
quantil

Two ongoing research papers of machine learning in health care

Álvaro Riascos and Natalia Serna

Predicting Length-Of-Stay and its Impact on Annual Health Costs in the Colombian Health Care System

Agenda

1. Introduction
2. Colombian health care system
3. Empirical framework
4. Data
5. Models and results

Introduction

- Increasing costs in the Colombian statutory health care system have raised many questions on whether insurers are actually setting forth promotion and prevention programs for their enrollees.
- One of the main sources of increased health expenditures are unnecessary and avoidable hospitalizations.
- Prediction of patient length-of-stay (LOS) will generate benefits in terms of resource allocation and patient health outcomes.
- Objective: predict the annual length-of-stay of users in the contributory health care system in Colombia and estimate the impact of prolonged LOS on health costs.

Colombian statutory health care system

- Contributory and Subsidized.
- Health care systems that rely on hospitalizations for early patient treatment are more expensive than those that use hospitalizations as a last resource (ACEMI).
- The Colombian health care system constantly faces a shortage of hospital beds that saturates emergency rooms (ER) and other levels of attention.
- During 2011, for every 100,000 enrollees, there were 3,500 hospitalizations.
- The frequency of hospitalizations is greater in pediatric units and some diagnosis-related groups.

Empirical framework

- The literature usually predicts $\log(LOS + 1)$.
- We want to compare ourselves with the winning team of the Heritage Health Prize (HHP).
- The competition required participants to predict annual days in hospital based on features of the claims during the previous year.
- Models were compared and evaluated using the Root Mean Squared Error (RMSE)
- RMSE: 0.4438, Average LOS: 0.4379 days, the average log LOS: 0.1782.

Empirical framework

- To predict patient LOS in year t with information of past years, we use a panel database of claims reported to the FOSYGA during 2009 to 2011.
- Impact evaluation for calculating the effect of an additional day in hospital over health costs incurred by the provider and the insurer.
- We argue this impact is highly nonlinear and non-parametrical.

Data

Table: Descriptive statistics in the train and test sets

	Train		Test		
Variable	Mean	sd	Mean	sd	diff
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

Data

Table: Descriptive statistics in the train and test sets

	Train		Test		
Variable	Mean	sd	Mean	sd	diff
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		

Models and results

Negative predictions are truncated at zero and predictions above ln(360) are truncated at ln(360).

Table: Coefficients of the linear ensemble

	Dependent variable: $\ln(LOS+1)$
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	$291{,}939^{***}$

