Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Risk Adjustment Revisited using Machine Learning Techniques

Álvaro J. Riascos[1]
University of Los Andes and Quantil, Bogotá, Colombia
Mauricio Romero[2]
University of California, San Diego, United States
Natalia Serna[3]
University of Wisconsin-Madison, Madison, United States

Abstract. Risk adjustment is vital in health policy design. Risk adjustment defines the annual capitation payments to health insurers and is a key determinant of insolvency risk for health insurers. In this study we compare the current risk adjustment formula used by Colombia's Ministry of Health and Social Protection against alternative specifications that adjust for additional factors. We show that the current risk adjustment formula, which conditions on demographic factors and their interactions, can only predict 30% of total health expenditures in the upper quintile of the expenditure distribution. We also show the government's formula can improve significantly by conditioning ex ante on measures indicators of 29 long-term diseases. We contribute to the risk adjustment literature by estimating machine learning based models and showing non-parametric methodologies (e.g., boosted trees models) outperform linear regressions even when fitted in a smaller set of regressors.
Keywords. Risk adjustment; Diagnostic related groups; Risk selection; Machine learning.

1 Introduction

The last three decades have seen several restructuring of public health systems around the world. In Colombia, law 100 of 1993, transformed the public health system into a competitive insurance market with three key components to it: 1) a benefits package (POS) that defines all the services, procedures, and medications each enrollee has the right to claim; 2) a group of health insurers (EPS) who configure a network of health service providers in charge of delivering all services listed in the benefits package; and 3) a mechanism for the payment of such services that controls for enrollee heterogeneity and risk. Enrollees' contributions configure a cross subsidies system that helps insurers mitigate their financial risk and reduce the incentives to "cream skim" [11]. Monthly risk premium fees are collected from all enrollees with formal employment and then redistributed in

[1]ariascos@uniandes.edu.co
[2]mtromero@ucsd.edu
[3]nserna@wisc.edu

form of capitation payments (UPC) to every health insurer in the market. The redistribution is based on the specific risk profile of the insurer's population of enrollees. The goal of a risk-adjustment mechanism is to reduce annual health expenditure uncertainty as much as possible by controlling for variables that are not subject to manipulation by health insurers. In this document, we propose a new risk-adjustment formula based on information available to the Ministry of Health. We propose adjusting for indicators of 29 long-term disease groups and indicators of hospitalizations, consults with specialists and admissions to the intensive care unit. Given their predictive ability, we use machine learning algorithms to predict annual health expenditures and show that models based on machine learning techniques predict more accurately the annual health expenditure of individuals in an specific quintile of the health expenditure distribution.

2 Risk adjustment literature

In a study for the Colombian health sector, [8] propose a new way to calculate the UPC by including additional risk factors describing the morbidity distribution of the enrollee population. The authors show that an ex ante morbidity risk adjustment using these long-term disease groups approximates better the empirical distribution of annual health expenditures and, thus, reduces capital requirements of health insurers in the contributory system. Other studies have shown that interactions between diagnoses and demographic variables improve the predictions of models for annual health expenditure. For instance, [10] compare two categorizations of ICD-9 ambulatory diagnoses and show that a model adjusted on demographics and diagnoses predicts more accurately the per capita health expenditure in the Medicare system in the United States. [3] also find that an index of weighted ICD-9 codes explain 10.66% of the variation in 1993 annual health expenditures in the United States. More recently, machine learning techniques have incurred in the field of health risk adjustment. [1] in a study for Germany, estimate a regression tree to find all relevant interactions between regressors, and then include such interactions in a linear regression model. The authors show including interactions from a machine learning algorithm improves the R^2 of the lineal regression by 10 percentage points compared to the government formula. [5] argue the non linear relation between risk factors is usually very difficult to capture with linear models. In that matter, the authors estimate a random forest model to find relevant interactions and patterns in the claims data of Medicare. The random forest model reaches a R^2 of 38% with an standard deviation on 0.008 while the linear regression model reaches a R^2 of 31% with a standard deviation on 0.01.

