No. 1
AGOSTO de 2016

quantil

Documentos de Trabajo

Edición electrónica

Predicting Readmissions, Mortality, and Infections in the ICU using Machine Learning Techniques

Natalia Serna
Álvaro J. Riascos
Marcela Granados
Fernando Rosso
Ramiro Guerrero

quantil | matemáticas aplicadas

Serie Documentos de Trabajo Quantil, 2016-1
Edición electrónica.

AGOSTO de 2016

Carrera 7 # 77 - 07. Oficina 901, Bogotá, D. C., Colombia
Teléfonos: 3718132 – (310)6791459 – (320)8461236
E-mail: info@quantil.com.co
http://www.quantil.com.co

Impreso en Colombia – Printed in Colombia

Predicting Readmissions, Mortality, and Infections in the ICU using Machine Learning Techniques*

Natalia Serna† Álvaro J. Riascos‡ Marcela Granados§ Fernando Rosso¶
Ramiro Guerrero‖

August 28, 2016

Abstract

Health care at the Intensive Care Unit (ICU) is both expensive for hospitals and strenuous for doctors. Early detection of risk factors associated to readmissions, mortality, and infections in the ICU, can improve patient care quality and reduce costs in the long-run. In this article we use machine learning techniques to predict those three outcomes using patient-level data of the ICU of a high complexity hospital in Colombia. Our results show that pathologies of the aorta, cancer, neurologic and respiratory diseases as well as invasive procedures such as dialysis, tracheostomy, and bronchoscopy are positively correlated to the probability of readmission, death, and catheter infections in the ICU. The area under the ROC curve for the first outcome ranges between 71 and 74%, for the second between 76 and 81%, and for the third between 88 and 92%. We estimate a model that competes against the APACHE II scoring system and achieve the same predictive power using less information about the patient.

Keywords: *Intensive Care Unit, machine learning, readmissions, mortality, catheter infections.*

*Acknowledgements: this study would not have been possible without the contribution of Robinson Gutierrez, Eliana Manzi, and Andrés Castro (Fundación Valle del Lili), involved in assembling the database. All opinions and possible errors or omissions are responsibility of the authors. The content of this article does not compromise any of the authors' affiliations

†Quantil. e-mail: natalia.serna@quantil.com.co

‡Universidad de los Andes and Quantil. e-mail: alvaro.riascos@quantil.com.co

§Fundación Valle del Lili, e-mail: mgranados@fcvl.org

¶Fundación Valle del Lili, e-mail: frosso@fcvl.org

‖PROESA - Universidad Icesi, e-mail: rguerrero@proesa.org.co

Contents

1 Introduction

Intensive Care Unit (ICU) services are amongst the costlier services provided by hospitals. Early detection of risk factors associated to readmissions, death, and infections in the ICU, can improve patient care quality and reduce costs in the long-run. On one hand, early discharge could be associated to a greater risk of readmission which at the same time represents lower quality in treatment. But prolonged length of stay could also increase the risk of infections which is detrimental for patients' health. On the other hand, mortality at the ICU is the gold standard for measuring hospital quality. Our objective in this study is two-fold: one is to find the risk factors that significantly correlate with the event of being readmitted, dying, and getting a catheter infection at the ICU; the other is to predict the risk scores for those outcomes using machine learning methods in order to provide quantitative tools for health management at the ICU.

To the best of our knowledge, very few health care literature borrows from the techniques of machine learning to make predictive exercises. Although making accurate predictions in the field of health care is fundamental to guide medical efforts and resources, and machine learning methods have proven to have greater predictive power than the usual statistical models in health care, machine learning remains lagged from this field. Counted studies have advanced in this type of modeling: Sujin et al. (2011) use artificial neural networks (ANN), support vector machines (SVM), and decision trees to predict mortality at the ICU; Ramon et al. (2007) study the evolution of patients in the ICU through outcomes such as survival, inflammation, and kidney dysfunction using decision trees, random forests, bayesian networks, naive bayesian networks, and tree-augmented naive bayesian networks; Fiahlo et al. (2012) use association rules or fuzzy modeling to predict ICU readmission between 24 and 72 hours after discharge; and Buchner et al. (2015) use regression trees to predict health care costs. Our study differs from the ones mentioned for various reasons: first, we are not only interested in estimating the risk score for each outcome but also in finding correlations that are significant from a medical point of view; second, we use ensemble techniques such as bagging and boosting to improve predictive power through variance reduction; third, we compare the usual logit model to machine learning methods to see how much predictive power we gain only through mathematical modeling; and fourth, we use a unique database of ICU patients in a high complexity[1] hospital in Colombia, but have no information regarding lab tests and physiological patient characteristics, therefore feature selection and machine learning modeling will be crucial to achieve high predictive power.

We study three different outcomes:

1. Readmissions to the ICU stratified as:
   - *Early readmissions*, those that occur within 72 hours after discharge.
   - *Median readmissions*, those that occur within 3 and 28 days after discharge
   - *Late readmissions*, those that occur past 28 days after discharge.
2. Mortality at the ICU, with two exercises:
   - Using variables known only at the moment of admission, which makes our models comparable to the APACHE II.
   - Using variables known at the moment of admission and during the stay, which makes our model predictions indicators of hospital quality.

[1] High complexity hospitals are institutions that meet at least the following requirements according the Ministry of Health and Social Protection (Decreto 1760 de 1990, art. 9.): (i) pathologies with high frequency and complexity, (ii) large population basis in the municipalities they cover, (iii) provision of health care in other municipalities where there is only presence of low complexity hospitals, (iv) high quality technology, and (v) specialized and sub-specialized personnel in charge of providing health care.

3. Catheter infections at the ICU.

The first outcome constitutes a multinomial predictive task while the other two are binomial. Some risk factors found in the literature are transversal to all of these outcomes. They include, for example, comorbidities and diagnoses (Badawi and Breslow (2012), Kramer and Higgins (2012), Jo et al. (2015)); complications such as lung failure, sepsis, bleeding, nosocomial infections, and accidental extubations (Kogan et al. (2003), Litmathe et al. (2009), Benetis et al. (2013)); procedures such as invasive and noninvasive ventilation and sedation; demographics such as gender and age (Campbell et al. (2008), Badawi and Breslow (2012), Bayati et al. (2014), Brown and Ratcliffe (2012), Sujin et al. (2011), Ramon et al. (2007)); lab tests such as red blood cell count, white blood cell count, platelet count, creatinine, etc. (Fiahlo et al. (2012), Badawi and Breslow (2012), Sujin et al. (2011)); and some physiological characteristics such as respiratory rate, heart rate, weight and height (Ramon et al. (2007)).

The remainder of this paper is organized as follows: section two, shows some descriptive statistics for each outcome and the data preprocessing. Section three, presents the machine learning models and describes them briefly. Section four shows the results and explain how the optimal parameters for certain models are selected. Section five presents a comparison with related literature. And section six outlines some conclusions, limitations, and ideas for further research.

2 Descriptive statistics and data preprocessing

To study readmissions, mortality, and catheter infections at the ICU we have a unique database of a high complexity hospital in Colombia with 53,841 admissions to the adult ICU from 1998 to 2015. We have information collected at the moment of admission such as: admission date, age, gender, health insurer, municipality of residence, admission diagnosis, patient origin, cause of admission, and Acute Physiology, Age, and Chronic Health Evaluation (APACHE) II. Variables collected during the patient's stay such as length of stay, catheter days, days of bladder catheterization, procedures, complications, medical monitoring, number of Swan Ganz catheters, hours of invasive and noninvasive ventilation, and indicators of platelet count above 150,000, red blood cell transfusion, and other blood transfusions. And finally we have variables collected at discharge such as discharge diagnosis. Admission and discharge diagnoses are not coded following any international standardization like the ICD-10 codes. Rather, they are described as syndromic diagnoses whose categories were created specifically for the needs and characteristics of this hospital's ICU.

