_0} = \alpha_{s,t<t_0} + \beta_{s,t<t_0} t$.[4] We decided to predict on a national level as in some cantons we only observe very few physicians. With respect to the second assumption, note that it would also be feasible to assume a more involved function than a simple trend. However, in that case, our model is likely to overfit as we only observe the number of physicians on a canton-specialty level. Thus, it is conservative to take a stylized model. In the following, we propose to use a weighted average of the two assumptions. Further, we suggest that the weight of the linear trend assumption, δ, should be higher when forming predictions about the far future. In contrast, one-year ahead predictions may perform well with a low value of δ. We propose to estimate three different scenarios which incorporate the following values of δ: 0.3, 0.5 and 0.7. Taken together, we will predict the number of physicians according to equation (7).

$$\tilde{N}_{s,c,t} = (1 - \delta)\tilde{N}_{A1,s,c,t} + \delta\tilde{N}_{A2,s,c,t} \tag{7}$$

[4] Technically, we only estimated a linear trend when we observed a positive number of physicians of a given specialty at least in five time periods before t_0. Otherwise, we switched to assumption 1 as estimating a trend would be unreliable.

3.2 Predicting the Cost of a Physician

As it is a priori unclear which variables and link functions may be crucial for the predictions, we have chosen to run a large variety of models which use different selection techniques. Apart from the standard Small Least Squares and Lasso models, we run Post- and Plugin-Lasso, Ridge regressions, Elastic Nets as well as Random Forests and Pruned Trees. The models can be categorized into a set of linear models with regularization and models that employ "pure" machine learning. The former can be motivated from standard regression theory while the latter have their roots in decision theory and computer science.[5]

The Lasso regression is suitable if the data is "dense" such that few variables are important drivers of the expenditures. Ridge regression on the other side works well if the data is "sparse" such that many variables contribute little to the expenditures. The Elastic Net, as a combination of Lasso and Ridge, captures both "dense" and "sparse" effects - though it performs inferior if the data is mostly one or the other. Further, we consider two improved versions of the Lasso: the Plug-In Lasso of Belloni et al. (2012) that is valid with non-Gaussian errors and the Post-Lasso of Belloni et al. (2013) which has a smaller bias than the Plug-In Lasso.

The Pruned Tree partitions the sample space via greedy algorithm and performs well if non-linear and complex interactions effect the outcome. We employ a stopping rule such that every partition contains at least 20 observations and use cost-complexity pruning to avoid overfitting. A Random Forest model is an assembly of Trees that overcomes the issue of path-dependency of the greedy algorithm and results in a more robust model. We grow 100 Trees where at every partitioning split, only a randomly drawn set $m \approx \sqrt{p}$ of the p predictors is considered. Predictions are then averaged over the 100 Trees.

The data input for our models contains dummies for year, month, physician specialty, canton, hospital and, for the regression based techniques, the first order interactions be-

[5] Hastie et al. (2009) provides a thorough introduction to machine learning methods from a statistical point of view.

tween year and canton. In a first step we try to select the most promising methods. We compare the models in their 5-fold Cross-Validation predicted Mean Squared Error (MSE). The preferred models are then used to make predictions that mirror the KOF predictions. For our sample period, the KOF released four publications of health cost growth in Switzerland: in June 2015, November 2015, June 2016 and November 2016.[6] In each of these publications, the KOF makes current year, one year ahead and sometimes two year ahead predictions. Given the nature of our data, we belief that current year predictions should be made rather with current billings than a predictive model. Consequently, the 2016, 2017 and 2018 KOF predictions are our benchmark. Two KOF reports predict each the cost growth in 2016 and 2017 but only November 2016 report predicts 2018 for a total 5 benchmarks.

4 Data

The present study employs an physician panel dataset based on monthly entries from 2014 to 2018. The data is provided by the health insurance association SASIS AG. Originally, the source of the data are individual claims covered by the Mandatory Health Insurance (MHI). Further, the data only covers services provided by physicians working in the outpatient sector. Those services are priced according to tariff scheme called *TARMED*, a contract between the healthcare providers and the insurance companies. The data is simplistic in the sense that it only captures five variables which are the monthly expenditures (CHF), the date, information on the physician specialty, a cantonal indicator and whether or not the provider is a hospital or not.[7] Tables X and XI in Appendix B provide detailed information on how expenditures are distributed by physician specialties and Cantons. An attractive feature of the data is that almost all insurance companies are

[6] Note that the publications from November 2015 and June 2016 are based on the same data. Thus, we drop the first prediction to prevent redundancies and only use the more recent prediction as it should be more accurate.

