No.27
MARZO DE 2017

Documentos CEDE

ISSN 1657-7191 Edición electrónica.

Risk Adjustment Revisited using Machine Learning Techniques

Alvaro J. Riascos
Mauricio Romero
Natalia Serna

CEDE
Centro de Estudios sobre Desarrollo Económico

Universidad de los Andes
Facultad de Economía

Serie Documentos Cede, 2017-27
ISSN 1657-7191 Edición electrónica.
Marzo de 2017

Risk Adjustment Revisited using Machine Learning Techniques*

Álvaro J. Riascos† Mauricio Romero‡ Natalia Serna§

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.

JEL Classification: I11, I13, I18, C45, C55

Keywords: risk adjustment; Diagnostic Related Groups; risk selection; machine learning

*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.

†Quantil — Matemáticas Aplicadas and Faculty of Economics, Los Andes University. e-mail: alvaro.riascos@quantil.com.co

‡University of California - San Diego. e-mail: mtromero@ucsd.edu

§Quantil — Matemáticas Aplicadas. e-mail: natalia.serna@quantil.com.co

Revisión del Ajuste de Riesgo con Técnicas de Aprendizaje de Máquinas*

Álvaro J. Riascos† Mauricio Romero‡ Natalia Serna§

Resumen

El ajuste de riesgo es un componente esencial en el diseño de la política del sector de la salud. El ajuste de riesgo define los pagos de capitación que se le hacen a las aseguradoras en salud y determina el riesgo de insolvencia de las mismas. En este estudio comparamos la fórmula usada actualmente por el Ministerio de Salud y Protección Social para determinar el ajuste de riesgo, contra especificaciones alternativas que controlan por factores de riesgo adicionales. Mostramos que la fórmula actual, la cual ajusta los pagos solamente a factores de riesgo demográficos, predice tan solo el 30 % del gasto en el quintil superior de la distribución del gasto. Nuestros resultados muestran que incorporar indicadores de 29 enfermedades de larga duración en un modelo lineal, como el que usa el gobierno, mejora su capacidad predictiva considerablemente. Finalmente, contribuimos a la literatura de ajuste de riesgo a través de la estimación de modelos de aprendizaje de máquinas y mostramos que modelos no paramétricos (e.g., boosting de árboles) tienen un mejor desempeño que los modelos lineales, incluso cuando se ajustan sobre un conjunto de regresores más pequeño.

Código JEL: I11, I13, I18, C45, C55

Palabras clave: ajuste de riesgo; Códigos Relacionados de Diagnóstico; selección de riesgo; aprendizaje de máquinas

*Agradecemos el apoyo financiero del Ministerio de Salud y Protección Social. También agradecemos los valiosos comentarios y sugerencias de 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, los participantes del taller "TAUB Center for Social Policy Studies" en Israel (financiado por el Banco Mundial), el CEDE de la Universidad de los Andes y un equipo médico en la Fundación Valle del Lili en Cali, sobre este artículo y sus versiones anteriores. Las consideraciones legales usuales aplican.

†Quantil — Matemáticas Aplicadas y Facultad de Economía, Universidad de los Andes. e-mail: alvaro.riascos@quantil.com.co

‡Universidad de California - San Diego. e-mail: mtromero@ucsd.edu

§Quantil — Matemáticas Aplicadas. e-mail: natalia.serna@quantil.com.co

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. This market structure has 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" (Wynand et al., 2000). Monthly risk premium fees are collected from all enrollees with formal employment and then redistributed in form of capitation payments to every health insurer in the market. The redistribution is based on the specific risk profile of the insurer's population of enrollees.

There are two noteworthy features of the capitation payments. First, payments are done ex-ante, i.e, before enrollees claim services. Second, payments are adjusted for each enrollee's health risk. In the absence of risk-adjusted payments, health insurers have perverse incentives to engage in risk selection. For example, they would discourage the enrollment of sick or high-risk individuals through strenuous formalities, large waiting lines, or low service quality.

Although there is great uncertainty revolving annual health expenditures, a part of it is predictable by socio-demographic variables. For example, childbearing age women are costlier than men in the same age group, elders are costlier than teenagers, and people living in urban areas are costlier than those living in rural areas.[1]. 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 Colombia, risk-adjustment is based on the risk pools formed by unique combinations of gender, age group and enrollee's location. However, the socio-demographic characteristics of individuals only explain 2% of the variation in health expenditure.

