

Predicting Annual Length-Of-Stay and its Impact on Health Costs

The Case of the Colombian Health Care System

Álvaro J. Riascos
University of los Andes and Quantil
Bogotá, Colombia
ariascos@uniandes.edu.co

Natalia Serna
Quantil
Bogotá, Colombia
natalia.serna@quantil.com.co

ABSTRACT

Avoidable hospitalizations are a source of increased health expenditures in many health systems. Prolonged length-of-stay is costly for providers, insurers, and patients in the sense it is associated to higher health service consumption and to the development of endangering states during the hospital stay. In this article we use machine learning techniques to predict annual patient length-of-stay in Colombia's contributory health care system and measure its impact on health costs by estimating the potential cost savings of a hospitalization prevention program. Results from the predictive modeling show tree-based methods outperform linear approximations and achieve lower out-of-sample error rates compared to the winning model of the Heritage Health Prize. We also show a prevention program where patient intervention is decided upon the predictions of the model can achieve significant cost savings relative to the best uniform policy (i.e., intervene all patients or no intervention), for program efficacies greater than 40% and intervention costs per patient ranging between 100,000 and 700,000 pesos.

KEYWORDS

Machine learning, public policy, health policy, prevention, cost effectiveness

ACM Reference format:

Álvaro J. Riascos and Natalia Serna. 2017. Predicting Annual Length-Of-Stay and its Impact on Health Costs. In *Proceedings of KDD 2017 Conference, Halifax, Nova Scotia - Canada, August 2017 (KDD'17)*, 8 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Avoidable hospitalizations are a source of increased health expenditures in many health systems. Prolonged length-of-stay is costly for providers, insurers, and patients in the sense it

is associated to higher health service consumption and to the development of endangering states during the hospital stay. In the Colombian statutory health care system increasing health costs due to avoidable hospitalizations have raised many questions on whether insurers are implementing prevention programs and on whether such programs significantly impact health costs. In this context, prediction of patient length-of-stay (LOS) is an important tool for resource allocation and for improvement patient health outcomes. Hence, the purpose of this paper is twofold: first, predicting the annual length-of-stay of users in the contributory health care system in Colombia and, second, estimating the potential cost savings of a preventive program whose main input is the prediction of annual LOS.

Most of the literature engaged in predicting LOS is developed from the providers' perspective rather than the insurers' perspective. Many authors predict LOS using a sample of patients with specific acute conditions or physiological traits that are often unobserved by the insurer. For example, [4] focus on individuals with cerebrovascular accident, [10] on patients that enter the intensive care unit following a cardiac surgery, [5] on patients with renal failure, and [6] on patients with hip fracture. Our study differs from the previous in that we predict LOS using information that is symmetrical between insurers, providers, and the government, and we do not focus on users with particular health conditions but instead analyze a representative sample of individuals in the contributory system with heterogeneous demographic and morbidity characteristics. We also lack data regarding specific patient physiological traits and we extend the analysis to measure the potential cost savings of a prevention program where the intervention decision depends on the predicted patient LOS. With regard to the techniques for predicting annual LOS in the area of machine learning, we take similar approaches to the ones used by [9] and [11], which include boosted trees, random forests, artificial neural networks, etc.

The remainder of this paper is structured as follows: after this introduction, section II describes the Colombian statutory health care system, section III provides the empirical framework, section IV describes our database and the data preprocessing, section V presents the results of machine learning techniques, section 6 presents the impact of LOS on health costs, and section VII concludes.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
KDD'17, Halifax, Nova Scotia - Canada
© 2017 Copyright held by the owner/author(s). 978-x-xxxx-xxxx-x/YY/MM...\$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

2 THE COLOMBIAN STATUTORY HEALTH CARE SYSTEM

The Colombian statutory health care system consists of two regimes: contributory and subsidized. The first covers 44 percent of the population and the second the remaining 56 percent. Each regime has its own network of health insurers and health service providers, which are responsible for providing a predetermined benefits package to all enrollees, known as the “Plan Obligatorio de Salud” (POS). In the contributory regime, enrollees (formal employees and individual contractors) pay for health care services a compulsory monthly tariff proportional to their income, while the subsidized regime is fully funded by the government.

Contributions of enrollees in the contributory regime are collected by a government agency called FOSYGA. This agency redistributes contributions to insurers at the beginning of the each year using a risk-adjusted premium per enrollee known as the “Unidad de Pago por Capitación” (UPC). The capitation premium adjusts health risk to demographic variables such as age, gender, and municipality of residence while being income neutral. Each year, all services provided must be reported to the FOSYGA in order to calculate the UPC for the next period. Our empirical analysis is based on this database of services in the contributory regime.