Models and results

Table: 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

Table: 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

Models and results

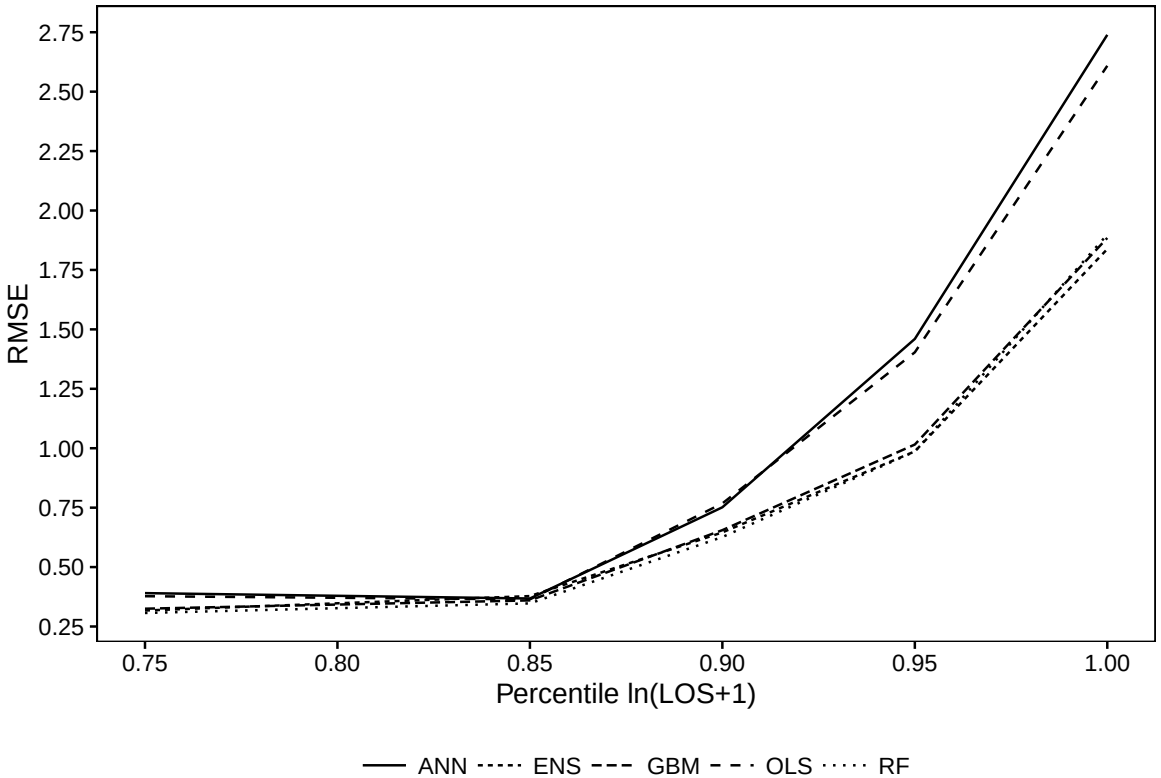


Figure: Variation in the RMSE by percentiles of the LOS distribution

Models and results

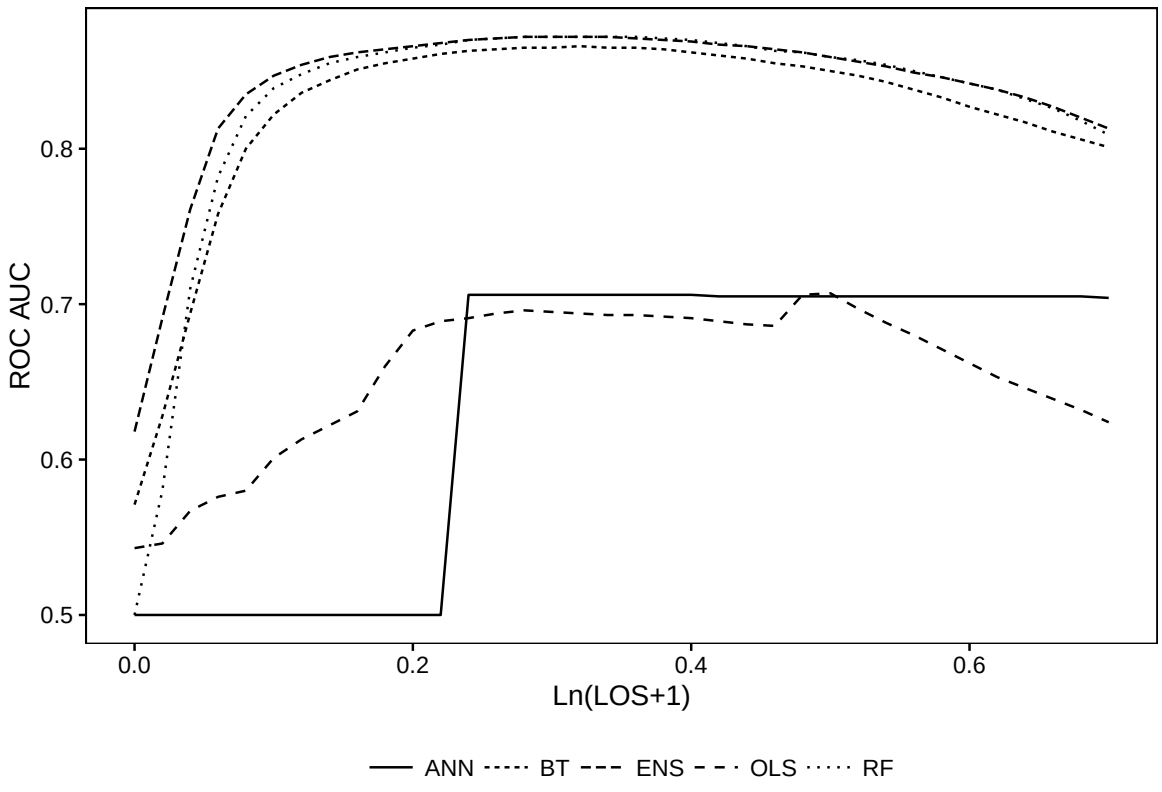


Figure: Prediction accuracy

Marginal effect of LOS on health costs

- We perform direct matching over the age group, location, insurer, and type of long-term disease.