We create other variables following some of the related literature: daily patient inflow as a proxy for ICU occupation; number of discharge diagnoses, procedures, and medical monitoring received by the patient; admission and discharge day of week; admission and discharge month; indicators of outliers in length of stay, catheter days, days of bladder catheterization, and hours of invasive ventilation, defining outliers as those patients whose measure lies above the 95th percentile of each variable distribution; and an indicator of past admissions.

For the exercise of predicting readmissions, the dependent variable is created as follows: first we locate the first entrance of each patient, then we look three and 28 days later whether there is an early, median, or late readmission, whichever happens first, and mark this first entry accordingly. Then we move forward to the second entry of the patient and look three and 28 days later whether there is another entry or not. We continue this process until the last entry of each patient for which

we will have no information. By definition, a patient who dies at the ICU will have no risk of being readmitted, therefore we eliminate those observations where the discharge condition is death. For estimating the probability of being readmitted to the ICU we are able to use all information known of the patient at the moment of discharge. This includes all discharge diagnoses, procedures, complications, monitoring, length of stay, catheter days, etc.

For the exercise of predicting mortality, our dependent variable takes the value of 1 if the discharge condition is death and 0 otherwise. Since the dependent variable is measured at the moment of discharge we are not able to use other variables measured at this same moment as predictors. Hence, for the mortality score that is comparable to the APACHE II we are left with only those variables that are measured at the moment of admission, but for the mortality score that serves as a quality indicator of the hospital, we are additionally able to use variables measured *during* the stay at the ICU. Also, for the first exercise, we exclude those patients that are submitted from another ICU because the APACHE II scores are designed to measure mortality risk at the first entry to the ICU.

In the case of predicting catheter infections, our dependent variable takes the value of 1 if the patient presents one of the following complications: catheter sepsis, catheter bacteremia, or insertion site infection. We think of the probability of getting a catheter infection as one that can be updated during the patient's stay but has no utility once the patient has been discharged. This implies we are not able to use the variables that are measured at the moment of discharge as predictors but we can use variables measured at the moment of admission and during the stay.

In this section we present some descriptive statistics for each outcome. Table (1) shows the descriptive statistics for each type of readmission and the base category (no readmissions). There are no significant differences between patients readmitted within the first 72 hours and those who are not readmitted. However, differences are marked in relation to patients readmitted after the third day of discharge. Patients with readmissions within 3 and 28 days are 4 years older than those who are not readmitted and have higher APACHE II scores. Patients with early readmissions have significantly less hours of invasive ventilation, days of bladder catheterization, and medical monitoring compared to the ones who are not readmitted. No differences are found regarding the number of diagnoses received by patients with early readmissions and no readmissions. But patients with late readmissions have significantly more hours of invasive ventilation relative to the base category as well as a higher quantity of diagnoses. Evidence also shows there is a higher proportion of readmitted patients (under any definition) that come from other ICUs, while all non readmitted patients come directly from their houses to the ICU.

Table 1: Descriptive statistics for the outcome of readmissions

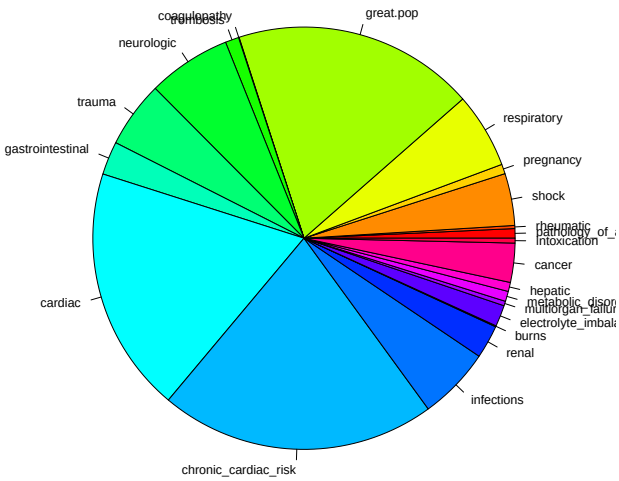
	No readmissions		Early		Median		Late	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Demographics								
Male	0.56	0.5	0.54	0.5	0.55	0.5	0.55	0.5
Age	55.88	35.22	59.36*	18.75	60.18	27.33	59.57	23.95
Medical/Hospital								
Apache	12.60	5.70	12.07*	5.37	13.32*	5.55	12.65	5.33
Length of stay	4.07	27.2	2.12*	2.27	4.42	4.56	4.16	5.92
Catheter days	1.59	3.64	0.53*	1.7	1.7	3.55	1.64	4.56
Days of bladder catheterization	2.31	4.79	0.8*	2.03	2.23	3.89	2.14*	5.01
# of central catheters	0.4	1.22	0.19*	0.45	0.4	1.24	0.36*	0.77
# of Swan Ganz	0.08	0.46	0.02*	0.14	0.08	0.34	0.08	0.38
# of arterial lines	0.5	0.66	0.33*	0.47	0.5	0.58	0.43*	0.59
Hours of invasive ventilation	28.99	96.98	6.74*	37.37	24.81*	68.78	29.44	118.56
Hours of noninvasive ventilation	0.01	0.47	0	0.17	0.02	1.55	0.01	0.24
Red blood cell transfusion	0.03	0.17	0.02	0.15	0.03	0.17	0.03	0.16
Other transfusions	0.02	0.14	0.02	0.13	0.02	0.13	0.02	0.13
Platelets>150000	0.02	0.15	0.02	0.13	0.02	0.15	0.02	0.14
Past admissions	0	0	0.37*	0.48	0.41*	0.49	0.46*	0.5
# of diagnoses	1.8	0.83	1.84	0.86	1.9*	0.83	1.93*	0.87
# of procedures	0.14	0.48	0.03*	0.18	0.13	0.45	0.14	0.49
# of complications	0.08	0.38	0.03*	0.2	0.08	0.34	0.09	0.4
# of medical monitoring	0.49	1.06	0.27*	0.76	0.38*	0.95	0.37*	0.97
Admission cause cardiology	0.27	0.44	0.33*	0.47	0.31*	0.46	0.34*	0.47
Admission cause medical	0.37	0.48	0.45*	0.5	0.46*	0.5	0.43*	0.5
Admission cause surgical	0.36	0.48	0.22*	0.41	0.23*	0.42	0.23*	0.42
Second floor hospitalization	0.95	0.21	0.94*	0.25	0.96	0.2	0.94*	0.23
Medical ICU hospitalization	0.05	0.21	0.06*	0.24	0.04	0.19	0.06*	0.23
House origin	0	0.01	0	0	0	0	0	0
Catheterizaton origin	0.02	0.15	0.01	0.11	0.02*	0.13	0.02	0.15
Surgery origin	0.26	0.44	0.16*	0.36	0.16*	0.36	0.18*	0.38
Consult origin	0	0.01	0	0	0	0.01	0	0.02
Other ICU origin	0.03	0.16	0.07*	0.26	0.08*	0.27	0.06*	0.24
Same hospital floor origin	0.05	0.21	0.06	0.24	0.09*	0.28	0.07*	0.26
Intermediate ICU origin	0	0.03	0.01*	0.09	0.01*	0.09	0	0.07
ER origin	0.63	0.48	0.66	0.47	0.62	0.48	0.64	0.48
N	24767		1253		5998		8678	

Note: The mean and standard deviation of each variable conditional to the readmission status are reported. (*) Indicates the difference in means or proportions between the type of readmission and the base category (no readmissions) is significant at a 99% confidence level. Authors calculations.

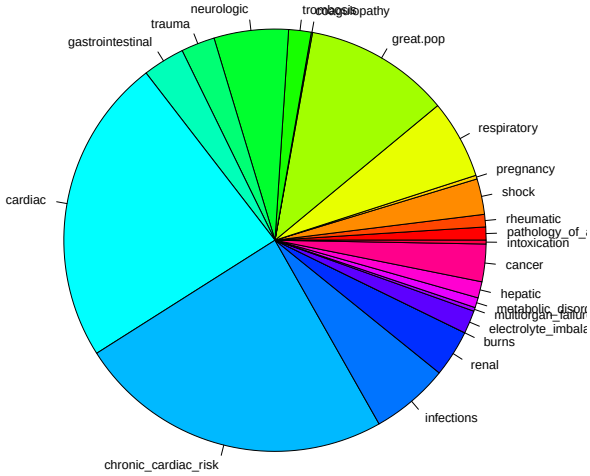
Figure (1) shows the distribution of discharge diagnoses among the different types of readmissions. A greater proportion of patients who are not readmitted enter the ICU following a major surgery, while among those who are readmitted under any definition, cardiac diseases and liver diseases are more recurrent.