At the same time, insurers and health service providers negotiate bilateral contracts from a fixed menu of contract types defined by the law. This menu of contracts defines the type of payment from the insurer to the provider for attending its population of enrollees, but additional arrangements between the parties are not observed. Forms of payment include capitation and fee-for-service, which distribute risk and incentives in opposite ways between insurers and providers. The insurer bears all financial and health risks when subscribing fee-for-service contracts, while providers bear all risks in capitation contracts.

In the context of increasing health costs and risks stemming from the UPC and bilateral contracts, evaluating promotion and prevention programs is important to identify sources of cost savings. One of them are hospitalizations. As suggested in [1], health care systems that rely on hospitalizations for early patient treatment are more expensive than those that use hospitalizations as a last resource. The Colombian health care system resembles to the former which is why it is facing problems such as hospital bed shortage and financial instability, all of which worsens patient health outcomes.

During 2011, for every 100,000 enrollees, there were 3,500 hospitalizations. The frequency of hospitalizations is greater in pediatric units (19,983 for every 100,000 enrollees less than one year old and 8,117 for every 100,000 enrollees between 1 and 4 years old) ([1]) and some diagnosis-related groups such as acute respiratory infection (31.73 percent of hospitalizations in children less than 5 years old during 2009 to 2012 were due to respiratory infections like pneumonia or acute bronchiolitis) and acute diarrheal disease (8.76 percent of hospitalizations in children less than 5 years old during 2009 to 2012 were associated to gastrointestinal diseases) ([8]).

Predicting annual patient length-of-stay is, therefore, an important tool for resource allocation and cost administration in hospitals and health insurers. Identifying the factors that increase the average patient LOS enables hospitals and insurers to engage in early interventions and prevention programs to mitigate hospitalizations.

3 EMPIRICAL FRAMEWORK

Our empirical framework is two-fold as well: first, we use machine learning techniques to predict annual patient LOS and, second, we model a decision rule that uses the predictions of the first stage to indicate whether a patient should be intervened to avoid her hospitalization and then measure the potential cost savings of such prevention program relative to several base scenarios.

Predicting patient LOS has motivated part of the literature of big data and machine learning in health care. It is usually done through regression methods by transforming the outcome variable to a logarithmic scale as $\ln(LOS + 1)$. Authors like [7] and [12] argue the distribution of patient LOS is usually lognormal and, therefore, linear regressions can be used to explain LOS conditional on patient and hospital characteristics. During 2013, an alliance of service providers in the United States launched the Heritage Health Prize, a competition to predict more accurately annual days in hospital based on claims data of previous years. The outcome variable was also measured in logarithmic scale. Most participants showed machine learning techniques outperformed ordinary linear regressions. In particular, ensemble methods proved to be the best models. Milestone winners used ensembles consisting of linear combinations of boosted trees models, random forests, artificial neural networks and linear regressions, restricting the sum of coefficients to 1 and truncating negative predictions. Models were compared and evaluated using the Root Mean Squared Error (RMSE). The winning team achieved a RMSE of 0.4438 which is, nonetheless, 2.5 times the average log LOS of the third year of data.

To predict patient LOS in year t with information of previous years, we use a panel database of claims reported to the FOSYGA during 2009 to 2011 called “Base de Suficiencia”. We are interested in a regression task as the one proposed in the Heritage Health Prize and use different machine learning techniques for this first objective: boosted trees (GMB), random forests (RF), artificial neural networks (ANN), linear regressions (OLS), and ensemble techniques (ENS).

For the second objective, we find risk factors and model a decision rule that indicates whether a patient should be intervened to prevent her LOS next year. Let \hat{y}_i be the prediction of $\ln(LOS + 1)$ for patient i . We transform the predictions into the probability of being hospitalized next year by estimating the logit model of equation (1):

$$p_i = Prob[y_i = 1] = \frac{e^{\beta_0 + \beta_1 \hat{y}_i}}{1 + e^{\beta_0 + \beta_1 \hat{y}_i}} \quad (1)$$

where,

Table 1: Descriptive statistics of weighted annual hospitalization costs

Gender	Location	Age group	Mean cost
Female	Urban	Age 0	1,124,958
Male	Urban	Age 0	1,119,639
Female	Urban	Age 19-44	957,419
Male	Urban	Age 19-44	1,214,199
Female	Urban	Age 70-74	3,605,493
Male	Urban	Age 70-74	4,958,471

This table shows the mean annual hospitalization cost for certain risk pools weighted by the number of days enrolled. Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

$$y_i = \begin{cases} 1 & \text{if } LOS > 0 \\ 0 & \text{if } LOS = 0 \end{cases} \quad (2)$$

and LOS is the observed annual length-of-stay.