[7] There are around 50 different physician specialties. However, the largest 10 account for about 75% of all revenue. Those are GPs, psychiatrists, group practices, ophthalmologists, OB GYNs, radiologists, pediatricians, medical practitioners, cardiologists and dermatologists (ordered by revenue). Further, note that we only observe hospitals as one entity. Consequently, we treat them as a separate physician specialty.

included. Thus, the data can be taken as representative for the Swiss population.

To ensure comparability of our predictions with the ones of the KOF, we limit ourselves to data available up to the same publication date. However, as the physician data is lagged by 6 months, we are using significantly less current data. Further, we assume that the pool of physicians stays constant.[8] The newly published predictions of the KOF no longer include specifics for the outpatient sector. Therefore, the latest comparable predictions stem from the end of 2016. Finally, note that our data only captures costs covered by the MHI whereas the KOF bases its predictions on the costs of the whole healthcare sector. The difference is roughly 24% in terms of aggregate costs. To guarantee comparability of all predictions, we thus only compare relative deviances from absolute costs. As a benchmark of the actual costs, we utilize the realized expenditures published by the Federal Statistical Office (FSO).[9] On overview over the different data samples can be found in Table IV in Appendix B.

Starting from the 1st of January 2018, a revised tariff scheme is in place. Long-lasting negotiations between payers and providers have been unsuccessful in reaching an agreement on how the outdated scheme should be reformed. Thus, the Federal Council intervened and decided to revise the *TARMED* announcing an annual saving target of 0.47 billion CHF which is roughly a 5% of overall costs. Thus, the time period affected by this revision can hardy be compared to preceding expenditure patterns and is therefore excluded from our baseline predictions. However, in an extension to pure cost predictions, we analyze the effect of the reform in Section 6 by including a variable measuring the physician's loss in revenue.

[8] More precisely, we assume that that all physicians working in one of the last three months before our cut-off date will keep on working. In this way, we are agnostic about physicians entering or leaving the market.

[9] In particular, we use the data in FSO (2019b) and FSO (2019a) for the whole healthcare sector and the costs covered by the MHI, respectively. Note that the FSO uses prediction trends to adjust their aggregate values. Several changes in the employed data altered the costs with retrospective effects. We decided to use the data published in 2019 but methodologically consistent with earlier years to ensure comparability.

5 Results

Table I reports 5-fold CV mean squared prediction error relative to the regular cross-validation Lasso. Both Plug-in and Post Lasso perform exceptionally as well as Pruned Tree and Random Forest models. Likely, tree-based methods are well suited for our prediction exercise because the data mainly captures indicators. From the other models, the Post-Lasso performs the best, though the difference appears to be marginal. With respect to the chosen variables, Canton specific Hospital effects[10] and the physician specialty of Anesthesiology seem to be the most influential predictors. Table XII in Appendix B provides detailed information on the relative importance of the predictors. Due to their superior performance, we only consider the Pruned Tree, the Random Forest and the Post-Lasso[11] for the growth predictions that will be compared to the ones of the KOF.

Table I: 5-fold CV MSE

	CV Lasso	Plug-In Lasso	Post Lasso	Ridge	Elastic-Net	Pruned Tree	Random Forest
MSE	1.0	0.970	0.970	1.0	1.0	0.975	0.980
SE	1.0	0.952	0.948	1.0	1.0	0.948	0.987

Notes: 5-fold CV mean squared prediction error and its standard error relative to CV Lasso.

Table II shows that our models perform reasonably well compared to the KOF prediction. As expected from the MSPE, the Random Forest model performs best. A $\delta = 0.5$, such that the synthetic sample is constructed to equal parts by the linear trend and the current number of physicans, yields the lowest average error and a reduction in the bias. Note also that our model performs worst in the first period and continuously imroves in predictive accuracy. Given that in the first period, only 8 months of data in 2014 are used to predict the costs in 2016, this is no surprise. In total, our Random Forest model outperforms the KOF prediction having an average deviation of 2.3% (for the $\delta = 0.5$ case) compared to the KOF 3.3%. This result is remarkable because of two reasons. First and foremost, simple out-of-the-box machine learning seems to outperform sophisticated pre-

[10] A result coming from the different states of Hospital centralization in the Cantons
[11] Since the Post Lasso basically nests the Plug-In Lasso, we restrict ourselves here.

dictions without any fine-tuning. Second, the employed data is highly simplified. There is substantial room for improvement by additionally including broader economic measures into the prediction models. For instance by including a physican ID as additional information would increase the performance even more.[12]