Figure 1: Distribution of discharge diagnoses among types of readmissions

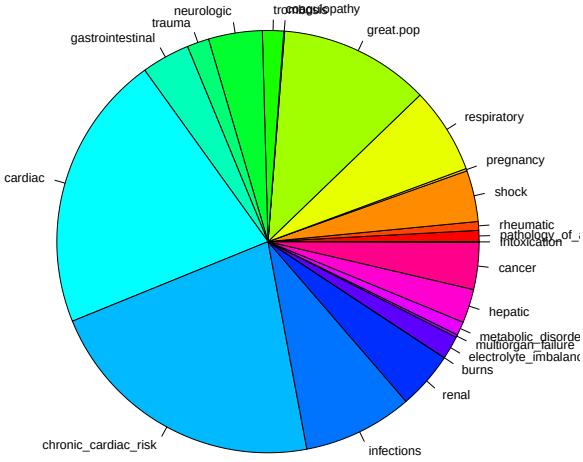
(a) No readmissions



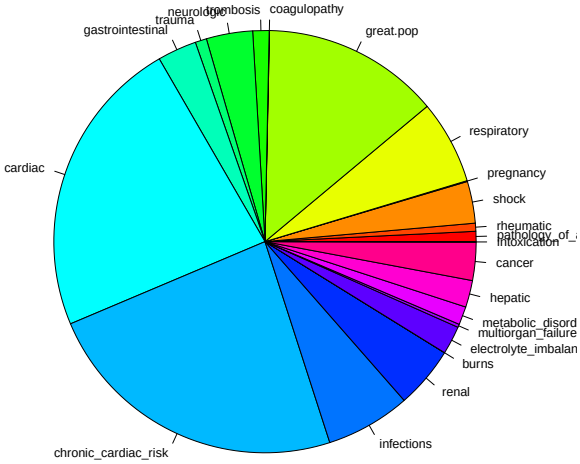
(b) Early readmissions



(c) Median readmissions



(d) Late readmissions



Greater differences between patients who are not readmitted and those who are readmitted stand out when comparing the type of procedures they receive. Figure (2) shows a significantly higher proportion of early readmitted patients who receive procedures related to cardiac arrhythmia and dialysis compared to those who enter only once. However, arrhythmia related procedures are less common between median and late readmitted patients than between those who are not readmitted or those who re-enter within the first 72 hours. Respiratory invasive procedures are also more prevalent among patients who enter once than among those who are early readmitted.

Figure 2: Distribution of procedures among types of readmissions

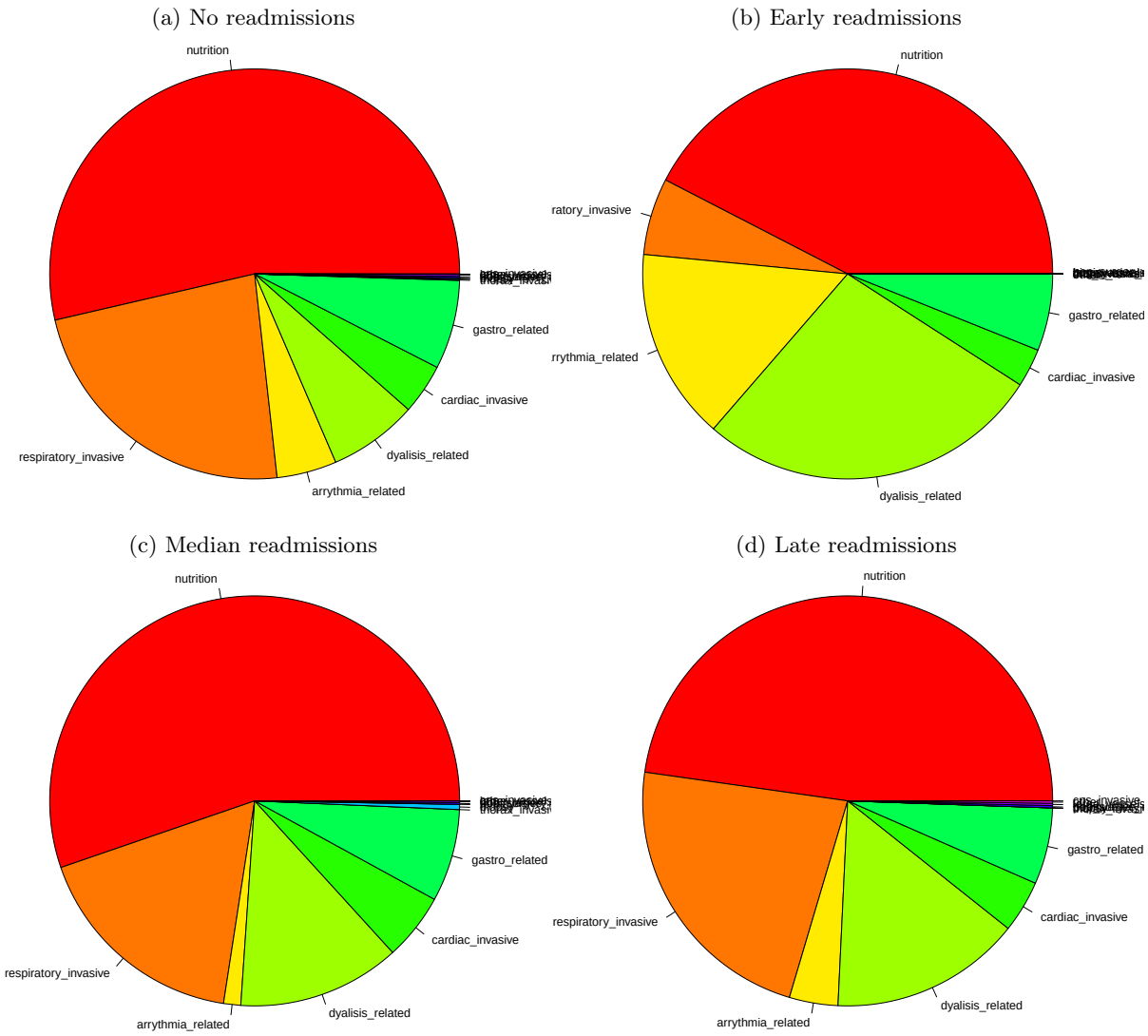
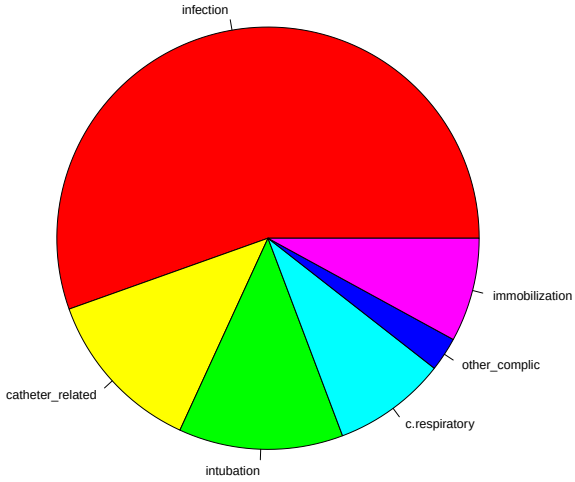


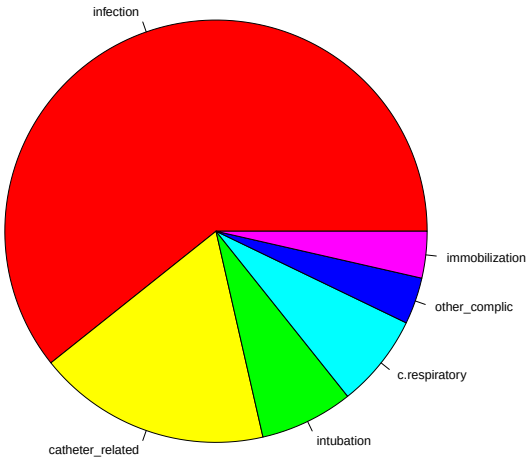
Figure (3) displays the distribution of complications by type of readmission. Four things are worth noticing from these figures: one is that complications related to infections (catheter infections and urinary infections) are more prevalent among early readmitted patients than among patients that enter only once; two, catheter related complications such as catheter tip colonization, pneumothorax by catheter, and hemothorax by catheter, are also more prevalent among early and late readmitted patients than among patients in the base category; three, immobilization is more recurrent in patients who will be readmitted within 3 and 28 days than in those who are not readmitted; and, four, patients who enter only once present more complications associated to intubation than patients who are readmitted.