3 Data and description

To predict the annual health expenditure of enrollees to the contributory system in Colombia we use the *Base de Suficiencia* of the Ministry of Health and Social Protection for the years of 2010 and 2011. We use the demographic characteristics and diagnoses received by each enrollee during 2010 to predict the annual health expenditure in 2011 adjusted by the number of enrolled days in this year. For each enrollee we observe gender, age, municipality of residence, insurer, provider, service cost, and ICD-10 diagnosis [4]. We categorize the municipality of residence in the three payment geographic areas defined by the National Administrative Department of Statistics (DANE) (urban, normal and special). ICD-10 diagnoses are grouped in 29 long-term disease pools [2]. Table 1 shows some descriptive statistics of the cross-sections and the panel of enrollees. During 2010 there were nearly 24 million enrollees in the contributory system and during 2011 this number increases to 25 million. The intersection of enrollees between these two cross-sections consists of 13 million people. Notice that the individuals who are active enrollees in both years have an average enrolled-days weighted health expenditure higher than the average expenditure in the cross-sections.

Table 1: Comparison of 2010 and 2011 cross-sections with the panel of enrollees

	2010 cross-section	2011 cross-section	Panel
Number of enrollees	24,354,254	25,695,491	13,652,533
Capitation Payment (UPC) 2010	565,563		732,499
Capitation Payment (UPC) 2011		568,417	695,776
% of enrollees with long-term diseases during 2010	26.03		25.99
% of enrollees with long-term diseases during 2011		26.19	26.17
% of enrollees with long-term diseases in 2010 that upturn in 2011	9.85		7.86
% of enrollees without long-term diseases in 2010 that are diagnosed in 2011		10.01	10.00
Average enrolled days in 2010	306.5		310.8
Average enrolled days in 2011		303.4	311.0
% of enrollees who claim services in 2010	64.14		98.9
% of enrollees who claim services in 2011		64.05	83.2

Note: This table shows some descriptive statistics of the 2010 and 2011 cross-sections of the Base de Suficiencia and the panel built from their intersection. We report the number of enrollees in each dataset, the UPC ($360 \times \sum_i x_i / \sum_i d_i$), the percentage of enrollees with long-term diseases, and the percentage of those whose health condition changes from one year to the other. We also show the percentage of enrollees who claim at least one health service every year.

For prediction of annual health expenditures and to avoid overfitting problems, we build two mutually exclusive datasets by randomly selecting 500,000 enrollees each, from the intersection of 13 million individuals. One dataset is the training set where we will fit all of our models and the other is the test set where we will compute the fitting measures.

4 Results

In this section we show the out-of-sample fitting metrics for the linear models and the machine learning based models both the full expenditure distribution as well as for the upper and lower quintiles of the observed expenditure distribution. To build the models we take the following individual traits: demographic characteristics ($Demog$), indicators of long-term disease groups (Dx) indicators of hospitalizations (H), consults with specialists (E) and admissions to the intensive care unit (U). The machine learning based models considered in this research were random forest (RF), boosted trees models (GBM) and artificial neural network (ANN). All the parameters of these models were obtained using cross validation. In the case of the RF model, each tree is estimated using a random subset of predictors of size $\sqrt{X}$, where X is the total number of predictors. As shown in [6] the RF model is robust to different tuning parameters including the cardinality of the set of predictors on which each tree in the forest is trained. We find that the