Table II: Results

$\delta = 0.3$	Growth BFS	Prediction KOF	Growth SASIS	Random Forest	Pruned Tree	Post-Lasso
2016 Prediction March 2015	-1.4%	2.8%	5.0%	0.8%	-1.8%	0.3%
2016 Prediction November 2015	-1.4%	3.8%	5.0%	7.0%	0.9%	2.8%
2017 Prediction November 2015	2.2%	4.0%	2.0%	2.6%	-3.5%	-1.0%
2017 Prediction May 2016	2.2%	4.1%	2.0%	-0.7%	-5.3%	-5.4%
2018 Prediction May 2016		4.2%	0.3%	-1.8%	-6.6%	-6.3%
Average Error		**3.3%**		**2.3%**	**6.1%**	**4.8%**
Bias		**3.8%**		**-1.3%**	**-6.1%**	**-6.1%**

$\delta = 0.5$	Growth BFS	Prediction KOF	Growth SASIS	Random Forest	Pruned Tree	Post-Lasso
2016 Prediction March 2015	-1.4%	2.8%	5.0%	1.0%	-2.2%	0.5%
2016 Prediction November 2015	-1.4%	3.8%	5.0%	7.7%	1.2%	3.4%
2017 Prediction November 2015	2.2%	4.0%	2.0%	4.0%	-2.8%	0.6%
2017 Prediction May 2016	2.2%	4.1%	2.0%	0.0%	-4.9%	-4.4%
2018 Prediction May 2016		4.2%	0.3%	-0.4%	-5.9%	-4.9%
Average Error		**3.3%**		**2.3%**	**5.8%**	**3.8%**
Bias		**3.8%**		**-0.4%**	**-5.8%**	**-3.8%**

$\delta = 0.7$	Growth BFS	Prediction KOF	Growth SASIS	Random Forest	Pruned Tree	Post-Lasso
2016 Prediction March 2015	-1.4%	2.8%	5.0%	-0.6%	-4.1%	-0.2%
2016 Prediction November 2015	-1.4%	3.8%	5.0%	6.8%	0.0%	3.5%
2017 Prediction November 2015	2.2%	4.0%	2.0%	3.6%	-3.6%	1.2%
2017 Prediction May 2016	2.2%	4.1%	2.0%	-0.8%	-6.0%	-4.3%
2018 Prediction May 2016		4.2%	0.3%	-0.9%	-6.7%	-4.2%
Average Error		**3.3%**		**2.6%**	**6.9%**	**3.6%**
Bias		**3.8%**		**-1.9%**	**-6.9%**	**-3.6%**

Notes: The KOF bases its predictions on all physicians and hospitals whereas the SASIS data only covers services provided in the outpatient sector.

[12] For two reasons we decided against it. First, we want to rely on information that might be feasible for a actual implementation - data privacy concerns on modelling retractable physicans costs would hinder such implementation. Second, we do not aim to provide the best prediction possible but rather want to demonstrate the possibilities and compare them in a “fair” setting.

6 Reform 2018

This section aims at estimating the effect of the TARMED revision on aggregate costs. For the specifics of the reform we refer the reader to the summary in Bischof and Meier (2020). We do not intend to model individual physician behaviour as in other studies (e.g. Clemens and Gottlieb, 2014), rather the focus lies on the proposed saving target of 0.47 billion CHF. Equation (8) specifies how we identify the anticipated loss in revenue for a provider i at time t.

$$\Delta\mathrm{rev}_{i,t} = \frac{\sum_s^S \hat{Q}_{i,s,t} \cdot P_{\mathrm{new},s} - \sum_s^S \hat{Q}_{i,s,t} \cdot P_{\mathrm{old},s}}{\sum_s^S \hat{Q}_{i,s,t} \cdot P_{\mathrm{old},s}} \tag{8}$$

$P_{\mathrm{old},s}$ and $P_{\mathrm{new},s}$ refer to the old and new prices of service s and $\hat{Q}_{i,s,t}$ represent predicted quantities. These predictions are described in detail in Appendix A. Intuitively, $\Delta\mathrm{rev}_{i,t}$ measures what fraction of revenue a provider loses due to the reform. As the reform is plausibly exogenous, we include $\Delta\mathrm{rev}_{i,t}$ as continuous treatment into our model. Note that all providers are treated, thus there is no classic control group. We estimate the effect of the treatment $\Delta\mathrm{rev}_{i,t}$ on monthly expenditures in a Causal Forest framework for two reasons. First, the Causal Forest is able to recover valid heterogeneous treatment effects, even with many covariates. Usual regression techniques, where complex interactions of the treatment effect with covariates are infeasible, implicitly assume sparsity. Second, the Causal Forest is "honest" in the sense that it yields valid confidence bands. Honesty is achieve by splitting the training sample into two parts where the first part is used to construct the tree (including the CV step) and the second part for estimating the treatment effects within the leaves of the tree. This contrasts conventional tree models which use adaptive (one-step) estimation of partition and leaf effects.[13]