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. These variables are not subject to manipulation by health insurers (they are easily verifiable by the Ministry of Health) and are available in the database that registers all claims made by the enrollees of the health system know as "Base

[1]The urban-rural divide in expenditure is driven both by demand (access) and supply

de Suficiencia".

We use machine learning to predict annual health expenditures. We compare the current government formula that adjusts payments only on socio-demographic variables against specifications that control for long-term diseases and indicators of health condition. We show models based on machine learning techniques predict more accurately the annual health expenditure of individuals in the upper quintile of the health expenditure distribution but tend to overestimate expenditures at the lower tail of the distribution, which is not necessarily problematic. We also show the predictive power of the current government formula can increase significantly after risk adjusting ex ante for the enrollees' morbidity distribution using indicators of long-term diseases.

This document is structured as follows: after this introduction, the second section describes the Colombian health sector. Section 3 presents the empirical framework of linear models and machine learning models. Section 4 describes our database and Section 5 analyzes estimation results. Finally, Section 6 highlights some conclusions and policy implications.

2 The Colombian health sector and the capitation payments

The Colombian health sector comprises a contributory and a subsidized system. All formal employees and their beneficiaries are enrolled to the former, while people with no sources of income and who are poor enough to qualify are enrolled to the latter. Of the 46 million people enrolled, 44% are in the contributory system and 56% in the subsidized system. Each month, enrollees to the contributory system pay a premium that is proportional to their income. Contributions are collected by the FOSYGA and then redistributed among health insurers through the capitation payment known as "Unidad de Pago por Capitación" (UPC). Contributions also finance the enrollment of people in the subsidized system. The UPC is paid to every insurer at the beginning of the year. In exchange, insurance companies must provide all health services listed in the benefits package that individuals claim.[2]

The risk-adjustment mechanism helps contain the system's expenditures and guarantees the financial solvency of health insurers. Currently the mechanism consists of a linear regression of annual health expenditure on sociodemographic risk factors including gender, age groups, location, and their two-way interactions. Age groups are defined by the Ministry of Health[3] and location is a categorization of the municipality of residence in three areas: urban (metropolitan areas), normal (municipalities surrounding metropolitan areas)

[2]Enrollees can have access to services outside of the benefits package by subscribing to prepaid medicine plans for a higher premium.

[3]The age groups are: 0, 1-4, 5-18, 19-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, and more then 75

and special (peripheral municipalities).

Let Y_{it} be the annual health expenditure of enrolle i during year t calculated as the sum of the cost of all services claimed during the year, and let d_{it} be individual i's number of enrolled days during year t. The annual health expenditure of enrollee i is:

$$g_{it} = 360 \times \frac{Y_{it}}{d_{it}} \tag{1}$$

Equation (2) represents the current government formula for risk-adjustment. This model is estimated with the claims data of year t to predict annual health expenditure for year $t+2$. Thus, there is a two-year lag between the gathering of claims data and the application of payments.

$$\log(g_i) = \beta_0 + \sum_{j=1}^{K} \beta_j D_j + \varepsilon_i, \tag{2}$$

where D is a set of K dummy variables indicating the presence of certain demographic characteristics (gender, age group, and location) and ε_i is an enrollee specific random shock. The coefficients in the model are the additional expenditure incurred by a risk pool with the combination of variables denoted by D relative to the base category, or the category for which all dummy variables take the value of zero simultaneously. The base category has an UPC of:

$$UPC_0 = \hat{\beta}_0 \tag{3}$$

which is equivalent to the global UPC defined by the Ministry of Health:

$$UPC_t = 360 \times \frac{\sum_{i=1}^{N} Y_{it}}{\sum_{i=1}^{N} d_{it}} \tag{4}$$

And, the UPC for risk pool r is:

$$UPC_r = UPC_0 + \sum_{j=1}^{K} \hat{\beta}_j D_j \tag{5}$$

which is equivalent to:

$$UPC_{rt} = 360 \times \frac{\sum_{i \in r}^{N_r} Y_{it}}{\sum_{i \in r}^{N_r} d_{it}} \tag{6}$$

The government formula controls for socio-demographic risk factors which are strongly correlated to

health expenditure. However, there are marked differences between the UPC and the observed health expenditure. Aditionally, there is evidence of risk selection strategies by health insurers (Castano & Zambrano, 2006; Gómez-Suárez, 2007). To mitigate these problems, a more precise risk-adjustment mechanism is necessary.