Now, suppose each insurer in the system undergoes a prevention program with an efficacy of α and a cost per patient of f . α can be interpreted as the reduction in the probability of hospitalization next year. Let g be a risk pool characterized by a unique combination of gender, location, and age group. These combinations make up a total of 72 risk pools. Let $X_g = \sum_{i \in g} x_i$ be the annual health cost of patients with $LOS > 0$ in risk pool g calculated as the sum of the cost of all claimed services during a year and let $D_g = \sum_{i \in g} d_i$ be the sum of the number of days every patient with $LOS > 0$ in risk pool g has been enrolled to the health system. The annual cost of hospitalizations for patients in risk pool g is:

$$c_g = 360 \times \frac{X_g}{D_g} \quad (3)$$

Table (1) shows the mean annual hospitalization cost for some risk pools. Annual hospitalization costs increase with age and, overall, are U-shaped. Costs decrease from newborns to people aged 15-18, and then increase monotonically for people aged 19-45 and people older than 75 years. Following [3] the expected cost of hospitalization for patient i is:

$$C_0 = \hat{p}_i c_g \quad (4)$$

If insurers undergo the prevention program for this user, then expected health costs would be:

$$C_1 = \alpha \hat{p}_i c_g + f \quad (5)$$

Thus, the patient must be intervened if $\pi(\hat{p}_i | \alpha, f) = C_1 - C_0 \geq 0$ or if:

$$\hat{p}_i \geq \frac{f}{\alpha c_g} \quad (6)$$

To measure the incremental cost-effectiveness of the prevention program, we compare the costs of the program where intervention is decided upon the inequality in expression (6) with two scenarios: the no-intervention policy and the best uniform policy given α and f .

In the first case, the incremental cost-effectiveness of the prevention program compared to the no-intervention case is:

$$\begin{aligned} CE_i &= \sum_{i=1}^N \mathbb{I} \left(\hat{p}_i \geq \frac{f}{\alpha c_g} \right) \pi(\hat{p}_i | \alpha, f) \\ &= \sum_{i=1}^N \mathbb{I} \left(\hat{p}_i \geq \frac{f}{\alpha c_g} \right) (C_0(\hat{p}_i) - C_1(\hat{p}_i)) \end{aligned} \quad (7)$$

In the second case, let p_u^* be the best uniform policy such that:

$$p_u^* = \min [C_0(\hat{p}_i) ; C_1(\hat{p}_i)] \quad \forall \hat{p}_i \quad (8)$$

The incremental cost-effectiveness per user between the intervention based on inequality (6) and the best uniform policy would be:

$$\begin{aligned} CE_i &= \max \left(\mathbb{I} \left(\hat{p}_i \geq \frac{f}{\alpha c_g} \right) (C_0(\hat{p}_i) - C_1(\hat{p}_i)) ; \right. \\ &\quad \left. \mathbb{I} \left(\hat{p}_i < \frac{f}{\alpha c_g} \right) (C_1(\hat{p}_i) - C_0(\hat{p}_i)) \right) \end{aligned} \quad (9)$$

Hence, total cost savings are:

$$CE = \sum_{i=1}^N CE_i \quad (10)$$

4 THE DATA

To predict patient LOS and estimate its impact on health costs, we have the yearly claims of a sample of 5.7 million enrollees in the contributory system during 2009 to 2011. The sample was built by the Ministry of Health and Social Protection, focusing on individuals who claim at least one service per year and do not change their insurer company during the time span. For ease of computation, we choose randomly 1 million enrollees and their associated claims.

Information per sampled individual includes: insurer to which she is enrolled, services she demands (claims) identified with a procedures code (CUPS by its Spanish acronym)¹, provider ID, cost per service, date of provision, diagnosis per service identified with the International Classification of Diseases (ICD) Codes (10th version), length-of-stay per claim, age, gender, and municipality of residence.

The municipality of residence is categorized as metropolitan, normal, or special following the definition of payment geographic areas of the National Administrative Department of Statistics (DANE). The first definition integrates metropolitan areas and its adjacent municipalities, the second includes small municipalities around metropolitan areas, and the third includes peripheral municipalities. Age is also categorized in 12 groups according to the Ministry of Health and Social Protection: 0, 1-4, 5-14, 15-18, 19-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, and 75 or older. Finally, ICD 10 codes are categorized in 29 long-term diseases proposed by [2].²

¹CUPS stands for "Código Único de Procedimientos" and is a dictionary of all services, procedures, and drugs included in the colombian benefits package.