By assumption, the Causal Forest model request i.i.d. data. In the present case, we argue that even though we use panel data, the monthly variation is substantial such that

[13] Additionally, several adjustments to the estimator, for which we refer to the original paper, are made such that conditional average treatment can be obtained.

fixed effects play a minor role. Since the reform affects all practitioners at the same time such that the sample is the population, a causal forest model is applicable even though the treatment is not randomized. Likely, the assumption on unconfoundedness of the treatment is not violated as the regressors contain information physician specialty. The latter is closely correlated with the treatment. The remaining variation in the treatment is plausibly unrelated to unobservables. In Table III, we report the conditional average partial effect (APE) developed in Athey et al. (2019). The conditional APE is statistically significant for both practitioners and hospitals and the effect size is larger for hospitals than for practitioners. Aggregating the individual APE for 2018 yields a total saving of 361 Mio. CHF. which is considerably lower than the proposed saving target of 470 Mio. CHF. This results is in line with others (e.g. Yip, 1998; Clemens and Gottlieb, 2014) arguing that physicians respond to price cuts by expanding their healthcare supply. With respect to policy decisions, the presented results suggest that policy makers should anticipate this behavioural response beyond the mechanical reduction of costs due to fee cuts.

Table III: (Conditional) Average Partial Effect

	Estimate	Std. Error	Aggregate Effect 2018
Practitioners	-1'105.2	417.8	-134.1 Mio. CHF
Hospitals	-33'484.2	6'917.6	-232.0 Mio. CHF
Total	-1'702.1	426.3	-364.1 Mio. CHF

Notes: Conditional average partial effect of the causal forest for the 2018 TARMED reform.

7 Conclusion

In this paper we introduced machine learning methods to improve upon existing healthcare cost predictions in Switzerland. We find that the Random Forest outperforms current models in a comparable setting. Predicting the next year health cost growth in six cases, the Random Forests increases the accuracy of the estimate by almost 30%. This result is remarkable as we based our predictions on a short and simplified panel data set of physicians.

Further, we employed a machine learning method for causal inference of a cost and billing revision in 2018. The Causal Forest directly incorporates heterogeneity in the causal effect such that many covariates can be used to fit the model to the data. Calculating the aggregate, average partial effect reveals that the proposed saving target was missed by approximately 23%. This result confirms existing findings on physician behaviour and suggests that policy makers should anticipate behavioural responses.

Our work demonstrates that machine learning techniques may provide promising tools to predict healthcare costs. Alternative methods such as Support Vector Machines, Multivariate Adaptive Regression Splines and Neural Networks, known to have comparable performance to the Random Forest, might be employed for such tasks and should therefore be studied. Finally, future research may tackle hybrid models in a formalized manner.

Bibliography

Athey, Susan, Julie Tibshirani, Stefan Wager et al., "Generalized random forests," *The Annals of Statistics*, 2019, *47* (2), 1148–1178.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen, "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 2012, *80* (6), 2369–2429.

_ , Victor Chernozhukov et al., "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 2013, *19* (2), 521–547.

Bischof, Tamara and Jonas Meier, "Physician Induced Demand and Financial Incentives - Evidence from Large-Scale Fee Changes," 2020.

Clemens, Jeffrey and Joshua D Gottlieb, "Do physicians' financial incentives affect medical treatment and patient health?," *American Economic Review*, 2014, *104* (4), 1320–49.

CMS, "National Health Expenditure Fact Sheet," Online 2020.

DeSalvo, Karen B., Tiffany M. Jones, John Peabody, Jay McDonald, Stephan Fihn, Vincent Fan, Jiang He, and Paul Muntner, "Health Care Expenditure Prediction with a Single Item, Self-Rated Health Measure," *Medical Care*, 2009, *47* (4), 440–447.

Einav, Liran, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer, "Predictive modeling of U.S. health care spending in late life," *Science*, jun 2018, *360* (6396), 1462–1465.

FSO, "Kosten des Gesundheitswesens nach Leistungserbringer und Leistungen bis 2017," 2019.

_ , "Kosten des Gesundheitswesens nach Leistungserbringern und Finanzierungsregimes bis 2017," 2019.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media, 2009.

Henderson, J Vernon, Adam Storeygard, and David N Weil, "Measuring economic growth from outer space," *American economic review*, 2012, *102* (2), 994–1028.