2.1 Risk adjustment literature

In a study for the Colombian health sector, Riascos et al. (2017) propose a new way to calculate the UPC by including additional risk factors describing the morbidity distribution of the enrollee population.[4] 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. Capital requirements are the minimum capital insurers must hold to keep the insolvency probability under certain thresholds.

Other studies have shown that interactions between diagnoses and demographic variables improve the predictions of models for annual health expenditure. For instance, Weiner et al. (1996) 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. Their model estimated with 1991 claims data explains 6.3% of the variation in annual health expenditures in 1992. Hughes et al. (2004) also categorize ICD-9 diagnoses into groups of chronic diseases and design an index of severity of illness to explain annual health expenditures in the United States. They find a linear regression including dummy variables of chronic diseases, their two-way interactions, and the severity of illness index, estimated with 1992 claims data, explains 10.66% of the variation in 1993 annual health expenditures. Moreover, the ratio of total predicted expenditure over observed expenditure for the fifth quintile of the expenditure distribution is 1.021, in other words, the model predicts quite accurately the expenditure of the most expensive individuals in the system.

The literature has also been motivated by the search of a smaller yet predictive set of regressors. DeSalvo et al. (2009) find that conditioning annual health expenditures on age and a self-reported measure of well-being classifies individuals in percentiles of the expenditure distribution more accurately than a model based only on diagnoses. The area under the ROC curve for the authors' model is 5 percentage points greater than a model based only on diagnoses, but is 5 percentage points lower than a model that risk adjusts for demographics, diagnoses and enrollment status.

[4] Specifically, they use indicators of 29 long-term disease groups built from ICD-10 diagnoses.

More recently, machine learning techniques have incurred in the field of health risk adjustment. Buchner et al. (2015) 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 that only includes interactions between demographics as age, gender, and enrollment status. Li et al. (2013) 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 authors build 70 diagnoses categories using the ICD-9 diagnoses and also include a gender dummy variable and age group dummy variables. 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.

In this study, we compare the current government formula in Colombia against alternative specifications that control for (i) the enrollees' morbidity characterized by 29 long-term disease groups following Alfonso et al. (2013)[5], (ii) the severity of health condition using indicators of hospitalizations, consult with specialist and admission to the intensive care unit, and (iii) all possible two-way interactions between the demographic variables in the government's base formula.

Our main contribution to the literature of risk adjustment is the estimation of novel models based on machines learning techniques, whose application to Colombia's health sector is still incipient[6]. We compare the predictive power of neural networks, random forests, and boosted trees against linear models, using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2, and predictive ratio (total predicted expenditure over total observed expenditure) for the upper and lower quintiles of the observed expenditure distribution.

3 Empirical framework

3.1 Prediction of annual health expenditure

To predict the annual health expenditure of enrollees to the contributory system in Colombia, we use a panel of claims from the individuals who were active enrollees during 2010 and 2011. To build the models we take the following individual traits: demographic characteristics ($Demog$) such as gender, age group as

[5]See www.alvaroriascos.com/research/healtheconomics for more on the construction of these groups.

[6]Some applications include Riascos & Serna (2017) in the prediction of annual length-of-stay of patients enrolled to the contributory system.

defined by the Ministry of Health, location as the categorization of the municipality of residence in three payment areas, and their two-way interactions; indicators of long-term disease groups (Dx); and indicators of hospitalizations (H), consults with specialists (E) and admissions to the intensive care unit (U). The first empirical approach is the estimation of linear models through enrolled-days weighted least squares and the second is the estimation of models based on machine learning techniques. Each model will be estimated on the set of regressors listed below:

1. $UPC = Demog$
2. $UPC + Dx$
3. $UPC + Dx + H$
4. $UPC + Dx + H + E$
5. $UPC + Dx + H + E + U$
6. $UPC \times H \times E \times U + Dx$