Figure 3: Distribution of complications among types of readmissions

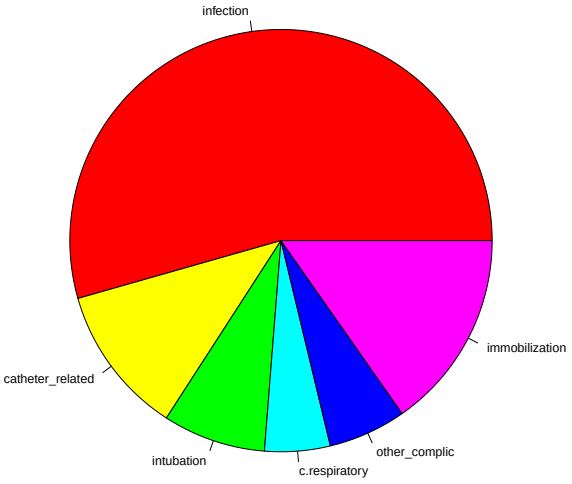
(a) No readmissions



(b) Early readmissions



(c) Median readmissions



(d) Late readmissions

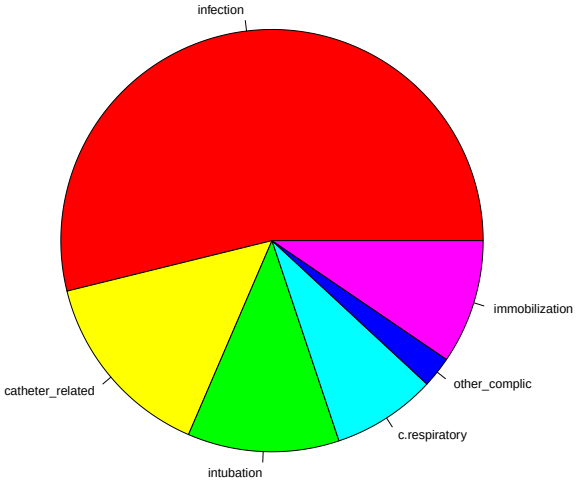


Table (2) shows the descriptive statistics, mean and standard deviation, of some variables by mortality status. Overall, patients who die at the ICU present worse health conditions than those that survive. For example, patients who die have longer length of stay and invasive ventilation, more catheter days and days of bladder catheterization, and greater number of arterial lines than patients that survive. They also present significantly more complications in the ICU and receive more procedures.

Table 2: Descriptive statistics for the outcome of mortality

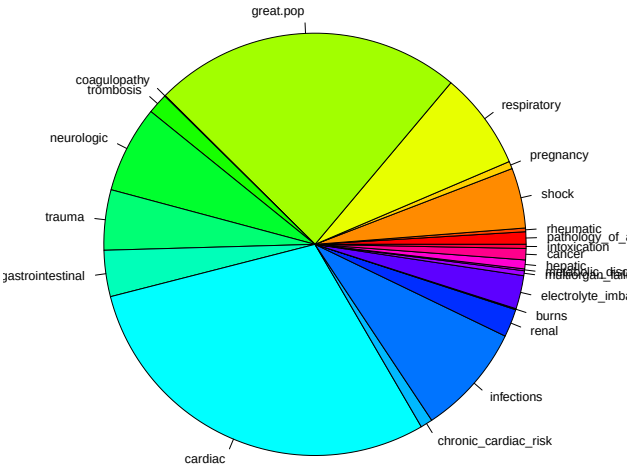
	Mortality=0		Mortality=1	
	Mean	SD	Mean	SD
Demographics				
Male	0.56	0.5	0.56	0.5
Age	57.97	30.88	58.99	31.74
Medical/Hospital				
Length of stay*	4.11	23.78	5.39	7.08
Catheter days*	1.57	3.75	3.44	5.76
Days of bladder catheterization*	2.19	4.57	4.25	6.71
# central catheters*	0.38	1.09	0.84	2.04
# of Swan Ganz*	0.08	0.45	0.26	0.59
# of arterial lines*	0.48	0.62	0.84	0.55
Hours of invasive ventilation*	26.92	94.54	81	162.2
Hours of noninvasive ventilation	0.01	0.86	0.1	3.61
Red blood cell transfusion*	0.03	0.17	0.07	0.26
Other transfusions*	0.02	0.14	0.05	0.23
Platelets>150000*	0.02	0.15	0.06	0.24
# of procedures*	0.13	0.47	0.37	0.81
# of complications*	0.08	0.37	0.18	0.58
# of medical monitoring*	0.43	1	1.08	1.49
Admission cause cardiology*	0.29	0.45	0.13	0.33
Admission cause medical*	0.41	0.49	0.66	0.47
Admission cause surgical*	0.3	0.46	0.21	0.41
Second floor hospitalization*	0.95	0.21	0.99	0.07
Medical ICU hospitalization*	0.05	0.21	0	0.07
House origin	0	0.01	0	0
Catheterization origin*	0.02	0.15	0.01	0.08
Surgery origin*	0.22	0.41	0.09	0.29
Consult origin	0	0.01	0	0
Other ICU origin*	0.06	0.23	0.1	0.3
Same hospital floor origin*	0.07	0.25	0.11	0.31
Intermediate ICU origin*	0	0.06	0.01	0.09
ER origin*	0.61	0.49	0.66	0.47
N	48525		4867	

Note: The mean and standard deviation of each variable conditional to the mortality status are reported. (*) Indicates the difference in means or proportions between patients that die and those who live is significant at a 99% confidence level. Authors calculations.

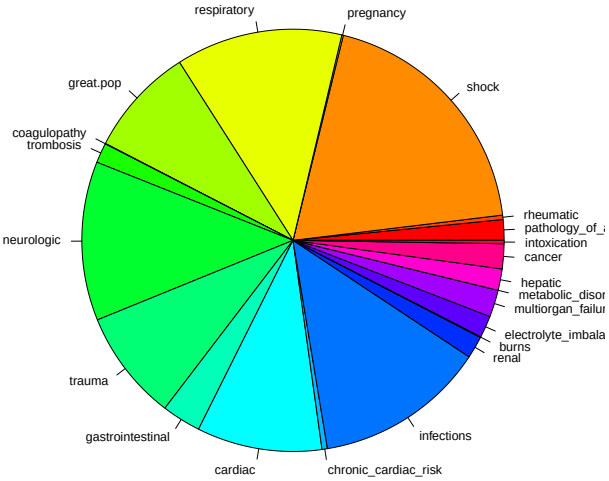
Figures (4a) and (4b) illustrate the distribution of admission diagnoses among patients who die at the ICU and those who survive. A greater proportion of patients who die at the ICU enter with diagnoses associated to shock, infections, lung diseases, and neurologic diseases than those who survive. This last group of patients, on the other hand, enter primarily with cardiac related diagnoses and because of a major surgery postoperative. There are also significant differences in the proportion of patients who die at the ICU that are admitted with diagnoses of cancer, multiple organ failure, pathologies of the aorta, and electrolyte imbalance contrary to those who survive.