Herschel, William, "XIII. Observations tending to investigate the nature of the sun, in order to find the causes or symptoms of its variable emission of light and heat; with remarks on the use that may possibly be drawn from solar observations," *Philosophical Transactions of the Royal Society of London*, 1801, (91), 265–318.

Jödicke, Annika M., Urs Zellweger, Ivan T. Tomka, Thomas Neuer, Ivanka Curkovic, Malgorzata Roos, Gerd A. Kullak-Ublick, Hayk Sargsyan, and Marco Egbring, "Prediction of health care expenditure increase: how does pharmacotherapy contribute?," *BMC Health Services Research*, dec 2019, *19* (1).

Kim, Yeonkook J. and Hayoung Park, "Improving Prediction of High-Cost Health Care Users with Medical Check-Up Data," *Big Data*, sep 2019, *7* (3), 163–175.

Köthenbürger, Marko and Marc Anderes, "KOF Prognose der Gesundheitsausgaben Herbst 2019," 2019.

Medeiros, Marcelo C, Gabriel FR Vasconcelos, Alvaro Veiga, and Eduardo Zilberman, "Forecasting Inflation in a data-rich environment: the benefits of machine learning methods," *Journal of Business & Economic Statistics*, 2019, pp. 1–22.

Morid, Mohammad Amin, Kensaku Kawamoto, Travis Ault, and Samir Abdelrahman, "Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation," *AMIA Annual Symposium Proceedings Archive*, 2017.

_ , Olivia R. Liu Sheng, Kensaku Kawamoto, Travis Ault, Josette Dorius, and Samir Abdelrahman, "Healthcare cost prediction: Leveraging fine-grain temporal patterns," *Journal of Biomedical Informatics*, mar 2019, *91*, 103113.

OECD, "System of Health Accounts 1997-2017," Online 2019.

Stearns, Sally C. and Edward C. Norton, "Time to include time to death? The future of health care expenditure predictions," *Health Economics*, apr 2004, *13* (4), 315–327.

Varian, Hal R, "Big data: New tricks for econometrics," *Journal of Economic Perspectives*, 2014, *28* (2), 3–28.

Yang, Chengliang, Chris Delcher, Elizabeth Shenkman, and Sanjay Ranka, "Machine learning approaches for predicting high cost high need patient expenditures in health care," *BioMedical Engineering OnLine*, nov 2018, *17* (S1).

Yip, Winnie C., "Physician response to Medicare fee reductions: changes in the volume of coronary artery bypass graft (CABG) surgeries in the Medicare and private sectors," *Journal of Health Economics*, dec 1998, *17* (6), 675–699.

Appendix

A Prediction for Individual Quantities

We predict physician-specific service volume based on pre-reform observations. In a first step, the aggregate average change in quantity was computed for service s and provider type p (physician or hospital). In a second step, a percentage prediction for every service and physician speciality pair was computed. Finally, it is assumed that all physicians and all hospitals separately share the same percentage changes of their quantities.

$$\text{Step 1:} \quad \overline{\Delta Q}_{p,s,j} = \frac{1}{T/4} \sum_{j}^{T/12} \Delta Q_{p,s,j+12k} \qquad j \in \{1, .., 12\}$$

$$\text{Step 2:} \quad \%\hat{Q}_{p,s,j_{2018}} = \left(Q_{p,s,j_{2018}-12} + \overline{\Delta Q}_{p,s,j}\right) / Q_{p,s,j_{2018}-12}$$

$$\text{Step 3:} \quad \hat{Q}_{i,s,j_{2018}} = \left(1 + Q_{i,s,j_{2018}-12}\right) \%\hat{Q}_{g,s,j_{2018}}$$

B Tables

Table IV: Healthcare Costs (Mio. CHF), Different Samples

	2014	2015	2016	2017
KOF FSO				
Physicians	8713	9308	9175	9374
Hospitals	6063	6303	6793	6956
Growth (%)		5.65	2.28	2.27
TARMED FSO				
Physicians	6781	7175	7460	7778
Hospitals	4290	4533	4944	5127
Growth (%)		5.76	5.95	4.04
TARMED SASIS				
Physicians	6051	6400	6719	6851
Hospitals	3473	3613	3899	3981
Growth (%)		5.13	6.04	2.03

Notes: The TARMED tariff only covers physicians working in the outpatient sector. In contrast, the KOF bases its predictions on all physicians, thus, aggregate costs are higher. Compared to the SASIS data, the FSO (2019a) additionally covers psychologists. Further, we had to assume how much of laboratory and radiological tests were done in the outpatient sector. We assumed it to be the proportion of outpatient costs to all healthcare costs.