[4]The 10th revision of the International Statistical Classification of Diseases and Related Health Problems is a medical classification compiled by World Health Organization

optimal number of trees is 680, which is close to what other authors have used when predicting health outcomes with claims data and an imbalanced number of patients in each diagnosis group [4]. We use the same number of trees for estimating the GBM model. This model consists of fitting decision trees sequentially unlike the RF model where it is done simultaneously. Each tree in the GBM is trained on a weighted sample from the previous tree. The weights are chosen such that observations that the previous tree misclassified receive a higher weight than those that were correctly classified. Subsequent trees are also trained on the predictions of previous trees multiplied by a parameter known as the shrinkage parameter, which helps in the regularization method to avoid overfitting the data. Cross-validation shows that a relative large shrinkage parameter and a relative small number of interactions between the variables achieved the best prediction metrics in the train sample. These parameters differ from what other studies have found including [9], who use 232 trees in the boosting algorithm to predict health costs using interactions between ICD-9 codes. For the ANN, we used 5 inner layers and the back-propagation algorithm to estimate the optimal weight for each predictor in each layer. We use a regularization method that consists of adding a penalty to the criterion function -the sum of squared residuals- that is equal to the sum of squared weights multiplied by a "decay" parameter. The decay of weights is chosen using cross-validation.

In any case, "Two stages" stands for the inclusion of the probability of claiming services in the variable set, WLS stands for weighted least squares, and "FS" indicates the model is fitted on the 21 variables resulting from the feature selection procedure.

Table 2: Out-of-sample fitting measures in the full distribution

Parameters	Model	RMSE (COP)	MAE (COP)	PR annualized	PR non annualized	R^2
	WLS UPC	3,506,658	720,587	0.896	0.999	1.57
	WLS UPC + Dx	3,440,928	694,404	0.892	0.999	5.23
	WLS UPC + Dx + H	3,437,175	694,005	0.894	1.000	5.45
	WLS UPC + Dx + H + E	3,435,470	691,169	0.892	0.999	5.53
	WLS UPC + Dx + H + E + U	3,431,842	688,771	0.892	0.999	5.73
	WLS UPC x H x E x U + Dx	3,432,097	683,209	0.893	0.999	5.71
5 + 0.05	ANN Demog + Dx (Two stages)	3,470,597	816,431	1.072	1.203	3.59
5 + 0.05	ANN Demog + Dx + H (Two stages)	3,455,518	785,698	1.058	1.189	4.42
5 + 0.05	ANN FS	3,455,366	774,190	1.064	1.179	4.43
680	RF Demog + Dx + H + E + U (Two stages)	3,607,413	735,027	0.981	1.097	-4.16
680	RF FS	3,465,301	712,820	0.975	1.087	3.88
680 + 3 + 0.1 + 50	GBM Demog + Dx + H + E + U (Two stages)	3,436,299	719,029	0.997	1.111	5.48
680 + 3 + 0.1 + 50	GBM FS	3,431,044	721,168	1.002	1.115	5.77

Note: This table shows the RMSE, MAE, R^2, and annualized and non-annualized predictive ratios in the full sample. The first column shows the parameters with which the machine learning based models were trained. For the neural networks (ANN) the first number corresponds to the number of neurons in the inner layer and the second is the weight decay parameter. For the random forest model (RF) the number indicates the number of trees, and for the boosted trees model (GBM) the correspond to the number of trees, the dimension of variable iteractions, the contraction parameters and the minimum number of observations in non-terminal nodes, respectively. *Two stages* indicates the model includes the probability of claiming a service and *FS* that the model is fitted over the set of variables chosen using feature selection. *WLS* stands for weighted least squares. The RMSE and the MAE are reported in 2011 colombian pesos. Authors' calculations from the *Base de Suficiencia*.