Figure 4: Distribution of admission diagnoses, procedures, and complications among patients who survive and die at the ICU

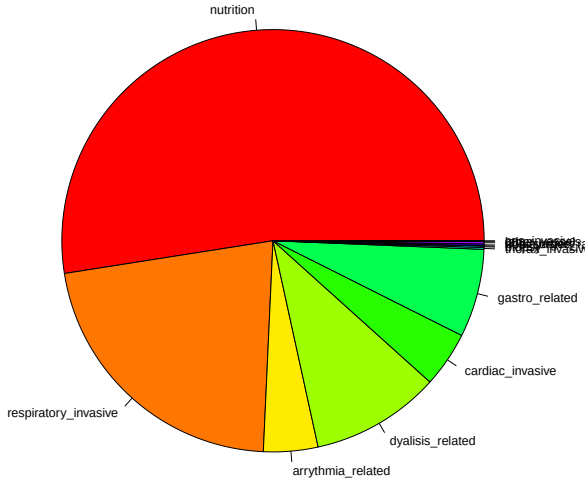
(a) Mortality=0



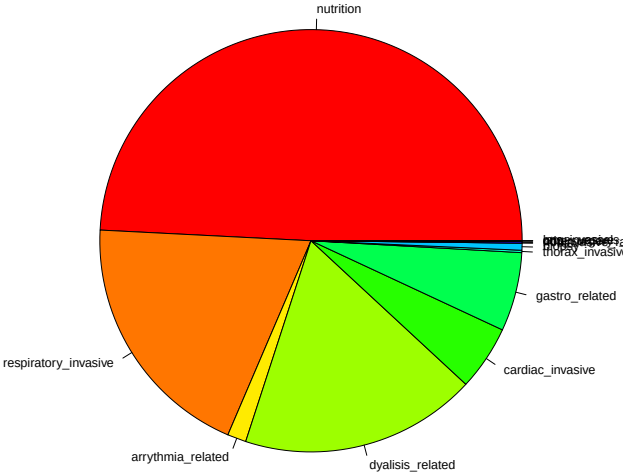
(b) Mortality=1



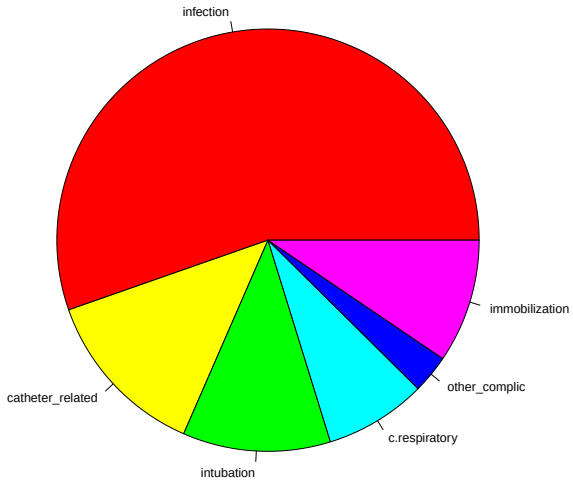
(c) Mortality=0



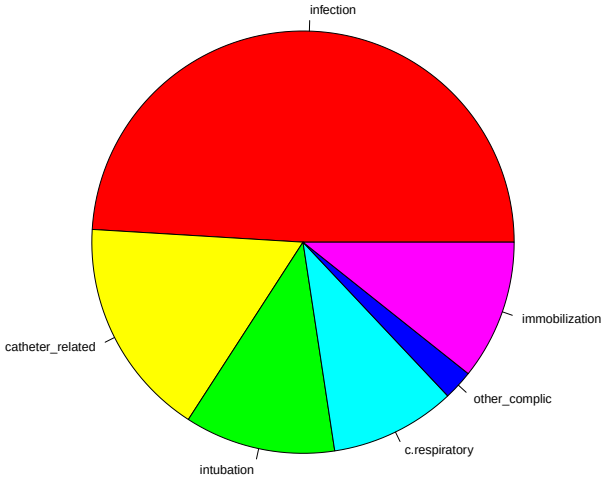
(d) Mortality=1



(e) Mortality=0



(f) Mortality=1



Figures (4c) and (4d) present the distribution of procedures for the group of patients that survive and die at the ICU, respectively. In this case, one thing is worth noticing: dialysis related procedures and arrhythmia related procedures are more prevalent among the group of patients that die than among those who survive. Finally, there are no significant differences in the proportion of patients per type of complication in each group as shown in figures (4e) and (4f).

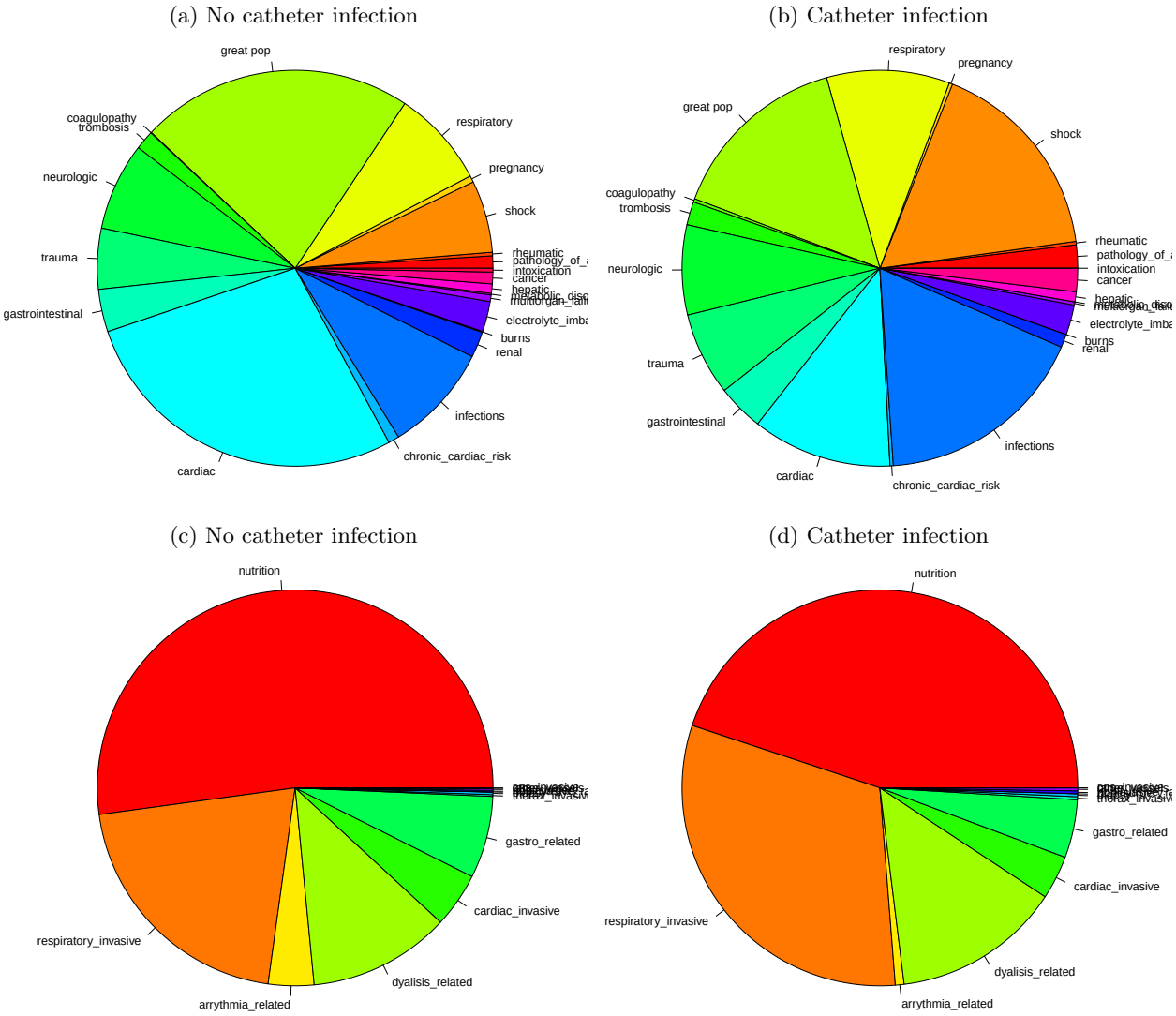
Table 3: Descriptive statistics for the outcome of catheter infection