The first set of regressors will make up the base model or the current government formula. ($\times$) indicates the variable can interact with the rest of included regressors and ($+$) that the variables are included in a linear additive fashion. The weighted least squares linear regression model is:

$$y_{i,2011} = \beta_0 + \sum_{k=1}^{K} \beta_k D_{ik,2010} + \varepsilon_i \tag{7}$$

where $y_{i,2011}$ is the annual health expenditure of enrollee i during 2011 weighted by her enrolled days in the year ($360 \times x_{i,2011}/d_{i,2011}$); $D_{i,2010}$ is a set of dummy variables indicating the presence of the traits listed above; and ε_i is an enrollee specific random shock that is normally distributed. The linear model is estimated through weighted least squares (WLS) such that:

$$y_{i,2011} = w_{i,2011} Y_{i,2011} \tag{8}$$

and,

$$w_{i,2011} = \frac{d_{i,2011}}{360} \tag{9}$$

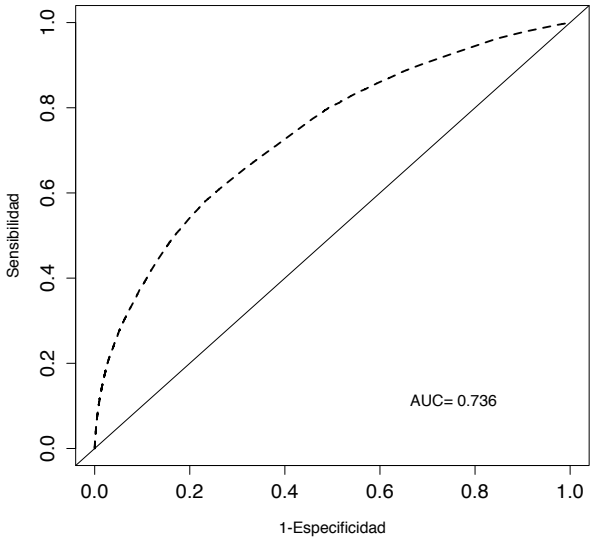
For the second approach we estimate 3 machine learning based models over different sets of regressors: *artificial neural networks* (ANN), *random forests* (RF), and *boosted trees* (GBM). In the case of neural

networks, we estimate a model with 3 layers and 5 neurons in the inner layer. We specify sigmoidal inner activation functions and a linear outer activation function. In order to use use sigmoidal functions in the inner layer we transform the dependent variable to a {0,1} scale. To do so, we divide the annual health expenditure by 10^{10} which is greater than the maximum observed expenditure, and then multiply predictions by this same value to return the variable to its original scale. In all machine learning based models we perform this same transformation of the annual health expenditure. In the training of neural networks we use the back propagation algorithm and find weights that minimize the RMSE in the training set.

In the random forest model we use the entire set of regressors for estimation $(Demog, H, E, U, Dx)$ because variable selection for recursive partitioning is done from a subset of M candidate variables chosen randomly from the full set of K variables, where $|M| = |K|/3$. Random forest models are estimated over 680 trees in a subsample of 40% of the training set due to computational issues. As in the random forest model, the boosted tress model is fitted over the entire set of regressors. In this case we also estimate 680 trees and iterate over a grid of values for the minimum number of observations in non-terminal nodes and for the contraction parameter of each tree grown in the sequence. In the boosted trees model we allow for three-way interactions between the variables.

In some specifications of the machine learning models we include an additional regressor: the probability of claiming a service. As shown in the descriptive section ahead, nearly 20% of enrollees to the contributory system do not claim any service during the year and, hence, their annual health expenditure is zero. This percentage of enrollees without expenditures suggest it is important to control for the selection bias of enrollees to health service demand. Individuals with zero health expenditure are the most healthy ones. To estimate the probability of claiming services we use a neural network model where the dependent variable takes the value of 1 if the enrollee has an annual health expenditure greater than 200,000 pesos and zero otherwise. This model has 3 layers and 5 neurons in the inner layer. Activation functions are sigmoidal in all cases. Figure (1) shows the out-of-sample ROC curve for the probability of claiming services:

Figure 1: ROC curve for the claiming of services



Negative predictions of machine learning models are truncated at zero. Overall, these models tend to overestimate the health expenditure in the lower percentiles of the expenditure distribution while they tend to underestimate the expenditure in the higher tail of the distribution. Overestimation of the lower tails is not problematic in the sense the average health expenditure of enrollees here is 2,000 pesos. On the contrary, underestimation of the upper tail is problematic since individuals represented in this portion of the distribution can cost more than 300 million pesos to the health system.