- We compute the Average Treatment effect on the Treated (ATT) where the treatment in iteration i is defined as having $i+1$ days in hospital and the control group are patients with i days in hospital.

$$\tau = \frac{1}{N^T} \sum_{i \in T} Y_i^T - \frac{1}{N^T} \sum_{j \in M} \frac{1}{N_i^M} Y_j^M \tag{1}$$

Marginal effect of LOS on health costs

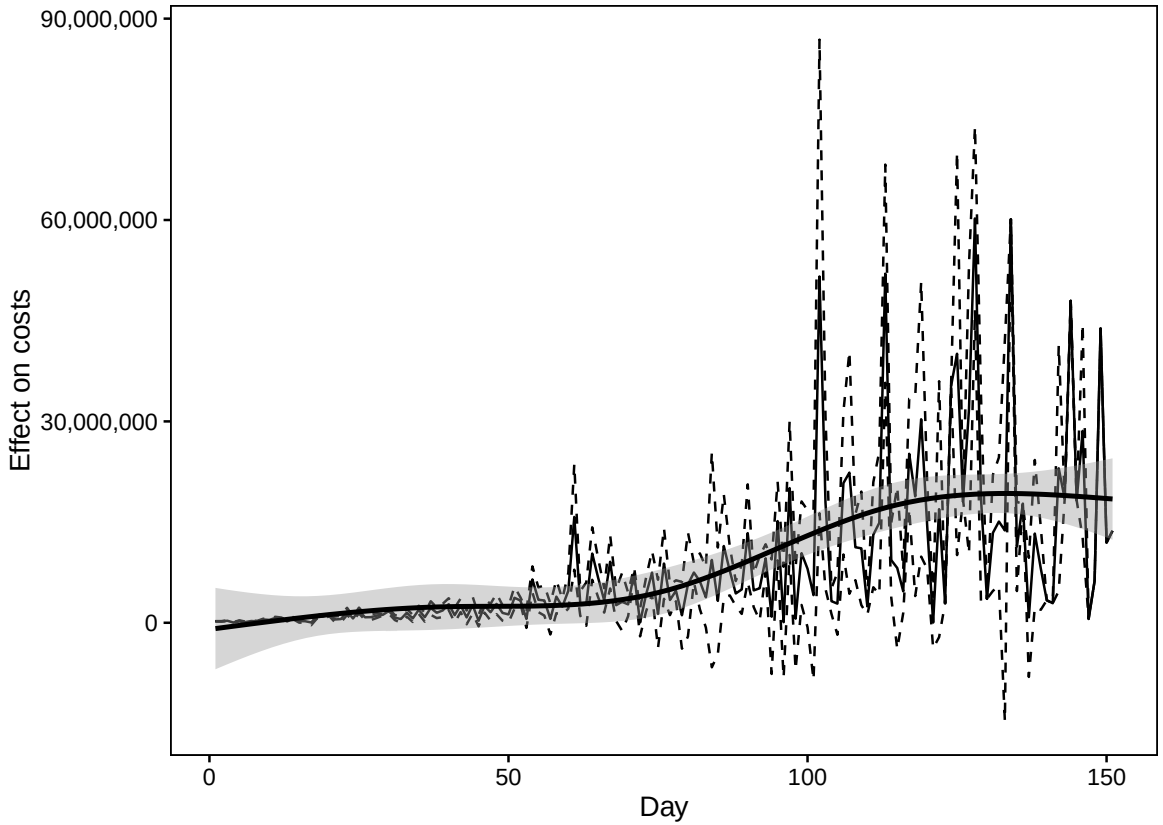


Figure: Marginal effect of an additional day in hospital on health costs in the test set