	No catheter infection		Catheter infection	
	Media	sd	Media	sd
Demographics				
Male	0.56	0.5	0.56	0.5
Age*	58.48	18.39	55.51	17.43
Medical/Hospital				
Apache*	13.36	6.38	17.84	6.57
Length of stay*	4.12	22.36	20.85	55.03
Catheter days*	1.67	3.85	11.36	10.68
Days of bladder catheterization*	2.3	4.65	14.02	12.49
# central catheters*	0.42	1.21	1.51	1.34
# of Swan Ganz*	0.09	0.38	0.73	3.41
# of arterial lines*	0.51	0.62	1.18	0.96
Hours of invasive ventilation*	30	97.8	303.73	336.41
Hours of noninvasive ventilation	0.02	1.37	0.02	0.32
Red blood cell transfusion*	0.03	0.18	0.14	0.34
Other transfusions*	0.02	0.15	0.11	0.31
Platelets>150000*	0.03	0.16	0.12	0.32
# of procedures*	0.14	0.49	1.34	1.31
# of complications	0.08	0.36	0.93	1.05
# of medical monitoring*	0.48	1.07	1.25	1.5
Admission cause cardiology*	0.27	0.45	0.15	0.35
Admission cause medical*	0.43	0.5	0.55	0.5
Admission cause surgical	0.29	0.46	0.3	0.46
Second floor hospitalization*	0.96	0.2	1	0.05
Medical ICU hospitalization*	0.04	0.2	0	0.05
House origin	0	0.01	0	0
Catheterization origin	0.02	0.14	0	0.05
Surgery origin	0.21	0.4	0.16	0.37
Consult origin	0	0.01	0	0
Other ICU origin*	0.06	0.24	0.14	0.34
Same hospital floor origin	0.07	0.25	0.1	0.3
Intermediate ICU origin	0	0.07	0.01	0.12
ER origin*	0.62	0.49	0.54	0.5
N	53032		360	

Note: The mean and standard deviation of each variable conditional on being infected or not are reported. (*) Indicates the difference in means or proportions between patients that get catheter infections and those who don't is significant at a 99% confidence level. Authors calculations.

Table (3) shows the descriptive statistics of several variables for the group of patients that get a catheter infection at the ICU and the group who doesn't. Significant differences are reported in age, APACHE II scores, length of stay, catheter days, days of bladder catheterization, hours on invasive ventilation, among others. Specifically, infected patients stay 16 more days at the ICU than those who do not get catheter infections, they have catheters for more than 10 days and bladder catheters for more than 14 days. On average, infected patients have 303 hours of invasive ventilation, while uninfected patients have 30 hours. 12% of infected patients have platelets over 150,000, 14% of them receive red blood cell transfusion and 14% of them come from other ICUs while only 6% of uninfected patients come from other ICUs.

Figure 5: Distribution of admission diagnoses and procedures among infected and uninfected patients



Panels (a) and (b) of figure (5) illustrate the distribution of admission diagnoses among the group of infected and uninfected patients. Shock, infections, and respiratory diseases at entry are more prevalent among patients who present catheter related infections during their stay at the ICU than among uninfected patients. Panels (c) and (d) also indicate a greater proportion of infected patients receive respiratory invasive procedures and dialysis.

The descriptive statistics presented in this section highlight several stylized facts in our sample of patients. First, admission diagnoses of shock, infections, respiratory diseases, cancer, and pathologies of the aorta are correlated with worse health outcomes after the ICU stay. Patients who present those diagnoses are more likely to be readmitted, to die, or to get a catheter infection at the ICU. Second, worse outcomes are also more likely among patients that receive dialysis and invasive procedures such as tracheostomy and bronchoscopy. Finally, patients who enter the ICU because of a major surgery post-op are less likely to be readmitted or to die at the ICU.

Before estimating the machine learning methods, further preprocessing is done to the data. All continuous variables are standardized (zero mean and unit standard deviation) and all categorical variables are dichotomized. With the purpose of avoiding over fitted predictions out-of-sample, the database is split into two mutually exclusive datasets whose observations are randomly selected from the pool of 53,841 admissions to the ICU. The training set comprises 70% of the database and the validation set comprises the remaining 30%.

3 Machine learning methods and feature selection

We borrow from the literature of machine learning the models and estimating techniques for predicting the outcomes of our study. For each outcome we estimate four models:

- Logit
- Random Forests (RF)
- Artificial Neural Networks (ANN)
- Boosted Trees (GBM)

The logit model is included because it is standard in medicine and because it facilitates the interpretation of the marginal effects of the predictors on the likelihood of being readmitted, dying, or getting a catheter infection at the ICU. In the case of readmissions, our dependent variable is categorical and takes four possible values: (0) no readmissions, (1) early readmissions, (2) median readmissions, and (3) late readmissions. Then, the problem of predicting the type of readmission is multinomial. For the other outcomes, mortality and infections, the dependent variable is binomial. We briefly describe each model and their estimation in the next subsections.

3.1 Logit model

Define y_i, for example, as the event of dying at the ICU, it takes the value of 1 if patient i dies and 0 otherwise, and let $\mathbf{x}_i'$ be the vector of patient characteristics. Equation (1) shows the relation between y_i and $\mathbf{x}_i'$.

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i \tag{1}$$

where ε_i is a patient-specific random shock. Assuming ε_i follows an extreme value type-I distribution, the probability of $y_i = 1$ conditional on $\mathbf{x}_i'$ will have the usual logit form as shown in equation (2).

$$\widehat{\pi_i} = Pr[y_i = 1|\mathbf{x}_i] = \frac{e^{\mathbf{x}_i'\widehat{\beta}}}{1 + e^{\mathbf{x}_i'\widehat{\beta}}} \tag{2}$$

The vector of parameters β is estimated by maximizing the log-likelihood function:

$$L(y) = \sum_{i=1}^{n} y_i log(\widehat{\pi_i}) + (1 - y_i) log(1 - \widehat{\pi_i}) \tag{3}$$

With $\widehat{\beta}$ we can compute the marginal effects of each predictor on the likelihood of dying at the ICU. Taking the derivative of $\widehat{\pi}_i$ with respect to x_i we have the following:

$$\frac{\partial \widehat{\pi}_i}{\partial x_i} = \widehat{\beta}\ \widehat{\pi}_i(1-\widehat{\pi}_i) \tag{4}$$

Hence, variations in x_i, change the likelihood of the event in $\widehat{\beta}\ \widehat{\pi}_i(1-\widehat{\pi}_i)$ percentage points. We can also compute the odds ratios as shown in equation (5).

$$Odds_i = \frac{\widehat{\pi}_i}{1-\widehat{\pi}_i} = \frac{\frac{e^{\mathbf{x}_i'\widehat{\beta}}}{1+e^{\mathbf{x}_i'\widehat{\beta}}}}{\frac{1}{1+e^{\mathbf{x}_i'\widehat{\beta}}}} = e^{\mathbf{x}_i'\widehat{\beta}} \tag{5}$$

In this case, variations in x_i should be interpreted in terms of the percentage increase or decrease in the relative risk of dying. Notice that if the confidence interval of the odds ratio includes the number 1, the odds is not significant.

For multinomial responses, if k is the number of categories in the dependent variable, the logit model requires estimating $k-1$ logistical regressions, one for each response relative to the base category, of the form $log(\pi_k/\pi_0) = \mathbf{x}_i'\widehat{\beta}$. The estimation is done through maximum likelihood and the marginal effects and odds ratios are calculated separately for each category. We will not go further on the assumptions underlying the estimation of a multinomial logit model, but the reader can refer to chapter 18 in Greene (2012).