On a final exercise, we estimate machine learning based models over a subset of variables chosen with statistical criteria excluding the probability of claiming health services. Variable selection consists of fixing a threshold for the relative importance metric in the boosted trees model. The relative importance is a measure of how many times is a variable chosen to divide the characteristics space in two over the non-terminal nodes of the trees in the boosting algorithm. We fix a threshold in 0.5 such that variables with a relative importance greater than or equal to 0.5 are chosen for estimation. The final subset includes 21 variables which is 90% smaller than the set of variables used for estimation of the current government formula. With this variable selection procedure we show the predictive power of machine learning algorithms outperform that of the linear models even if estimated on significantly less information.

3.2 Out-of-sample fitting metrics

We compare the models using the following metrics:

- Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} w_i(y_i - \hat{y}_i)^2}$$

- Mean Absolute Error:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} w_i|y_i - \hat{y}_i|$$

- Predictive ratio for annualized expenditure (PR - annualized):

$$PR = \frac{\sum_{i=1}^{N} \hat{y}_i}{\sum_{i=1}^{N} y_i}$$

- Predictive ratio for non annualized expenditure (PR - non annualized):

$$PR = \frac{\sum_{i=1}^{N} w_i\hat{y}_i}{\sum_{i=1}^{N} w_i y_i}$$

- R-squared (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N} w_i(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} w_i(y_i - \overline{y}_i)^2}$$

The predictive ratios are calculated in the full sample as well as conditional on the first and last quintiles of the observed expenditure distribution. The best model will be the one that achieves the lowest RMSE and MAE, the highest R-squared, and the closest to 1 predictive ratio over the full sample and the last quintile of the expenditure distribution.

4 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. Each year's database is a cross-section of claims that can be merged together to create a panel of enrollees. 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 ICD10 diagnosis. We categorize the age variable in the twelve groups defined by the Ministry of Health (0, 1-4, 5-18, 19-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, older than 75) and 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 following Alfonso et al. (2013).

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: the UPC in the 2010 cross section is 565,563 pesos while in the panel is 732,499 pesos in the same year. For 2011 the numbers are 568,417 pesos and 695,776 pesos respectively.

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
UPC 2010	565,563		732,499
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 the percentage of enrollees who claim at least one health service every year.

Individuals who are active enrollees both years have higher health expenditures than those who are enrolled only one year. In fact, the percentage of enrollees with long-term diseases during 2010 that upturn in 2011 is lower in the panel (7.86%) than in the 2010 cross-section (9.85%). However, the overall percentage of individuals with long-term diseases is stable both in the panel and the cross-sections (26%). During 2010, in the cross-section, 64% of enrollees claimed at least one health service as well as during the 2011 cross-section. In the intersection of enrollees for both years, the percentage of individuals claiming services in 2010 is 30 percentage points larger than in the cross-section and 20 percentage points larger in 2011, which suggests there is a selection bias of sicker individuals to be active enrollees both years.

For prediction of annual health expenditures and to avoid overfitting problems, we build two mutually ex-

clusive 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. Table 2 shows the variable means in each dataset and the third column reports if differences in means between both datasets is significant. Overall, the table shows there are nonsignificant differences between the training set and the test set. Cardiovascular diseases such as hypertension are the most prevalent followed by cervical cancer and diabetes. The majority of enrollees live in urban municipalities and an average of 38% of them is remitted to the specialist. Also, 22% of enrollees has been hospitalized during 2010 and 0.3% is admitted to the intensive care unit. In the first panel of the table where we report cost measures, there is evidence that there are no significant differences in the enrolled-days weighted annual health expenditure during 2011. The mean reported in the table corresponds to the following calculation: $360 \times \frac{1}{N} \sum_{i=1}^{N} x_i/d_i$. There are no significant differences either in the unweighted annual health expenditure $\frac{1}{N} \sum_{i=1}^{N} x_i$, nor in the number of enrolled days during 2011 between the training set and the test set.