²For more details on the construction of these long-term disease groups see www.alvaroriascos.com/research/healthconomics

Table 2: Descriptive statistics in the train and test sets

Variable	Train		Test		diff
	Mean	sd	Mean	sd	
Dependent variable					
LOS t	1.891	8.387	1.894	8.346	0.811
Demographics					
Male	0.445	0.497	0.446	0.497	0.432
Age 0	0.034	0.180	0.034	0.180	0.690
Age 1-4	0.054	0.225	0.054	0.226	0.388
Age 5-14	0.103	0.305	0.104	0.305	0.378
Age 15-18	0.020	0.138	0.020	0.139	0.570
Age 19-44	0.403	0.491	0.402	0.490	0.054
Age 45-49	0.082	0.275	0.082	0.275	0.745
Age 50-54	0.069	0.254	0.070	0.255	0.084
Age 55-59	0.060	0.238	0.060	0.237	0.487
Age 60-64	0.052	0.221	0.052	0.222	0.175
Age 65-69	0.041	0.199	0.041	0.199	0.687
Age 70-74	0.033	0.178	0.033	0.178	0.541
Age >75	0.048	0.214	0.048	0.214	0.985
Urban location	0.535	0.499	0.535	0.499	0.633
Normal location	0.438	0.496	0.438	0.496	0.550
Special location	0.027	0.161	0.026	0.161	0.715
Claims' characteristics					
Average cost	29,706.1	194,898.3	30,106.1	222,212.1	0.177
Average LOS $t-1$	3.369	6.352	3.368	6.356	0.871
St. Dev. cost	58,556.0	292,593.7	58,462.1	285,711.2	0.819
St. Dev. LOS	5.620	18.007	5.613	19.389	0.804
LOS $t-1$	19.006	26.772	19.024	26.875	0.639
LOS $t-1 > 30$	0.217	0.412	0.217	0.412	0.837
Max LOS	0.707	3.589	0.708	3.597	0.802
Second max LOS	0.150	1.333	0.149	1.351	0.591
Hemograms	0.620	1.628	0.621	1.635	0.709
Pressure tests	0.006	0.210	0.006	0.174	0.714
CTs	0.080	0.432	0.079	0.435	0.934
Creatinine tests	0.469	1.410	0.472	1.417	0.146
Thyroid tests	0.220	0.744	0.221	0.746	0.679
ER services	2.382	6.001	2.383	6.083	0.855
Ambulatory services	25.617	37.849	25.625	37.705	0.873
Hospital services	2.664	18.161	2.668	18.030	0.872
Domiciliary services	0.127	6.955	0.140	7.452	0.209
Average contribution income	1,020,238.0	291,184.2	1,020,367.0	291,343.8	0.754
St. Dev. contribution income	1,075,115.0	394,921.6	1,075,271.0	395,142.5	0.780
Drugs	10.72	20.45	10.72	20.52	0.942
N	993,857		993,711		

This table shows the mean and standard deviation of some of the features in the train and test sets. Column "diff" shows the p-value of the test of differences in means between both datasets. Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

Since the data needs to be aggregated from claims-level to patient-level, we create the following features with information from $t - 2$ to $t - 1$: annual LOS, average LOS, maximum LOS, second maximum LOS, indicator of annual LOS greater than 30 days, standard deviation of the LOS, average cost, standard deviation of the cost, average income of enrollees in each insurer, standard deviation of the income in each insurer, indicators of the 10 costlier diagnoses in the sample, number of hemograms, pressure tests, CTs, creatinine tests, thyroid tests, ER services, ambulatory services, hospital services, domiciliary services, drug claims, and the number of different long-term diseases affecting each patient. We also create the number of claims per month and per day of week, indicators of long-term diseases, and interactions between indicators of hospital services, ER services, domiciliary services and ambulatory services. The dependent variable is the logarithm of patient length-of-stay during year t .

To avoid over fitted predictions in the train set because models are estimated on the relevant patterns and features of this sample, we build a test set with information of a new sample of 1 million individuals chosen randomly and mutually exclusive from the train set. Table (2) shows some descriptive statistics of both datasets and shows whether differences in variable means are significant between them.

For all features reported in the table, the train and test sets are not statistically different from each other at a 95 percent confidence level. The average length-of-stay during year t is 1.89 days. Of those who claim at least one health service, 22 percent remained more than 30 days in the hospital during years $t - 2$ and $t - 1$. On average users claim 30 services per year, of which 83 percent correspond to ambulatory services and an average of 10.72 are due to drugs and medications. The majority of individuals live in urban municipalities and earn around 1 million pesos.

Further preprocessing of the database consists of deleting observations with more than 360 days in hospital during year t , deleting observations with more than 720 days in hospital from $t - 2$ to $t - 1$, and dichotomizing all categorical variables.