Marginal effect of LOS on health costs

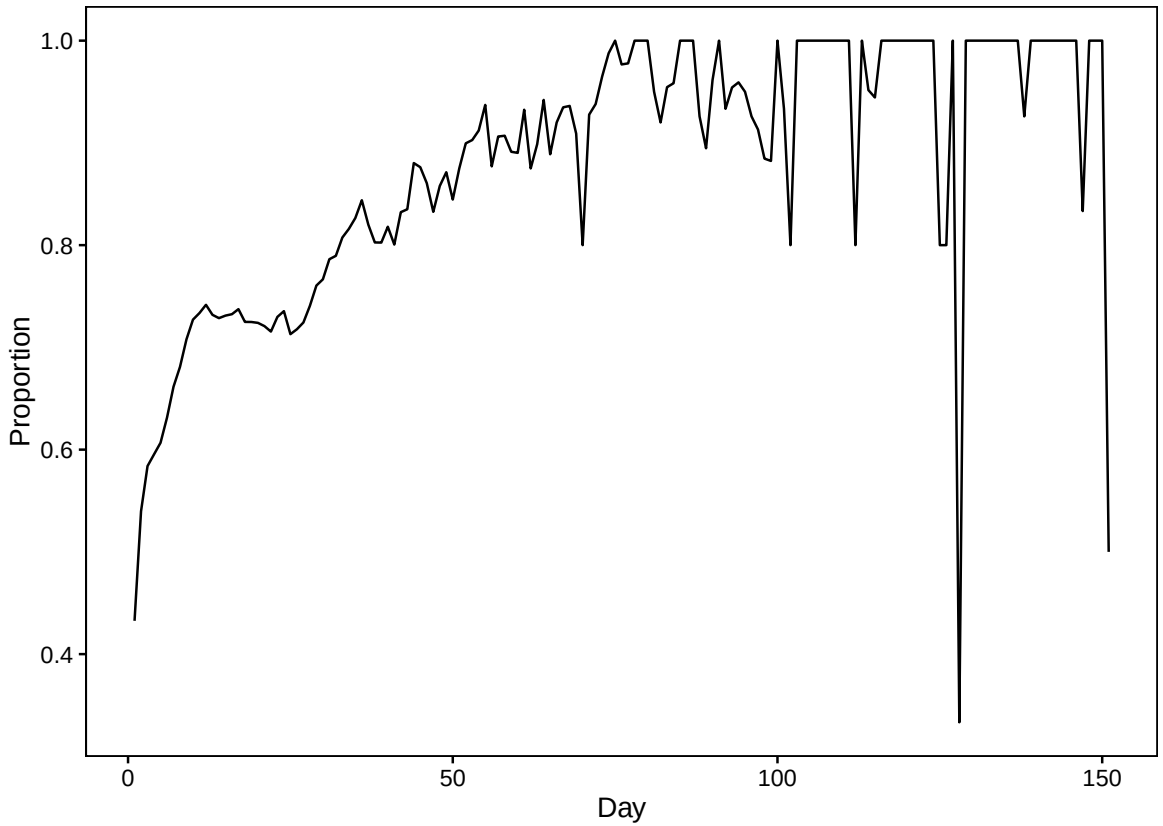


Figure: Proportion of patients with long-term diseases in the test matched sample

Marginal effect of LOS on health costs

Table: Statistics of the distribution of the marginal effect on annual health costs