3.2 Random forests

Assuming a particular distribution for the probability of observing an event does not necessarily improves prediction. In fact, parametric distributions often do not fit well the data. Many machine learning methods are mostly concerned with making accurate predictions in a non-parametric setting. This is the case of decision trees.

Suppose y_i is a function of $\mathbf{x}_i'$. Given a variable k and a realization of this variable, s, we split the data into two regions: $R_1(k,s) = \{X|X_k \leq s\}$ y $R_2(k,s) = \{X|X_k > s\}$. k and s are selected recursively through the minimization of a loss function, that in the case of a classification problem, is usually the Gini impurity. Let c_j be the proportion of observations in region j classified as 1, the loss function we wish to minimize with every k and s is:

$$I(R) = \min_{k,s}[c_1(1-c_1) + c_2(1-c_2)] \tag{6}$$

Each region is divided again into two subregions by choosing k and s from the set of all predictors looking to minimize the Gini impurity. The recursive partitioning continues until some convergence criteria is met, for example, after the reduction in the Gini impurity reaches a threshold or until the number of observations in a parent node reaches a threshold.

Since decision trees are sensible to the convergence criteria fixed by the researcher, their predictions are also volatile. High variance is problematic for out-of-sample predictions. In order to reduce variance several methods have been introduced in the field of machine learning. One of them is bagging, which consists in creating bootstrap samples in each of which a decision tree is grown. The out-of-sample predictions are averaged between trees to obtain one single predictor. This is the random forest, a bagging of decision trees where the variables selected at each node are also extracted from a random subset of variables. In this application we train 500 trees in the forest.

3.3 Artificial Neural Networks

Artificial neural networks (ANN), although parametric, are highly nonlinear models used in machine learning for both regression and classification tasks. Resembling the way the brain works, the ANN receives several inputs or predictors, whose information is processed in some inner neurons. Then, based on whether the information is relevant to predict the outcome, the neuron passes the prediction to the outer layer. An ANN with one inner layer of neurons can be written as:

$$f_k(X) = \sigma_k \left(\sum_{l=1}^{L} w_{2lk} \; \sigma \left(\sum_{j=1}^{J} w_{1jk} x_j + \theta_j \right) + \theta_l \right) \tag{7}$$

where L are the number of derived features, J are the number of inputs, σ_k and σ are activation functions, θ_l and θ_j are bias indicators at each node, and x_j is input j. We define the activation functions as sigmoid functions as follows:

$$\sigma_k(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

The parameters of the ANN, usually known as weights, $\beta = [w_{1jk}, w_{2lk}]$, are estimated using the backpropagation algorithm. The loss function we wish to minimize is the cross-entropy or deviance in equation (9):

$$R(\beta) = -\sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \; log f_k(x_i) \tag{9}$$

In the case of multiclass classification, activation functions are defined as:

$$\sigma_k(x) = \frac{e^{X_k}}{\sum_{l=1}^{K} e^{X_l}} \tag{10}$$

which is the same transformation of the multinomial logit model. In this study we go further in the estimation of ANNs by doing bagging. We estimate 30 ANNs in random subsets of 70% of the training set and then average their predictions. All ANNs are estimated with the same parameters. We discuss the way we fix these parameters at the section of results.

3.4 Boosted Trees

Boosted trees (GBM) are a special case of decision trees, in which trees are grown sequentially in weighted samples. It differs from random forests in the sense that trees in forests are grown simultaneously and not sequentially and in the sense misclassified observations in the training set receive higher weight in subsequent iterations. The boosting algorithm consists of the following steps:

1. Set $f_0(x) = 0$
2. For m=1 to M:
   (a) Compute $(\beta_m, \gamma_m) = arg\min_{\beta,\gamma} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \beta T(x_i, \gamma))$
   (b) Set $f_m(x) = f_{m-1}(x) + \beta_m T(x, \gamma_m)$

where γ represents all the parameters of tree T (the split variables and points at the internal nodes, and the predictions at the terminal nodes), L is any loss function, and β is a scale parameter. The latter is estimated using cross-validation in a grid of β's.

If the loss function is an exponential function $L(y, f(x)) = e^{-yf(x)}$ then step 2(a) can be rewritten as:

$$\begin{aligned}(\beta_m, \gamma_m) &= arg\min_{\beta,\gamma} \sum_{i=1}^{N} e^{-y_i(f_{m-1}(x_i)+\beta T(x_i,\gamma))} \\ &= arg\min_{\beta,\gamma} \sum_{i=1}^{N} e^{-y_i f_{m-1}(x_i)} e^{-y_i \beta T(x_i,\gamma)} \\ &= arg\min_{\beta,\gamma} \sum_{i=1}^{N} w_i^m \; e^{-y_i \beta T(x_i,\gamma)}\end{aligned} \tag{11}$$

The last line in equation (11) follows since $e^{-y_i f_{m-1}(x_i)}$ does not depend on the parameters over which the minimization is done, which is equivalent to finding weights w_i^m for each observation i in iteration m. Moreover, for misclassified observations for which $-y_i f_{m-1}(x_i)$ is less than 1, the weight will be higher in subsequent iterations.

The solution to the minimization problem yields predictions equal to one-half of the weighted log-odds in each region. The final predictor or boosted tree $g(x)$ is then, basically, a weighted average of trees.

3.5 Comparison of classification models

The predictive power of all the models presented in the previous subsections is compared using the Area Under the Receiver Operating Characteristics Curve (AUC). The closest to 1 the better the model. For the comparison to be fair, we also need the subset of variables on which the models are trained to be the same. To choose the variables we perform two feature selection techniques: decision trees and boosted trees. For the first technique we estimate 20 trees on the training sample fixing the complexity parameter in 0.001 and allowing for a minimum of 50 observations in the nodes. Since the parameters are fixed at a relatively small level, the trees will be over fitted and we need

to prune them. After pruning, we register the variables that were finally used to split the area of characteristics and select those that appear in more than 4 trees.

For the second technique we estimate a boosting of 500 trees and compute the relative importance of each variable. The relative importance is a measure of the influence of feature x_j on the variation of $g(x)$ over the joint input variable distribution, in other words, it measures how much does the loss function decreases when using feature x_j to split the area of characteristics in two regions. Breiman et al. (1983) propose the next approximation to the relative influence:

$$\hat{I}_j(T) = \sum_{t=1}^{J-1} \hat{l}(y, f(x)) \mathbf{1}(x_t = j) \tag{12}$$

where $J-1$ are the non-terminal nodes of tree T, x_j is the splitting variable used at the non-terminal node t and $\hat{l}(y, f(x))$ is the loss function.[2] Hence, the relative importance of variable x_j is the sum of the improvements in the loss function over all the non-terminal nodes where x_j is used as the splitting variable.[3] Generalizing to M trees in the boosting collection, the relative influence of feature x_j is:

$$\hat{I}_j = \frac{1}{M} \sum_{m=1}^{M} \hat{I}_j(T_m) \tag{13}$$

In this case we select those variables whose relative importance is greater than or equal to 0.1. The final subset of variables will be the union of the ones selected by the two methods.

For the exercise of predicting the type of readmission to the ICU, 90 out of 167 variables were selected for estimation, 55 out of 104 in the case of mortality in the ICU comparable to the APACHE II score, 60 out of 145 for the outcome of mortality as an indicator of hospital quality, and 38 out of 145 in the case of catheter infections at the ICU.

4 Results

In this section we present the estimation results. In the case of ANN and GBM several parameters have to be fixed exogenously. In ANN these are the number of neurons in the hidden layer and the weight decay parameter when predictions pass from one layer to the other. In GBM these are the number of trees, the shrinkage parameter, the number of observations at parent nodes, and the interaction depth that refers to the number of interactions between features. We choose these parameters optimally from a grid of values using 10-fold cross-validation on the training set. The resulting measures are those that maximize the AUC. Table (4) shows the optimal parameters.