5 RESULTS

For the prediction of patient annual LOS we use different machine learning techniques. In the case of neural networks, we set linear activation functions, one hidden layer, and estimate input weight parameters using a back-propagation algorithm. The number of neurons in the hidden layer (12) and the weight decay between layers (0.125) are fixed exogenously using repeated cross validation in a grid of values. For the boosted trees model, we use repeated cross validation to

Table 3: Coefficients of the linear ensemble

	<i>Dependent variable: $\ln(LOS + 1)$</i>
ANN	-0.058*** (0.003)
BT	0.246*** (0.004)
RF	0.857*** (0.004)
OLS	-0.047*** (0.002)
Constant	0.002* (0.001)
Observations	993,927
Residual Std. Error	0.559
F Statistic	291,939***

This table shows the coefficients of the linear ensemble of the Ordinary Least Squares (OLS), Artificial Neural Network (ANN), Random Forest (RF), and Boosted Trees (BT) predictions. Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01. Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

find the optimal parameters for the number of trees (8000), minimum observations in nodes (100), shrinkage (0.1), and interaction depth (2). In both models parameters are chosen to minimize the RMSE in the train set. For the random forest, we fix the number of trees to 7500. Finally, we use an ensemble method consisting of the linear combination of all the previous models, without any restriction on the sum of the coefficients. In all cases, negative predictions are truncated at zero and predictions above $\ln(360)$ are truncated at $\ln(360)$.

Table (3) shows the coefficients of the linear combination in the ensemble. Tree-based methods have a positive correlation with the final predictor while ANN and OLS have a negative correlation. Among the tree-based methods the random forest has the highest correlation while among the linear methods the ANN has the lowest.

Table (4) shows the out-of-sample Mean Absolute Error (MAE), RMSE and R-squared of different models calculated on the test set and table (5) presents some statistics of the distribution of patient LOS generated by each model versus the observed scenario in the test set. In terms of the MAE, the linear ensemble outperforms the rest of the models while the random forest seems to be the best predictor in terms of RMSE and R-squared, 53.5 percent of the variance in patient LOS is explained by this model. Overall, models fit the data well until the 75th percentile of the LOS distribution but prediction of higher percentiles is less accurate. The 25th and 75th percentiles of the linear ensemble prediction distribution are more similar to the corresponding percentiles of the observed distribution than that of other models. At the 25th percentile there is a difference of 0.008 days with respect to the observed distribution and of 0.452 days at the 75th percentile. However, the maximum LOS predicted by the linear ensemble is 48.8 days while the observed maximum is 360 days, which suggests the model significantly underestimates the upper tail of the distribution. On the contrary, the ANN overestimates LOS at lower percentiles and the difference between the maximum predicted LOS and the observed one is 13 days. Figure (1) shows the variation in RMSE for different percentiles of the observed LOS distribution. Tree-based

Table 4: Out-of-sample model fit

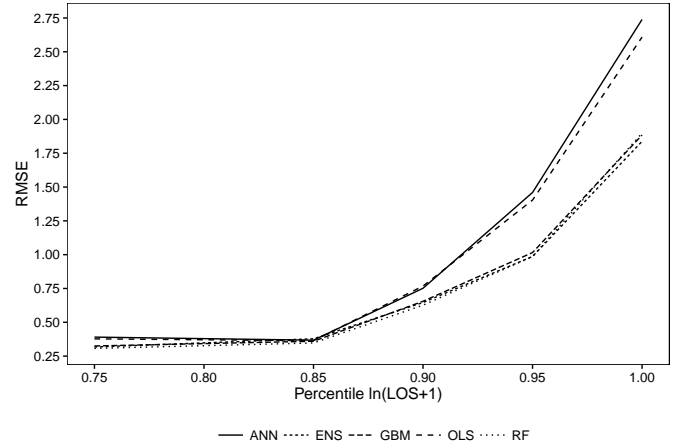
Model	MAE	RMSE	R-squared
OLS	0.4546	0.7502	0.1731
ANN	0.5032	0.7824	0.1006
RF	0.2634	0.5623	0.5354
BT	0.2721	0.5720	0.5192
ENS	0.2523	0.5609	0.5179

This table shows the out-of-sample MAE, RMSE, and R-squared of different models. OLS: Ordinary Least Squares, ANN: Artificial Neural Networks, RF: Random Forest, BT: Boosted Trees, ENS: Linear ensemble. Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

Table 5: Comparison of percentiles of patient LOS distribution

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Observed	0.333	0.828	0.000	0.000	0.000	5.889
LOS	0.338	0.318	0.000	0.128	0.482	5.886
ANN	0.370	0.247	0.000	0.229	0.753	5.851
RF	0.332	0.562	0.004	0.028	0.376	3.213
BT	0.335	0.580	0.000	0.022	0.390	5.886
ENS	0.334	0.605	0.000	0.008	0.373	3.909

This table shows the mean, standard deviation, minimum, 25th percentile, 75th percentile, and maximum of patient LOS distribution generated by each model and the observed scenario in the test set. OLS: Ordinary Least Squares, ANN: Artificial Neural Networks, RF: Random Forest, BT: Boosted Trees, ENS: Linear ensemble. Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

**Figure 1: Variation in the RMSE by percentiles of the LOS distribution**

Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

methods show a lower increase in RMSE at the right tail of the distribution compared to OLS and ANN models.