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Day	76	44	1	38.5	113.5	151
Marginal effect	8,426,253	11,966,632	29,192	916,213	11,302,376	60,239,912
Lower bound	3,985,101	9,899,354	−14,675,523	168,581	4,416,080	60,115,671
Upper bound	12,867,405	16,346,489	94,962	2,197,998	18,029,431	86,869,312

The Impact of Number of Beds on the Risk of Readmission to the ICU: A Machine Learning Approach

Agenda

1. Introduction
2. Colombian health care system
3. Empirical framework
4. Data
5. Models and results

Introduction

- Readmissions to the Intensive Care Unit (ICU) are costly for hospitals and patients.
- Early detection of risk factors associated to readmissions can help improve patient care quality and reduce costs in the long-run.
- Readmissions due to ICU bed demand have long interested critical care physicians.
- Hypothesis: the odds of being readmitted to the ICU increases as the number of available ICU beds decrease.

Empirical framework

1. We estimate a model that predicts the most probable type of readmission (early, median or late) per patient conditional on ICU occupation, patient and hospital characteristics, using machine learning techniques.
2. We identify the effect of the number of beds on the risk of readmission by evaluating the impact of a policy that increased the number of ICU beds during 2010 in a high complexity hospital in Colombia, from which our data comes from. (Methodology pending).

Empirical framework

We predict three types of readmissions as three binomial classification tasks:

1. *Early readmissions*, those that occur within the first 72 hours after discharge.
2. *Median readmissions*, those that occur within 3 and 28 days after discharge
3. *Late readmissions*, those that occur past 28 days after discharge.

Empirical framework

Table: Classification in the literature of readmissions

Autor	Outcome	AUC
Gajic et al. (2008)	7-day readmission	0.70
Badawi and Breslow (2012)	48h readmission	0.71
Ferreira et al. (2014)	Readmission	0.76
Goulart et al. (2015)	48h readmission	0.74
Ouanes et al. (2011)	7-day readmission	0.74
Fiahlo et al. (2012)	72h readmission	0.72
Bayati et al. (2014)	30-day readmission	0.66
Campbell et al. (2008)	48h readmission	0.67
Jo et al. (2015)	Readmission	0.76

Data

- We have a unique database of a high complexity hospital in Colombia with 53,841 admissions to the adult ICU from 1998 to 2015.
- Building the dependent variable.
- Building new features.