Table 2: Variable means in the training set and the test set

		(1) Train	(2) Test	diff (1)-(2)	
Costs	2011(weighted)	756,422	754,046	2,376	
	2011(unweighted)	600,611	601,162	551	
Enrolled days	2011	310.8	311.0	0.2	
% of enrollees who claim services	2011	83.1	83.2	-0.1	
Demográphics	Male	44.34	44.43	-0.0009	
	Age 1-4	6.92	6.86	0.0006	
	Age 19-44	42.24	42.25	-0.0001	
	Age 70-74	2.26	2.25	0.00008	
	Age ¿ 75	3.41	3.35	0.0005	
	Urban	54.20	54.20	0.000003	
	Normal	43.17	43.18	-0.0002	
	Special	2.64	2.62	0.0002	
Long-term diseases	Genetic anomalies	1.67	1.65	0.0001	
	Arthritis	1.20	1.15	0.0005	**
	Pyogen arthritis	0.08	0.08	0.00001	
	Arthrosis	2.06	2.11	-0.0005	**
	Asthma	1.75	1.79	-0.0004	
	Autoimmune	0.40	0.40	-0.00001	
	Cervical cáncer	3.18	3.15	0.0003	
	Invasive cervical cáncer	0.06	0.06	-0.000002	
	Male genitalia cancer	0.35	0.36	-0.0002	
	Breast cancer	1.42	1.42	0.00008	
	Skin cancer	0.23	0.24	-0.00003	
	Digestive organ cancer	0.14	0.13	0.00004	
	Respiratory organ cancer	0.05	0.05	-0.00003	
	Other cancer	0.93	0.94	-0.0001	
	Other female genitalia cancer	0.17	0.18	-0.00007	
	Lymphatic tissue cancer	0.17	0.16	0.00006	
	Cancer therapy	0.01	0.01	0.00007	
	Diabetes	2.77	2.80	-0.0002	
	Hypertension	10.68	10.72	-0.0004	
	Other cardiovascular disease	5.40	5.45	-0.0004	
	Pulmonary disease	1.65	1.68	-0.0003	
	Chronic renal insufficiency	0.59	0.58	0.0001	
	Other renal insufficiency	0.15	0.14	0.0006	
	Other renal disease	0.30	0.31	-0.0002	
	Long-term renal disease	0.02	0.02	-0.00002	
	HIV-AIDS	0.22	0.24	-0.0002	
	Epilepsy	0.70	0.70	0.00003	
	Transplant	0.06	0.06	0.000006	
	Tuberculosis	0.20	0.20	-0.000002	
Other	Hospitalization	21.44	21.59	-0.0014	
	Specialist	38.11	38.15	-0.0004	
	ICU	0.34	0.33	0.00003	

Note: This table shows the variable means in the train and tests sets. The first panel shows the average health expenditure weighted by enrolled days ($360 \times \frac{1}{N}\sum_{i=1}^{N} x_i/d_i$) and the unweighted average. The second panel shows the number of enrolled days, the third the percentage of enrollees who claim at least one health service during 2011, the fourth presents the variable means for some demographic variables the fifth shows the proportion of enrollees with the indicated long-term disease, and the sixth shows the proportion of enrollees who have been hospitalized, who have visited the specialist and who have been admitted to the ICU during 2010. Column *diff (1)-(2)* shows the differences in means between the two datasets and whether the difference is significant. ***p<0.01, **p<0.05, *p<0.1

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

Table 3: 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. The RMSE and the MAE are reported in 2011 colombian pesos. Authors' calculations from the *Base de Suficiencia*.

WLS are the linear models estimated with weighted least squares, ANN are the neural networks, RF are the random forests, and GBM are the boosted trees models. Table 3 shows the RMSE, MAE R^2 and predictive ratios in the full sample. The first column shows the parameters with which the models were fitted. In the case of neural networks, the first number corresponds to the number of neurons in the inner layer and the second stands for the weight decay parameter as a variable passes from one inner layer to the outer layer. In the case of the boosted trees model, the first number corresponds to the number of trees, the second to the number of interactions between variables, the third is the contraction parameter of each subsequent tree, and the fourth stands for the minimum number of observations in non-terminal nodes. In the case of the random forest model the number reported corresponds to the number of trees.

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.