The MAE of the linear ensemble represents 75 percent of the average $\ln(LOS + 1)$ in the test set and the RMSE 168 percent. Compared to the winning team in the Heritage Health Prize (HHP), our best model outperforms the best model in the competition, which achieves a RMSE of 0.4438 or 249 percent of the average $\ln(LOS + 1)$ in year 3 data. However, there are several differences in the approach to LOS prediction and data preprocessing between the competition and the present study: first, in the HHP competition, training

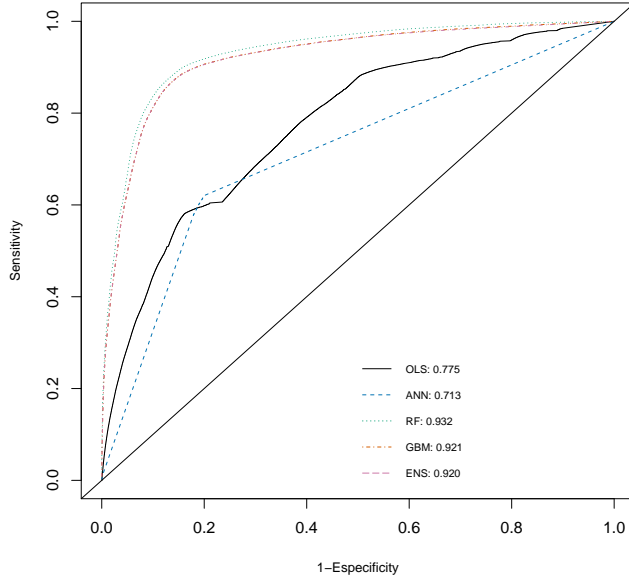


Figure 2: Prediction accuracy

Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

sets comprise only one year of data while we use at most two years; second, we lack information regarding the Charlson Index, lab counts, and drug counts included in the HHP; and third, we have a larger sample of patients compared to the HHP. Despite the differences, features built for the present study and machine learning techniques are similar to the ones used by Milestone winners.

A test of model accuracy for a classification task is the Receiver Operating Characteristics Curve (ROC). In figure (2) we build the ROC curve and calculate the area beneath it (AUC) for each model. Predicted proportions are calculated as the linear predictions divided into $\ln(360)$. The binary observed outcome takes the value of 1 if the annual LOS is greater than zero. The random forest has the highest AUC, followed by the boosted trees model and the linear ensemble: 0.932, 0.921, 0.920, respectively. Notice the linear regression outperforms the ANN and the reason is that we defined linear activation functions in the latter, so it basically amounts to estimating a linear regression with nonlinear parameters.

Figure (3) shows the most important predictors in the random forest model or risk factors as measured by the variation in node purity. Results shown in the figure should not be interpreted in terms of the direction of the effect but in terms of variable importance. The number of hospital services followed by the maximum LOS associated to a claim and the standard deviation of the insurer's average user income during the previous year are the most relevant predictors of next year's LOS. Comorbidities such as cardiovascular

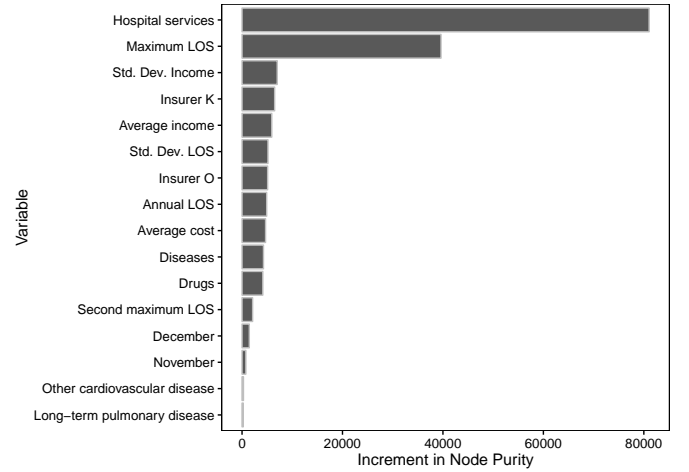


Figure 3: Risk factors in the random forest model
Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

diseases and long-term pulmonary diseases explain little of the variation in annual LOS, while fixed effects for insurers K and O are more significant.