Data

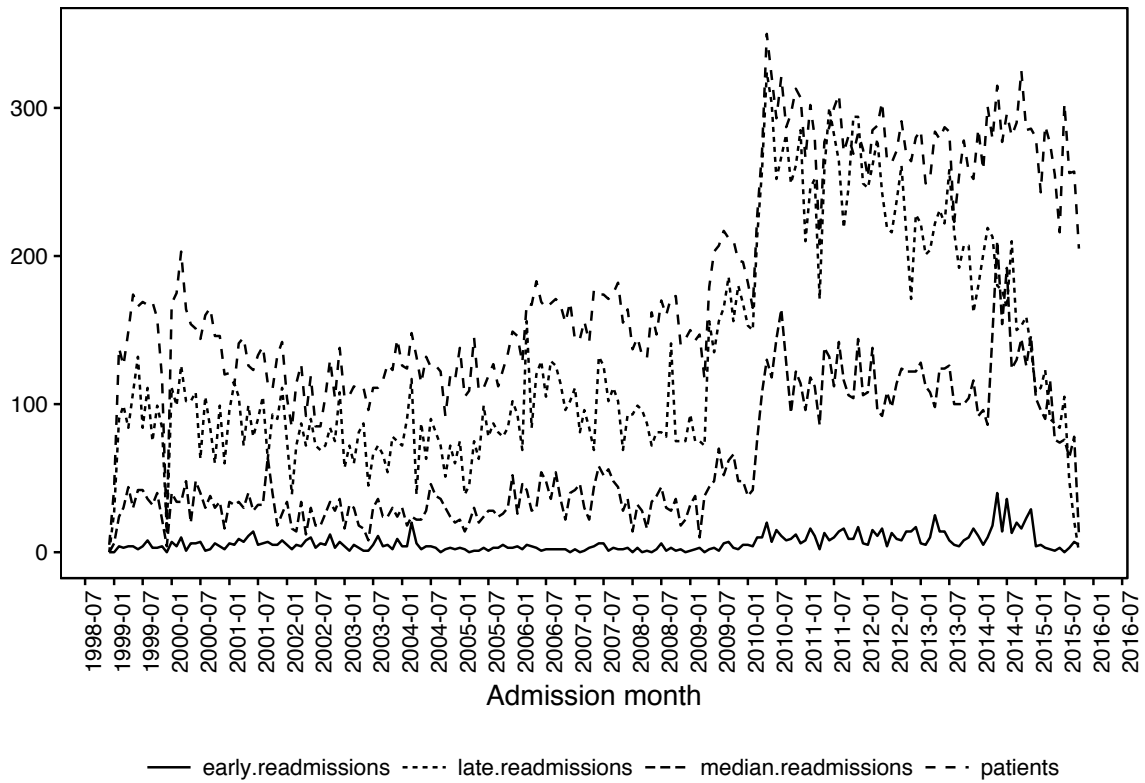


Figure: Number of patients and readmissions each month at the ICU

Data

Table: Descriptive statistics

	No readm.	Early		Median		Late	
	mean	mean	diff (1)-(0)	mean	diff (2)-(0)	mean	diff (3)-(0)
Diagnoses							
Pathology of the aorta	0.01	0.02	0.01	0.01	0.00	0.01	0.00
Rheumatic	0.00	0.02	0.02***	0.01	0.01***	0.01	0.01***
Shock	0.07	0.05	-0.02***	0.07	0.00	0.06	-0.01***
Pregnancy	0.01	0.00	-0.01	0.00	-0.01***	0.00	-0.01***
Respiratory	0.10	0.10	0.00	0.11	0.01***	0.11	0.01***
Major Post-Op	0.31	0.19	-0.12***	0.20	-0.11***	0.24	-0.07***
Trombosis	0.02	0.03	0.01**	0.03	0.01***	0.02	0.00
Neurologic	0.11	0.09	-0.02	0.07	-0.04***	0.06	-0.05***
Trauma	0.08	0.04	-0.04**	0.03	-0.05***	0.02	-0.06***
Gastrointestinal	0.04	0.05	0.01	0.06	0.02***	0.05	0.01**
Cardiac	0.31	0.39	0.08***	0.37	0.06***	0.40	0.09***
Chronic cardiac risk	0.35	0.40	0.05**	0.38	0.03	0.41	0.06***
Infections	0.09	0.10	0.01	0.15	0.06***	0.11	0.02***
Renal	0.04	0.06	0.02	0.08	0.04***	0.08	0.04***
Burns	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Electrolyte imbalance	0.03	0.03	0.00	0.03	0.00	0.04	0.01***
Multiple organ failure	0.01	0.00	-0.01	0.00	-0.01	0.00	-0.01
Metabolic disorder	0.01	0.01	0.00	0.02	0.01	0.02	0.01***
Hepatic	0.01	0.02	0.01	0.04	0.03***	0.04	0.03***
Cancer	0.05	0.05	0.00	0.06	0.01***	0.05	0.00
Poisoning	0.01	0.00	-0.01	0.00	-0.01***	0.00	-0.01***