Table V: Random Forest Variable Importance

Name	Node Impurity $e10^8$
Hospital	8005913.836
Month	307790.761
Luzern	196921.226
Solothurn	192174.469
Graubunden	175102.636
Year	133645.655
Freiburg	124037.384
Appenzell Ausserrhoden	88150.445
Bern	84864.046
Gruppenpraxen	83540.635
Wallis	83032.573
Basel-Stadt	66061.572
Radiologie	63151.250
Schwyz	55307.793
Basel-Landschaft	49137.142
Thurgau	48174.084
Allg. Innere Medizin	46895.573
canton.f16	39013.447
canton.f22	37687.402
canton.f1	36693.626
canton.f17	34360.133
canton.f21	33515.928
canton.f9	33014.864
canton.f4	31805.344
canton.f19	31395.652
canton.f25	30809.200
undergroup.f10	28970.662
canton.f8	25889.615
undergroup.f13	20424.723
canton.f24	17419.027
undergroup.f45	14990.065
undergroup.f53	13903.801
undergroup.f65	12682.861
canton.f14	7955.777
canton.f26	7332.864
Orthopadische Chirurgie	6950.987
Chirurgie	6811.305
Gynakologie	6189.754
Radio-Onkologie	4407.134
Kardiologie	3974.739
Jugendpsychiatrie	3696.973
AnÃ¤sthesiologie	2615.623
Kinder- und Jugendmedizin	2111.742

Notes: Variable importance based on node impurity of the Random Forest trained on the first three years of data for the prediction. Mean Decrease in Node Impurity $e10^8$.

Table VI: Random Forest Variable Importance

Name	Node Impurity $e10^8$
undergroup.f3	1589.341
undergroup.f22	1281.069
canton.f7	1280.240
canton.f6	1278.121
undergroup.f15	1236.989
undergroup.f99	1203.467
undergroup.f67	910.518
undergroup.f6	844.870
undergroup.f7	778.018
undergroup.f51	753.225
undergroup.f59	658.414
undergroup.f36	627.783
undergroup.f20	612.813
undergroup.f63	448.400
undergroup.f64	367.675
undergroup.f8	342.216
undergroup.f9	331.004
undergroup.f56	305.299
undergroup.f19	248.574
undergroup.f52	160.186
undergroup.f57	153.332
undergroup.f5	121.188
undergroup.f54	118.396
undergroup.f71	114.469
Arbeitsmedizin	107.829
Kinderchirurgie	92.820
Intensivmedizin	71.341
Reisemedizin	62.551
Medizinische Genetik	45.504
PrÃ¤vention	33.264
Pharmazeutische Medizin	28.651
Gefasschirurgie	17.424
Klinische Pharmakologie	8.361
OKP-Ausstand	5.906
Thoraxchirurgie	1.942

Notes: Variable importance based on node impurity of the Random Forest trained on the first three years of data for the prediction. Mean Decrease in Node Impurity $e10^8$.

Table VII: Heterogeneous CATE, Cantons

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	-1757.314	1137.287	0.122
H	-130130.708	1576.599	0.000
Zurich	-1809.757	1231.067	0.142
Bern	-3356.544	1708.739	0.049
Luzern	-4628.628	2078.986	0.026
Uri	-843.909	2269.553	0.710
Schwyz	-1593.550	1612.093	0.323
Obwalden	-7089.163	1700.084	0.000
Nidwalden	-965.012	3207.930	0.764
Glarus	-10223.870	3756.891	0.007
Zug	-2060.213	1943.785	0.289
Freiburg	-2570.703	1450.755	0.076
Solothurn	-2284.833	1617.372	0.158
Basel-Stadt	-3066.788	1273.829	0.016
Basel-Landschaft	-4906.678	1337.121	0.000
Schaffhausen	-4829.277	1676.276	0.004
Appenzell Ausserrhoden	3565.707	2555.895	0.163
Appenzell Innerrhoden	-4935.073	2039.577	0.016
St. Gallen	-10511.018	4344.600	0.016
GraubÃ¼nden	-2254.171	1325.904	0.089
Aargau	-4253.983	1259.759	0.001
Thurgau	-3063.042	1425.730	0.032
Tessin	-1452.725	1270.318	0.253
Waadt	-1556.671	1771.074	0.379
Wallis	-2031.006	1367.631	0.138
Neuenburg	-724.438	1343.909	0.590
Genf	-81.913	1218.360	0.946

*Notes:*Let $tau(Xi) = E[Y(1) - Y(0)|X = Xi]$ be the CATE, and Ai be a vector of user-provided covariates. This function provides a (doubly robust) fit to the linear model $tau(Xi)\ beta_0 + Ai * beta$. Best linear projection of the conditional average treatment effect. Confidence intervals are cluster- and heteroskedasticity-robust (HC3).