6 POTENTIAL COST SAVINGS OF A PREVENTION PROGRAM

To measure potential cost savings of a prevention program where patient intervention is decided based on her predicted proportion as in equation (6), we estimate π for different combinations of program efficacy and intervention cost per patient. Figure (4), shows the contour plots of the cost savings per patient due to the decision rule based on predictions of the random forest versus the no intervention case, for $0 \leq \alpha \leq 1$ and $0 \leq f \leq 700,000$. The decision rule consists of assigning C_1 to patient i if the inequality in equation (6) is binding and C_0 otherwise. For every combination of efficacy and intervention cost, the decision rule based on the predictive model generates significant cost savings per patient. An intervention that costs 200,000 pesos per patient generates 50,000 pesos of cost savings per patient if program efficacy is greater than or equal to 30%. For intervention costs greater than that, the savings amount per patient can only be attainable with greater program efficacy compared to the no intervention case.

In figure (5) we show the contour plots of the cost savings per patient due to the decision rule versus the best uniform policy for each combination of α and f . The best uniform policy is the policy that generates the highest cost saving conditional on α and f between intervening all patients (assigning C_1 to all patients) and not intervening them (assigning C_0 to all patients). For intervention costs less than 50,000 pesos per patient and efficacies greater than 10%, it is cheaper to intervene all patients than to use the decision rule. The best uniform policy in this case would generate 20,000 pesos of cost savings per patient. However, when the

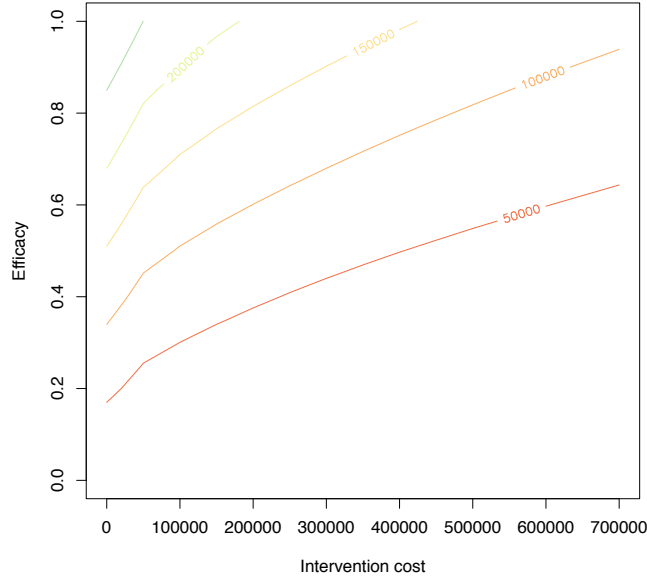


Figure 4: Cost savings over no policy

Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

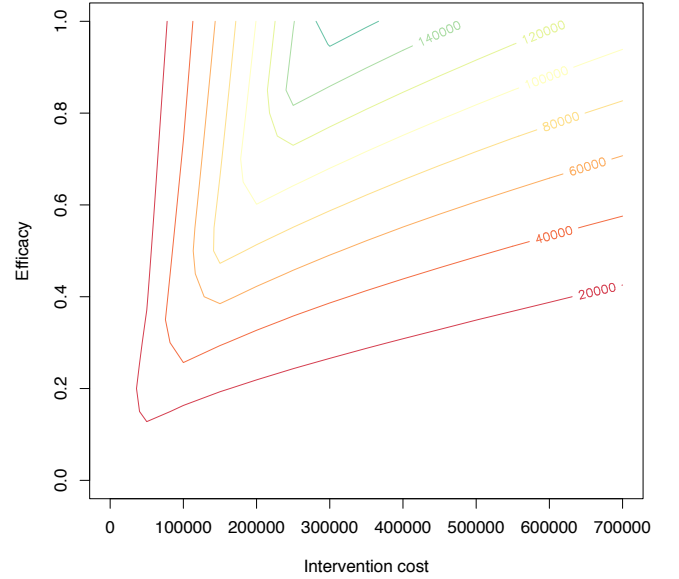


Figure 5: Cost savings over best uniform policy

Source: Base de Suficiencia, Ministry of Health and Social Protection. Authors' calculations.

intervention cost increases, benefits of using the decision rule are greater than the best uniform policy, and in any case greater than no intervention at all for program efficacies of more than 20%. If program efficacy falls from this threshold for any intervention cost then the program is not beneficial since it would be better to simply not intervene any patient.