Table 2 shows two of the most common metrics used to measure accuracy for continuous variables, the root-mean-square error (RMSE) and the mean absolute error (MAE) (both measure the differences between values predicted and the values actually observed), coefficient of determination (R^2) and predictive ratios (PR) in the full sample are also presented. Results show linear models, and in particular the current government formula, tend to underestimate the entire health expenditure distribution by nearly 11%. Underestimation is problematic because it leaves a portion of the health risk unassured. Although inclusion of an ex ante morbidity risk adjustment with the dummy variables for the 29 long-term diseases reduces the MAE by 30,000 pesos and the RMSE by 60,000 pesos compared to the government formula, the inclusion does generate an increase in the overall predictive ratio. On the contrary, machine learning models achieve a predictive ratio that is closer to 1 than the linear models. Notice that models that exceed the predictive ratio threshold of 1 are not suitable because they are overcompensating all insurers and increasing the health system's expenditures. By guaranteeing the full sample predictive ratio is less than or equal to one we are focusing on more efficient redistributions of the current level of health expenditure. The best model in this table is the boosted trees model that is fitted over the set of variables chosen through feature selection (GMB FS), which reaches an annualized predictive ratio of 1.002, a MAE of 721,168 pesos, a RMSE of 3,431,044 pesos and a R^2 of 5.77% that outperforms the linear models fitted over the entire set of regressors[5]. These results are also consistent with findings in [7] who mention that the boosting algorithm is able to explain almost 90% of the variance in individual health costs for the Medicare Prospective Payment System.

Table 3 shows the out-of-sample fitting metrics in the highest quintile of the observed expenditure distribution. Underestimation in this case increases the insolvency risk of those insurers with the sickest population of enrollees. Therefore, the criteria to choose the best model will be the one that achieves the closest to 1 predictive ratio. The neural network model that includes the probability of claiming a service and the hospitalization indicator predicts 53% of the non annualized health expenditure and 45% of the annualized health expenditure in this tail of the distribution. These percentages represent an improve of 7 percentage points compared to the linear models. The GBM FS model fitted on a much smaller set of variables is able to predict 50% of total health expenditures which represents an improve of 5 percentage points relative to the government's formula. Results in this table suggest the current government formula (WLS UPC) is little predictive of the annual health expenditure at the upper tail of the expenditure distribution, only 33.5%. Results also show the ex ante morbidity risk adjustment using the 29 long-term disease groups improves by 10 percentage points the government's formula predictive ratios.

Figure 1 presents the variable relative importance (an average of how many times is a variable used for the recursive partitioning of the trees) in the GBM FS model. The most relevant variables for prediction of annual health expenditures are in order: the probability of claiming a service, the indicator of consults with specialist, the indicator of hospitalizations, the 45 to 49 age group, and the indicator of urban areas. In relation to diagnoses categories, the most predictive groups are: cardiovascular diseases, cervical cancer, hypertension and breast cancer.

4.1 Incentives to risk selection

Overestimation of the annual health expenditure of enrollees in the lower tail of the observed expenditure distribution generates incentives to risk selection because insurers

[5]Although for the lower quintile of the observed expenditure distribution in 2011 all the models highly overestimate the annual health expenditure.

Table 3: Out-of-sample fitting metrics in the upper quintile of the expenditure distribution

Parameters	Model	RMSE	MAE	PR annualized	PR non annualized
	WLS UPC	7,749,235	1,920,486	0.291	0.335
	WLS UPC + Dx	7,580,659	1,983,269	0.367	0.426
	WLS UPC + Dx + H	7,569,594	1,974,414	0.380	0.440
	WLS UPC + Dx + H + E	7,565,491	1,963,559	0.388	0.450
	WLS UPC + Dx + H + E + U	7,555,009	1,965,934	0.390	0.452
	WLS UPC x H x E x U + Dx	7,555,760	1,981,378	0.393	0.456
5 + 0.05	ANN Demog + Dx (Two stages)	7,576,299	2,045,276	0.444	0.517
5 + 0.05	ANN Demog + Dx + H (Two stages)	7,558,045	2,000,360	0.454	0.526
5 + 0.05	ANN FS	7,582,293	1,962,318	0.412	0.474
680	RF Demog + Dx + H + E + U (Two stages)	7,780,452	2,118,572	0.445	0.520
680	RF FS	7,580,672	1,988,824	0.424	0.490
680 + 3 + 0.1 + 50	GBM Demog + Dx + H + E + U (Two stages)	7,532,498	1,988,610	0.436	0.505
680 + 3 + 0.1 + 50	GBM FS	7,517,520	1,961,026	0.430	0.500