Models and results

Table: Odds ratios for the multinomial logit model

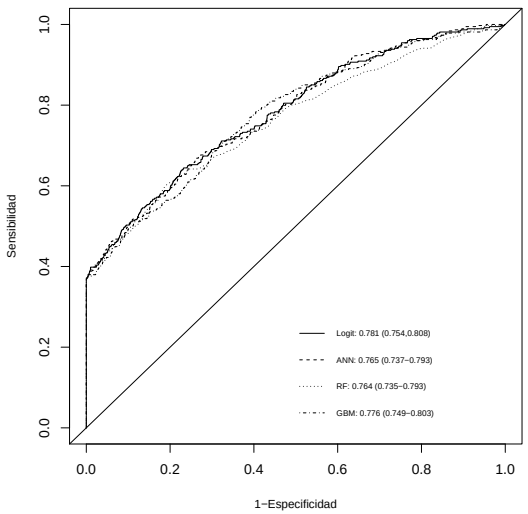
	Early		Median		Late	
Hospital and demographics						
Age	1.056	(0.978,1.141)	1.065	(1.021,1.111)	1.030	(0.991,1.071)
Apache	1.101	(1,1.212)	1.080	(1.03,1.132)	1.016	(0.972,1.061)
Length of stay	0.088	(0.057,0.136)	1.073	(0.992,1.16)	1.046	(0.967,1.131)
Catheter days	1.090	(0.814,1.459)	1.119	(1.052,1.19)	1.101	(1.041,1.164)
Arterial lines	0.953	(0.857,1.06)	1.048	(1.001,1.097)	0.925	(0.883,0.968)
Hrs invasive ventil.	0.890	(0.619,1.28)	0.881	(0.821,0.946)	1.023	(0.972,1.076)
# of patients	1.030	(0.96,1.105)	1.022	(0.983,1.063)	0.936	(0.903,0.971)
# of diagnoses	1.210	(1.058,1.384)	1.072	(0.993,1.157)	1.075	(1.004,1.151)
# of procedures	0.897	(0.753,1.068)	1.064	(1.014,1.117)	1.020	(0.976,1.066)
# of complications	1.198	(1.056,1.358)	1.002	(0.956,1.051)	1.034	(0.992,1.079)

Models and results

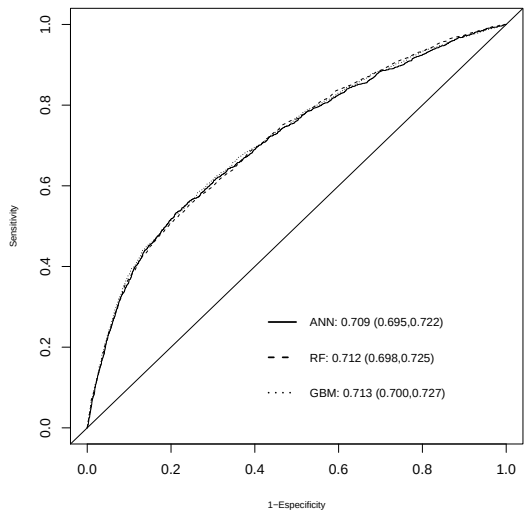
Table: Odds ratios for the multinomial logit model

	Early		Median		Late	
Diagnoses						
Rheumatic	3.791	(2.052,7)	2.573	(1.689,3.92)	2.200	(1.456,3.325)
Major Post-Op	0.531	(0.408,0.691)	0.654	(0.571,0.75)	0.919	(0.814,1.036)
Shock	0.756	(0.514,1.111)	0.885	(0.733,1.07)	0.808	(0.674,0.968)
Neurologic	1.029	(0.767,1.382)	0.760	(0.639,0.904)	0.731	(0.621,0.86)
Trauma	0.735	(0.497,1.088)	0.488	(0.387,0.616)	0.288	(0.224,0.37)
Cardiac	0.772	(0.585,1.017)	1.131	(0.974,1.312)	1.193	(1.044,1.364)
Renal	0.857	(0.604,1.217)	1.297	(1.085,1.55)	1.470	(1.248,1.731)
Electrolyte imbalance	0.561	(0.354,0.891)	0.690	(0.535,0.89)	0.956	(0.763,1.198)
Hepatic	1.402	(0.822,2.393)	3.334	(2.599,4.276)	2.742	(2.15,3.497)
Cancer	1.011	(0.712,1.436)	1.232	(1.024,1.481)	1.059	(0.887,1.263)
Poisoning	0.793	(0.276,2.274)	0.145	(0.042,0.498)	0.138	(0.045,0.423)