These results suggest that for any intervention cost from 100,000 to 700,000 pesos per patient and efficacies greater than 40%, an automated decision rule based on predictive modeling is an important source of cost savings for every insurer in Colombia's contributory health care system. The decision rule and results presented in this section account for patient heterogeneity in two ways: first, the predictive model is trained on patient demographic and morbidity characteristics, and second annual health costs are allowed to vary per patient.

7 CONCLUSIONS

Hospitalizations are one of the main sources of health costs in Colombia's statutory health care system. Relying on hospitalizations for patient treatment increases the risk of bed shortages in hospitals and the risk of worse health outcomes in patients. Predicting annual patient length-of-stay is, therefore, an important tool for cost administration and resource allocation for insurers and providers. In this paper we use machine learning techniques to predict annual patient LOS based on their characteristics from previous years. We show tree-based models such as boosted trees, random forests and

an ensemble of their predictions outperform linear models such as artificial neural networks and linear regressions, in measures like the RMSE, MAE and R-squared. Relative to the average LOS in our sample, we achieve lower error rates compared to the results obtained by the winning team of the Heritage Health Prize, although there are differences in the way our data is processed. Compared to several international efforts in predicting annual LOS, our best model is highly predictive and suitable for every agent in Colombia's health care market, since it is trained with information that is symmetric between the providers, the insurers, and the government.

Using the predictions of the model we build a decision rule that suggests when to intervene a patient to prevent her hospitalization and achieve cost savings. To measure potential cost savings we compute the difference in total annual health costs between a prevention program whose intervention decision relies on the predictions of the model and the best uniform policy that consists of intervening all patients or not intervening them, conditional on the percentage of program efficacy and the intervention cost per patient. Results suggest Colombia's contributory health care system would achieve significant cost savings if insurers implemented prevention programs based on predictive modeling with efficacies of more than 40% and for any intervention cost between 100,000 and 700,000 pesos per patient.

This article contributes to the growing literature of machine learning in health care and provides evidence that is

crucial for the understanding of sources of increased health expenditures that are undermining Colombia's health market financial stability.

REFERENCES

- [1] ACEMI. 2013. *Cifras e indicadores del Sistema de Salud*. Technical Report. Asociación Colombiana de Empresas de Medicina Integral.
- [2] E. Alfonso, A. Riascos, and M. Romero. 2013. The performance of risk adjustment models in Colombia competitive health insurance market. (2013).
- [3] M. Bayati, M. Braverman, M. Gillam, K. Mack, G. Ruiz, M. Smith, and E. Horvitz. 2014. Data-Driven decisions for reducing readmissions for heart failure: general methodology and case study. *PLOS One* 9, 10 (2014), 1–9.
- [4] K. Chang, M. Tseng, H. Weng, Y. Lin, C. Liou, and T. Tan. 2002. Prediction of Length of Stay of First-Ever Ischemic Stroke. *Stroke* 33, 11 (2002), 2670–2674.
- [5] G. Chertow, E. Burdick, M. Honour, J. Bonventre, and D. Bates. 2005. Acute Kidney Injury, Mortality, Length of Stay, and Costos in Hospitalized Patients. *Journal of the American Society of Nephrology* 16, 11 (2005), 3365–3370.
- [6] J. Clague, E. Craddock, G. Andrew, M. Horan, and N. Pendleton. 2002. Predictors of outcome following hip fracture. Admission time predicts length of stay and in-hospital mortality. *International Journal of Care Injured* 33 (2002), 1–6.
- [7] P. Cleary, S. Pauker, and B. Mcneil. 1991. Variations in Length of Stay and Outcomes for Six Medical and Surgical Conditions in Massachusetts and California. *The Journal of the American Medical Association* 266, 1 (1991), 73–79.
- [8] MinSalud. 2014. *Análisis de Situación de Salud*. Technical Report. Ministerio de Salud y Protección Social.
- [9] H. Peyman, A. Maryam, A. Somayyeh, and S. Frahnaz. 2013. Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients. *Healthcare Informatics Research* 19, 2 (2013), 121–129.
- [10] J. Tu and M. Guerriere. 1993. Use of a Neural Network as a Predictive Instrument for Length of Stay in the Intensive Care Unit Following Cardiac Surgery. *Computers and biomedical research* 26, 3 (1993), 220–229.
- [11] P. Walsh, P. Cunningham, S. Rothenberg, S. O'Doherty, H. Hoey, and R. Healy. 2004. An artificial neural network ensemble to predict disposition and length of stay in children presenting with bronchiolitis. *European Journal of Emergency Medicine* 11, 5 (2004), 259–264.
- [12] C. Winslow, R. Bode, D. Felton, D. Chen, and P. Meyer. 2002. Impact of Respiratory Complications on Length of Stay and Hospital Costs in Acute Cervical Spine Injury. *CHEST Journal* 121, 5 (2002), 1548–1554.