Table 4 show the out-of-sample fitting metrics for the lower quintile of the observed expenditure distribution in 2011. All the models highly overestimate the annual health expenditure of individuals in this part of the distribution. Linear models have the smallest RMSE and predictive ratios, for instance the model with all the set variables and their interactions achieves a RMSE of 552,219 pesos, a MAE of 319,218 pesos and an annualized predictive ratio of 267.2. In other words, this model predicts almost 270 times more expenditure in the first quintile of the expenditure distribution than what is actually realized. Machine learning based models do not seem to improve this metric. The GBM FS model achieves an annualized predictive ratio of 371.1 and a non annualized predictive ratio of 253.9

Table 4: Out-of-sample fitting measures in the first quintile of the expenditure distribution

Parameters	Model	RMSE	MAE	PR annualized	PR non annualized
	WLS UPC	585,387	423,863	344.0	285.9
	WLS UPC + Dx	544,839	344,251	285.0	232.5
	WLS UPC + Dx + H	549,384	331,398	275.4	223.8
	WLS UPC + Dx + H + E	541,922	314,979	272.6	212.6
	WLS UPC + Dx + H + E + U	560,868	317,613	261.6	214.4
	WLS UPC x H x E x U + Dx	552,219	319,218	267.2	215.6
5 + 0.05	ANN Demog + Dx (Two stages)	674,653	401,014	331.3	270.3
5 + 0.05	ANN Demog + Dx + H (Two stages	633,907	362,857	301.3	244.7
5 + 0.05	ANN FS	646,313	439,889	369.8	296.7
680	RF Demog + Dx + H + E + U (Two stages)	792,871	347,280	291.2	234.5
680	RF FS	658,254	361,954	304.7	244.4
680 + 3 + 0.1 + 50	GBM Demog + Dx + H + E + U (Two stages)	630,485	364,222	307.6	245.9
680 + 3 + 0.1 + 50	GBM FS	652,681	376,233	317.1	253.9

Note: This table shows the RMSE, MAE and annualized and non-annualized predictive ratios in the lower quintile of the observed annual health expenditure distribution for 2011. The models and their parameters as the same as in table (3). Authors' calculations from the *Base de Suficiencia*.

Table 5 shows the out-of-sample fitting metrics in the highest quintile of the observed expenditure distribution. Recall underestimation of health expenditure in this portion of the expenditure distribution is problematic because it implies insufficient coverage of the health and financial risks attainable to the insurers' economic activity. 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 in this part of the distribution which represents an improve of 5 percentage points relative to the government's formula. In general, the predictive ratios for the non annualized health expenditure of machine learning based model oscillates between 50 and 53%, and for the annualized health expenditure between 42 and 46%. In any case, 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. This suggests morbidity is a important risk factor and, hence, risk adjusting the UPC for any indicator of the enrollees' morbidity would significantly improve the description of the insurers' risk profile.

Table 5: Out-of-sample fitting metrics in the upper quintile of the expenditure distribution

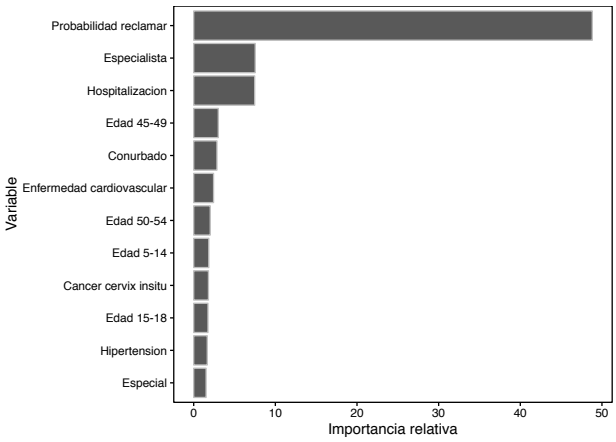
Parameters	Model	RMSE	MAE	PR annualized	PR non annualized
	WLS UPC	7,749,235	1,920,486	0.291	0.335
	WLS UPC + Dx	7,580,659	1,983,269	0.367	0.426
	WLS UPC + Dx + H	7,569,594	1,974,414	0.380	0.440
	WLS UPC + Dx + H + E	7,565,491	1,963,559	0.388	0.450
	WLS UPC + Dx + H + E + U	7,555,009	1,965,934	0.390	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 (3). Authors' calculations from the *Base de Suficiencia*.

Figure 2 presents the variable relative importance in the GBM FS model. Since this measures is an average of how many times is a variable used for the recursive partitioning of the trees, it should not be interpreted in terms of the direction of their effect on annual health expenditures rather than by how predictive they are. 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.