Table VIII: Heterogeneous CATE, Specialty 1/2

	D_n%	Estimate	Std. Error	Pr(>|t|)
Anasthesiologie	1.4	3687.816	336.367	0.000
Chirurgie	1.3	3329.500	323.455	0.000
Dermatologie	1.5	1978.511	356.052	0.000
Gynakologie	1.6	5025.033	616.121	0.000
Innere Medizin	5.4	-382.721	392.046	0.329
Endokrinologie	-3.0	2000.366	353.826	0.000
Pneumologie	2.7	1626.567	328.541	0.000
Neurochirurgie	-0.4	3335.834	480.328	0.000
Neurologie	-4.0	2322.432	359.907	0.000
Psychiatrie	-3.1	4959.799	1051.785	0.000
Jugendpsychiatrie1	-2.9	2851.495	396.497	0.000
Ophthalmologie	2.2	-5040.605	4584.361	0.272
Orthopadische Chirurgie	2.1	13076.065	4905.131	0.008
HNO	2.3	5552.080	364.200	0.000
Jugendmedizin	3.7	6525.197	791.321	0.000
Radiologie	-9.2	-7723.662	470.124	0.000
Reisemedizin	1.7	2260.584	488.617	0.000
Urologie	4.0	2999.414	372.833	0.000
Rheumatologie	-1.4	3345.425	364.240	0.000
Allgemeine Innere Medizin	4.3	7645.539	370.173	0.000
Angiologie	-3.6	318.409	292.515	0.276
OKP-Ausstand	-	445.084	539.846	0.410
Kardiologie	-9.3	-94.379	948.166	0.921
Hamatologie	-1.0	3074.922	393.972	0.000
Gastroenterologie	-9.3	-828.989	345.036	0.016

Notes: Best linear projection of the conditional average treatment effect. Confidence intervals are cluster- and heteroskedasticity-robust (HC3).

Table IX: Heterogeneous CATE, Specialty 2/2

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	-1757.314	1137.287	0.122
H	-130130.708	1576.599	0.000
Radio-Onkologie	-11888.237	1326.636	0.000
Nuklearmedizin	-8469.190	1529.626	0.000
Rehabilitation	2860.343	465.976	0.000
praktischer Arzt	8183.380	1322.848	0.000
Allergologie	766.345	421.046	0.069
Arbeitsmedizin	1233.059	498.866	0.013
Herz- und Gefasschirurgie	2264.034	394.054	0.000
Infektiologie	1419.661	379.617	0.000
Intensivmedizin	1435.170	547.046	0.009
Kiefer- und Gesichtschirurgie	3785.470	535.515	0.000
Kinderchirurgie	1955.401	443.872	0.000
Pharmakologie	-482.609	462.641	0.297
Medizinische Genetik	9284.349	1847.502	0.000
Medizinische Onkologie	1435.395	348.439	0.000
Nephrologie	2804.821	430.109	0.000
Pathologie	-6196.121	573.286	0.000
Pharmazeutische Medizin	4590.823	2363.679	0.052
Plastische Chirurgie	2060.935	348.077	0.000
Pravention	589.260	429.887	0.170
Handchirurgie	1519.602	418.278	0.000
Gefasschirurgie	1683.272	1044.057	0.107
Thoraxchirurgie	2681.059	1252.720	0.032
Gruppenpraxen	239.437	407.009	0.556
Spezialfalle	-2350.693	387.453	0.000

Notes: Best linear projection of the conditional average treatment effect. Confidence intervals are cluster- and heteroskedasticity-robust (HC3).

Table X: Healthcare Costs (Mio. CHF), 10 largest Pysician Specialties

	2014		2015		2016		2017		2018	
	Abs.	Δ%	Abs.	Δ%	Abs.	Δ%	Abs.	Δ%	Abs.	Δ%
Hospitals	3473		3613	4.03	3899	7.92	3981	2.1	3857	-3.11
General Practitioner	1537		1606	4.49	1591	-0.93	1525	-4.15	1509	-1.05
Psychiatrists	663		689	3.92	726	5.37	757	4.27	773	2.11
Group Practices	373		458	22.79	558	21.83	636	13.98	708	11.32
Ophthalmologists	520		554	6.54	591	6.68	623	5.41	598	-4.01
OB GYNs	393		411	4.58	425	3.41	422	-0.71	423	0.24
Radiologists	352		360	2.27	368	2.22	383	4.08	365	-4.7
Pediatricians	296		322	8.78	345	7.14	349	1.16	363	4.01
Medical Practitioner	238		267	12.18	284	6.37	290	2.11	288	-0.69
Cardiologists	191		199	4.19	212	6.53	224	5.66	226	0.89
Rest	1486		1534	3.23	1620	5.61	1640	1.23	1614	-1.59
Total	**9522**		**10013**	**5.16**	**10619**	**6.05**	**10830**	**1.99**	**10724**	**-.98**