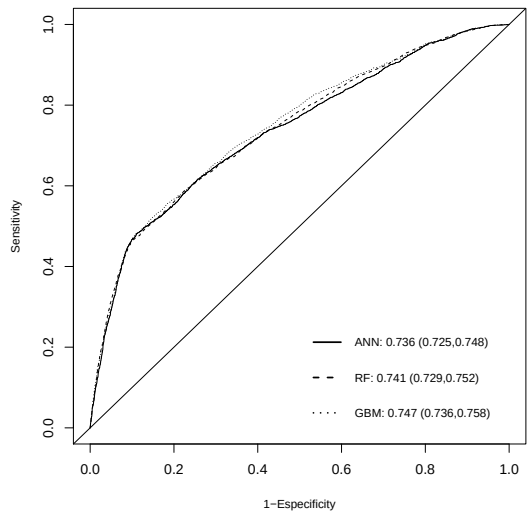
Models and results



Early readmissions



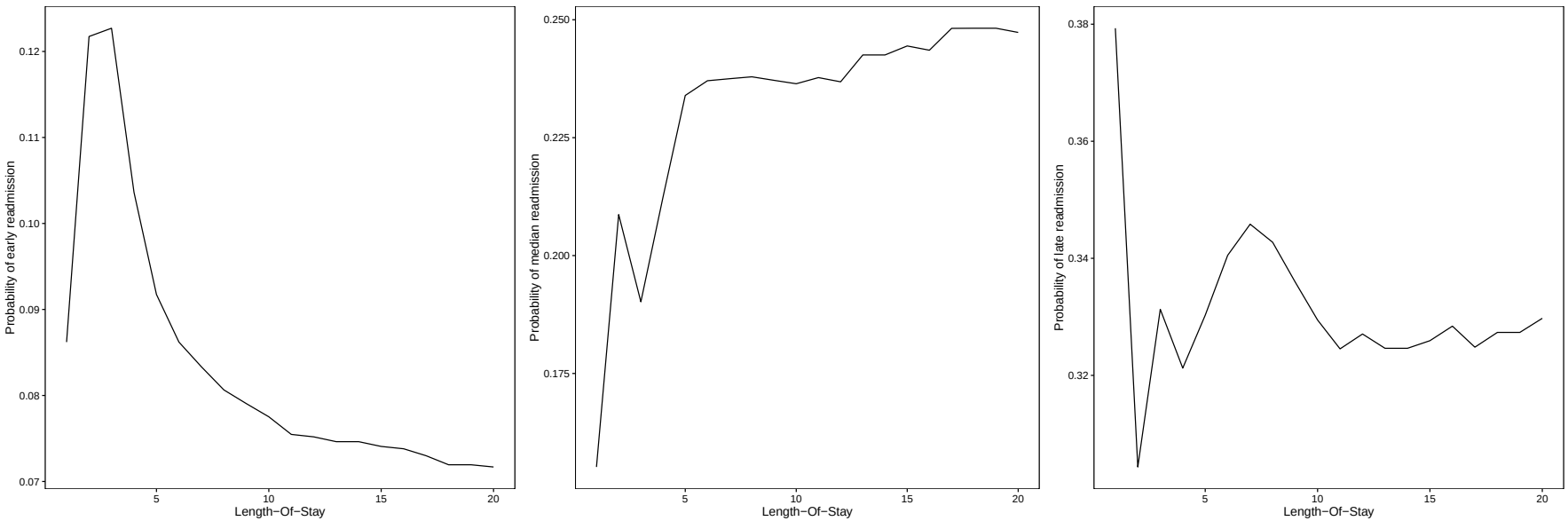
Median readmissions



Late readmissions

Figure: ROC curve for each type of readmission

Patients with rheumatic diseases



Early readmissions Median readmissions Late readmissions

Figure: Dynamics of the readmission probability

GRACIAS