Figure 2: Variable relative importance in the GBM FS model



5.1 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 will tend to enroll only those individuals whose health expenditure is low but whose prediction, on which per capita payments are based, is high. Risk selection is inefficient from the societal perspective because it disincentives insurers to enroll individuals for which the health expenditure is underestimated and who are, at the same time, the sickest population. 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 6 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: (i) if the percentage of enrollees and the percentage of profit with respect to income is similar and (ii) if the absolute level of profit is low compared to the rest of the models. In the first case, notice that if the percentage of enrollees is less than the percentage of profit they generate, this would mean the insurer can achieve a large proportion of its profit by enrolling only a small portion of the population of individuals leaving a large portion uninsured. In the second case, we want the overall level of profit due to overestimation of health expenditure to be low because it is directly related to the level of incentives to risk selection.

Table 6: 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.*

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.

6 Conclusions

In this study we compare the current risk adjustment formula by Colombia's Ministry of Health and Social Protection against alternative specifications that risk adjust for additional factors. Risk adjustment is vital in health policy design because it is one of the main causes of financial problems in Colombia's health sector. Risk adjustment consists of predicting the annual health expenditure of every enrollee based on income-neutral variables and then define the annual capitation payment based on this prediction. Capitation payments are the main source of income for health insurers and, therefore, underestimation of health expenditures for the riskier individuals can increase their insolvency risk. Since the definition of the risk adjustment formula, the government has been reluctant to the inclusion of more predictive factors. Although reluctancy is well founded as it can generate perverse incentives for "up-coding", for the provision of unnec-

essary health services, and for the unjustified billing to the FOSYGA, many health systems around the world (United States and Germany to name a few) are already risk adjusting capitation payments to indicators of the morbidity distribution such as ICD-9 diagnoses. We argue, first, that risk adjusting for morbidity ex ante using the 29 long-term diseases groups does not necessarily increase the risk of perverse incentives because this information is easily verifiable by the government and, second, that the ex post mechanism proposed by the Colombian government for risk adjusting high cost diseases is still insufficient to manage the system's health and financial risks. The liquidation of insurers like Cafesalud, Caprecom, SaludCoop, Humana Vivir, Salud Cóndor, and Solsalud, evidence the ex post and ex ante mechanisms as they are defined nowadays are not efficient.

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. Even though part of the annual health expenditure is unpredictable, the government's formula significantly underestimates the portion that is in fact predictable, compared to alternative specifications. For example, 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. Machine learning models maintain their predictive power even if fitted in a much smaller set of regressors. 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.

References

Alfonso, E., Riascos, A., & Romero, M. (2013). *The performance of risk adjustment models in Colombia competitive health insurance market.*

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

Castano, R., & Zambrano, A. (2006). Biased selection within the social health insurance market in colombia. *Health Policy*, *79*(2), 313-324.

DeSalvo, K. B., Jones, T. M., Peabody, J., McDonald, J., Fihn, S., Fan, V., ... Muntner, P. (2009). Health care expenditure prediction with a single item, self-rated health measure. *Medical care*, *47*(4), 440–447.

Gómez-Suárez, R. (2007). Cream-skimming and risk adjustment in colombian health insurance system: The public insurer case. *Archivos De Economá*.

Hughes, J. S., Averill, R. F., Eisenhandler, J., Goldfield, N. I., Muldoon, J., Neff, J. M., & Gay, J. C. (2004). Clinical risk groups (CRGs): a classification system for risk-adjusted capitation-based payment and health care management. *Medical care*, *42*(1), 81–90.

Li, L., Bagheri, S., Goote, H., Hasan, A., & Hazard, G. (2013). Risk adjustment of patient expenditures: A big data analytics approach. In *Big data, 2013 ieee international conference on* (pp. 12–14).

Riascos, A., & Serna, N. (2017, February). *Predicting annual length-of-stay and its impact on health costs: the case of the colombian health care system.*

Riascos, A., Serna, N., & Guerrero, R. (2017). Capital requirements of health insurers under different risk-adjusted capitation payments.

Weiner, J. P., Dobson, A., Maxwell, S. L., Coleman, K., et al. (1996). Risk-adjusted medicare capitation rates using ambulatory and inpatient diagnoses. *Health care financing review*, *17*(3), 77.

Wynand, P., De Ven, V., & Ellis, R. P. (2000). Risk adjustment in competitive health plan markets. *Handbook of health economics*, *1*, 755–845.