Notes: Data: SASIS AG. The physician specialties are ordered by 2017 gross costs. Absolute costs refer to Mio. CHF and the growth is measured in percentages.

Table XI: Healthcare Costs (Mio. CHF), 10 largest Cantons

	2014		2015		2016		2017		2018	
	Abs.	Δ%	Abs.	Δ%	Abs.	Δ%	Abs.	Δ%	Abs.	Δ%
ZH	1919		2013	4.9	2151	6.86	2190	1.81	2178	-0.55
BE	1165		1228	5.41	1281	4.32	1297	1.25	1276	-1.62
VD	1087		1140	4.88	1219	6.93	1249	2.46	1218	-2.48
GE	780		831	6.54	887	6.74	910	2.59	904	-0.66
AG	616		647	5.03	686	6.03	699	1.9	696	-0.43
SG	507		527	3.94	556	5.5	564	1.44	559	-0.89
TI	406		430	5.91	458	6.51	475	3.71	484	1.89
LU	423		437	3.31	460	5.26	474	3.04	467	-1.48
BS	411		426	3.65	458	7.51	467	1.97	460	-1.5
BL	313		328	4.79	342	4.27	344	0.58	339	-1.45
Rest	1898		2005	5.64	2120	5.74	2165	2.12	2147	-0.83
Total	**9525**		**10012**	**5.11**	**10618**	**6.05**	**10834**	**2.03**	**10728**	**-.98**

Notes: Data: SASIS AG. The Cantons are ordered by 2017 gross costs. Absolute costs refer to Mio. CHF and the growth is measured in percentages.

Table XII: Variable Importance

	Model			
	OLS	CV Lasso	CV Ridge	CV Elastic Net
1	Date (158.3)	SO×H (1.1)	SO×H (1.1)	SO×H (1.0)
2	SO×H (1.5)	H (0.78)	LU×H (0.51)	H (0.72)
3	H (1.1)	Pneumology (0.33)	FR×H (0.44)	Pneumology (0.31)
4	HMO-Group (0.58)	LU×H (0.33)	Inner Medicin (0.41)	LU×H (0.27)
5	Pathology (0.54)	Anesthesiology (0.15)	BS×H (0.32)	Anesthesiology (0.13)
6	LU×H (0.52)	VD×H (0.1)	SG×H (0.31)	VD×H (0.09)
7	Hand-Surgery (0.46)	BS×H (0.08)	VD×H (0.28)	BE×H (0.04)
8	Vascular-Surgery (0.45)	FR×H (0.06)	BE×H (0.26)	BS×H (0.03)
9	Public Health (0.43)	BE×H (0.06)	TI×H (0.23)	SG×H (0.02)
10	Plastical-Surgery (0.42)	SG×H (0.06)	Pneumology (0.22)	Radiology (0.01)
	Model			
	Plug-In Lasso	Post Lasso	Pruned Tree	Random Forest
1	H (1.0)	SO×H (1.5)	Pneumology	Anesthesiology
2	SO×H (0.8)	H (1.1)	Anesthesiology	H
3	Pneumology (0.34)	Pneumology (0.38)	H	Pneumology
4	Inner Medicin (0.29)	Inner Medicin (0.38)	VS	VS×H
5	Anesthesiology (0.17)	Anesthesiology (0.22)	VD	VS
6	Radiology (0.1)	Pathology (0.12)	GE	BE×H
7	Pathology (0.1)	Radiology (0.11)	LU	TI×H
8	HMO-Group (0.09)	HMO-Group (0.09)	BE	VD×H
9	Nuclearmedicin (0.05)	Nuclearmedicin (0.09)	Radio-Onkology	BE
10	Radio-Onkology (0.05)	Radio-Onkology (0.08)	TI	MONTH

Notes: With coefficients in Mio. CHF if available. For the Tree based models, we report the 10 most important variables by measuring the total decrease in node impurities from splitting on the variable. For the regression based methods, we report the 10 variables with largest coefficients without weighting.