Note: This table shows the RMSE, MAE and annualized and non-annualized predictive ratios in the upper quintile of the observed annual health expenditure distribution for 2011. The models and their parameters as the same as in table (2). Authors' calculations from the *Base de Suficiencia.*

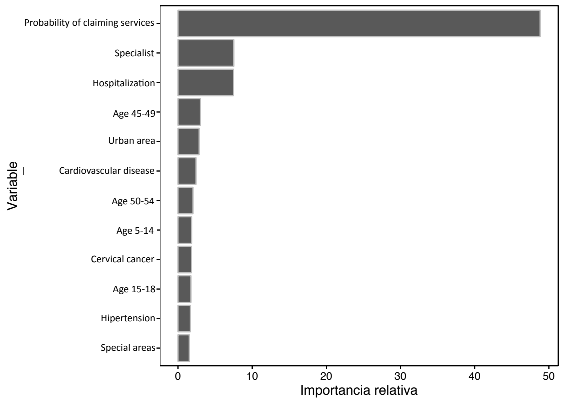

Figure 1: Variable relative importance in the GBM FS model

will tend to enroll only those individuals whose health expenditure is low but whose prediction, on which per capita payments are based, is high. In this subsection we study the incentive to risk selection by comparing the profit generated by those enrollees for which the models overestimate the annual health expenditure. Table 4 shows the percentage of enrollees for which the prediction of annual health expenditure is higher than the observed expenditure ($(1/N)\sum_{i=1}^{N} w_i I(\hat{y}_i > y_i)$) and the profit they would generate ($\sum_{i=1}^{N} w_i(\hat{y}_i - y_i)$). In terms of these results, a model is desirable if: (i) the percentage of enrollees and the percentage of profit with respect to income is similar and (ii) the absolute level of profit is low compared to the rest of the models.

Linear models generate the lowest profits due to risk selection compared to the rest of the models. However, the government's formula is dominated by the linear models that adjust ex ante for the morbidity distribution of the population of enrollees. The linear model with indicators of long-term diseases, hospitalization, consult with specialist, and admission to the intensive care unit, generates 172,051 million pesos of profit by enrolling individuals for which they overestimate the annual health expenditure. In this model, there is 5 percentage point difference between the percentage of enrollees and the percentage of profits, which is larger than what machine learning based models achieve. The boosted trees model fitted over the entire set of predictors overestimates the annual health expenditure for 70% of the population of enrollees, but this percentage of individuals is responsible for 73% of the system's profits.

Table 4: Incentives to risk selection

Parameters	Model	Enrollees (%)	Profit ($)*	Profit (%)
	WLS UPC	70.39	180,046	73.39
	WLS UPC + Dx	67.82	173,432	72.32
	WLS UPC + Dx + H	65.66	173,464	72.29
	WLS UPC + Dx + H + E	64.02	172,739	71.93
	WLS UPC + Dx + H + E + U	64.50	172,051	71.92
	WLS UPC x H x E x U + Dx	67.52	175,454	71.55
5 + 0.05	ANN Demog + Dx (Two stages)	57.96	234,627	76.19
5 + 0.05	ANN Demog + Dx + H (Two stages)	60.15	224,829	74.62
5 + 0.05	ANN FS	57.01	251,677	76.68
680	RF Demog + Dx + H + E + U (Two stages)	67.85	198,359	73.97
680	RF FS	69.77	191,198	72.33
680 + 3 + 0.1 + 50	GBM Demog + Dx + H + E + U (Two stages)	70.53	196,517	72.38
680 + 3 + 0.1 + 50	GBM FS	70.21	197,569	72.58

Note: This table shows the percentage of enrollees for which models overestimate the annual health expenditure $(1/N)\sum_{i=1}^{N} w_i I(\hat{y}_i > y_i)$, the profit they would generate $\sum_{i=1}^{N} w_i(\hat{y}_i - y_i)$, and the profit as percentage of income. Authors' calculations from the *Base de Suficiencia.*

5 Conclusions

We show the current risk adjustment formula that conditions on sociodemographic factors and their interactions, can only predict 30% of total health expenditures in the upper quintile of the expenditure distribution. We show the government's formula can improve significantly by conditioning ex ante for any measure of the morbidity distribution of enrollees such as the indicators of 29 long-term diseases, going from a 30% to a 40% prediction of total annual health expenditure in the upper tail of the distribution. Our main finding is that non parametric models based on machine learning techniques like the boosted trees model, outperforms by 5 percentage points the predictive ratio, by 40,000 pesos the RMSE and by 20,000 the MAE in the last quintile of the expenditure distribution, compared to the government's formula. This document evidences how the risk adjustment policy in Colombia can redistribute resources more efficiently by adjusting for the enrollees' health condition ex ante and by using non parametric specifications that capture the non linear relation between risk factors better than the linear models.

Acknowledgments

Financial support from the Ministry of Health and Social Protection is greatly appreciated. This research agenda and previous version of this paper have benefit greatly from comments and suggestions from Eduardo Alfonso, Luis Gonzalo Morales, Maria Clara Correa, Giovanni Hurtado, Alvaro Lopez, Juan Carlos Linares, Diana Estupinhan, Dov Chernichovsky, Wynand van den Ven, Mark Basset, Fernando Montenegro, Ramiro Guerrero, as well as helpful discussions with workshop participants at the TAUB Center for Social Policy Studies in Israel (workshop funded by the World Bank), the CEDE economics seminar at the University of Los Andes and a medical team from Fundación Valle del Lili in Cali. The usual disclaimer applies.

References

[1] F. Buchner, J. Wasem, and S. Schillo. Regression trees identify relevant interactions: Can this improve the predictive performance of risk adjustment? *Health economics*, 2015.

[2] A. Eduardo, A. Riascos, and M. Romero. The performance of risk adjustment models in colombia competitive health insurance market. 2013.

[3] J. Hughes, J. Averill, R.and Eisenhandler, N. Goldfield, J. Muldoon, J. Neff, and J. Gay. Clinical risk groups crgs: a classification system for risk-adjusted capitation-based payment and health care management. *Medical care*, 42(1):81–90, 2004.

[4] M. Khalilia, S. Chakraborty, and M. Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, 11(51):1–13, 2011.

[5] L. Li, S. Bagheri, H. Goote, A. Hasan, and G. Hazard. Risk adjustment of patient expenditures: A big data analytics approach. In *Big Data, 2013 IEEE International Conference on*, pages 12–14. IEEE, 2013.

[6] D. Palmer, N. O'Boyle, and J. Mitchell. Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modelling*, 47(1):150–158, 2007.

[7] D. Relles, G. Ridgeway, and G. Carter. Data mining and the implementation of a prospective payment system for inpatient rehabilitation. *Health Services and Ooutcomes Research Methodology*, 3(4):247–266, 2002.

[8] A. Riascos and N. Serna. Predicting annual length-of-stay and its impact on health costs: the case of the colombian health care system. February 2017.

[9] J. Robinson. Regression tree boosting to adjust health care costs predictions for diagnostic mix. *Health Services Research*, 43(2):755–772, 2008.

[10] J. Weiner, A. Dobson, S. Maxwell, K. Coleman, B. Starfield, and G. Anderson. Risk-adjusted medicare capitation rates using ambulatory and inpatient diagnoses. *Health care financing review*, 17:77, 1996.

[11] P. Wynand, V. De Ven, and R. Ellis. Risk adjustment in competitive health plan markets. *Handbook of health economics*, 1:755–845, 2000.