[2] Loss misclassifications can be exponential $e^{-yf(x)}$ or binomial deviance $log(1 + e^{-2yf(x)})$ in the case of categorical or binomial dependent variables.

[3] For more details on the relative influence the reader can refer to chapter 10 section 13 in Hastie et al. (2012)

Table 4: Optimal parameters for the ANN and the GBM

	ANN		GBM			
	Size	Decay	N trees	Obs in node	Shrinkage	Interactions
Readmissions	NA	0	500	120	0.05	3
Mortality (Apache)	5	1	1000	30	0.05	2
Mortality (Hospital quality)	5	1	1000	30	0.05	2
Catheter infections	3	1	500	120	0.05	3

4.1 Readmissions

Table 5: Odds ratios of the logit model for readmissions

	Early			Median			Late		
	OR	2.5%	97.5%	OR	2.5%	97.5%	OR	2.5%	97.5%
Demographics									
Male	1.060	0.917	1.224	1.106	1.024	1.195	1.070	0.997	1.148
Age 45-49	1.272	0.921	1.755	1.229	1.033	1.463	1.286	1.101	1.503
Age 50-54	1.329	0.989	1.785	1.062	0.898	1.256	1.272	1.100	1.470
Age 55-59	1.260	0.952	1.667	1.348	1.160	1.566	1.279	1.115	1.466
Age 60-64	1.025	0.763	1.377	1.534	1.330	1.770	1.311	1.147	1.498
Age 65-69	1.821	1.414	2.345	1.635	1.421	1.882	1.360	1.192	1.551
Age 70-74	1.148	0.864	1.525	1.473	1.275	1.703	1.359	1.190	1.553
Age greater than 75	1.261	1.003	1.586	1.209	1.066	1.371	1.123	0.999	1.261
Medical/Hospital									
Apache	1.026	0.930	1.132	1.067	1.022	1.114	1.000	0.957	1.046
Length of stay	0.007	0.003	0.019	1.002	0.976	1.028	0.953	0.809	1.122
Catheter days	1.128	0.831	1.529	1.115	1.044	1.190	1.151	1.086	1.220
Days of bladder catheterization	0.941	0.678	1.307	0.986	0.917	1.060	0.948	0.885	1.017
# central catheters	0.806	0.647	1.005	0.983	0.929	1.040	0.991	0.939	1.047
# arterial lines	0.896	0.798	1.007	1.049	1.007	1.092	0.904	0.862	0.949
Hours of invasive ventilation	0.951	0.667	1.354	0.866	0.806	0.930	1.072	1.022	1.124
RBC transfusion	1.767	1.100	2.839	1.093	0.869	1.375	1.088	0.870	1.360
Patient inflow	1.100	1.026	1.181	1.082	1.041	1.124	0.962	0.927	0.998
Past admission	4.E+09	0.000	NA	5.E+09	0.000	NA	6.E+09	0.000	NA
# of diagnosis	1.043	0.918	1.185	0.948	0.884	1.018	1.052	0.988	1.120
# of proced.	1.082	0.877	1.336	1.001	0.924	1.083	1.036	0.964	1.112
# of monitoring	1.028	0.924	1.143	1.015	0.969	1.063	0.972	0.932	1.014
Adm. Cause surgical	1.223	0.828	1.807	0.906	0.743	1.104	0.858	0.716	1.027
Adm. Cause medical	1.176	0.901	1.535	0.828	0.717	0.956	1.066	0.937	1.213
Same floor origin	1.841	1.044	3.248	1.673	1.257	2.226	1.249	0.961	1.622
Surgery origin	1.428	0.806	2.530	0.977	0.736	1.298	1.038	0.807	1.336
ER origin	1.632	0.993	2.681	1.240	0.969	1.586	1.177	0.946	1.464
Other ICU origin	3.301	1.878	5.800	1.897	1.401	2.569	1.260	0.950	1.671
Fast response origin	2.208	1.119	4.358	1.492	1.030	2.162	1.240	0.871	1.765
Diagnoses									
Rhematic	3.995	2.193	7.279	2.394	1.563	3.668	2.111	1.412	3.156
Neurologic	1.217	0.915	1.619	1.001	0.850	1.180	0.844	0.723	0.984
Trauma	1.104	0.728	1.672	0.597	0.469	0.760	0.381	0.296	0.490
Hepatic	1.777	1.070	2.951	3.245	2.550	4.129	2.641	2.091	3.335
Infections	0.927	0.691	1.244	1.274	1.099	1.478	0.953	0.827	1.098
Respiratory	0.957	0.710	1.288	1.125	0.962	1.316	1.047	0.907	1.208
Metabolic disorder	0.750	0.403	1.398	1.150	0.849	1.556	1.376	1.056	1.794
Gastrointestinal	1.071	0.745	1.540	1.427	1.186	1.716	1.125	0.942	1.344
Renal	0.968	0.689	1.360	1.302	1.090	1.556	1.272	1.079	1.499
Cardiac	1.102	0.830	1.462	1.164	1.002	1.352	1.203	1.051	1.376

Shock	1.061	0.742	1.517	1.074	0.899	1.284	0.857	0.721	1.018
Cancer	1.134	0.799	1.610	1.365	1.143	1.630	1.062	0.895	1.260
Trombosos	1.152	0.704	1.883	1.461	1.123	1.900	1.053	0.814	1.361
Pregnancy	0.361	0.122	1.070	0.323	0.162	0.644	0.128	0.054	0.300
Chronic cardiac risk	1.023	0.792	1.321	1.142	0.994	1.313	1.039	0.916	1.179
Major POP	0.636	0.451	0.896	0.837	0.705	0.995	1.118	0.962	1.300
Procedures									
Dialysis and related	1.399	0.588	3.330	1.672	1.163	2.404	1.478	1.058	2.065
Nutrition	0.374	0.150	0.931	1.200	0.907	1.588	0.872	0.671	1.133
Arrythmia related	0.773	0.219	2.727	0.411	0.189	0.897	0.922	0.564	1.506
Complications									
Infections	1.390	0.737	2.620	0.911	0.728	1.141	1.186	0.970	1.450
Immobilization	0.000	0.000	.	1.235	0.769	1.982	0.729	0.457	1.165
Intoxication	1.129	0.426	2.989	0.173	0.050	0.596	0.204	0.079	0.531
Others									
Admission day Monday	1.073	0.835	1.379	1.053	0.917	1.209	0.962	0.849	1.090
Admission day Tuesday	1.101	0.866	1.400	0.997	0.870	1.141	0.939	0.832	1.060
Admission day Thursday	0.718	0.550	0.938	1.082	0.946	1.236	0.843	0.746	0.953
Admission day Friday	0.935	0.722	1.210	1.084	0.945	1.243	0.863	0.762	0.978
Admission day Saturday	0.935	0.715	1.222	1.036	0.899	1.194	0.915	0.804	1.042
Admission day Sunday	0.972	0.743	1.273	0.891	0.766	1.036	0.878	0.767	1.006

Note: We exclude discharge day of week, discharge month, admission month, and type of insurer because of convergence issues. The base category of the multinomial logit are the non readmitted patients.

Table (5) presents the odds ratios of the multinomial logit model for each outcome: early, median, and late readmissions. Because of convergence issues, we had to exclude the indicators of discharge day of week and month, admission month, and type of insurer. Several things are worth highlighting from these results. On one hand, the literature on readmissions to the ICU is not conclusive on whether a longer stay can be associated to better care, in which case the risk of readmission should be lower, or to worse admission conditions, in which case the risk of readmission may not necessarily be lower. Our results confirm one of these effects: prolonged length of stay is associated to lower risk of early readmissions. However for median and late readmissions the odds ratio is not significant. The effect of other variables reflect the expectations from a medical point of view. Longer catheter days increase the risk of readmission, and this effect is stronger for late than for median and early readmissions. Receiving red blood cell transfusions is also a significant risk factor as well as having past admissions. For past admissions we are unable to report the upper limit of the 95% confidence interval of the odds ratio because of its large standard error. Patients that come from other ICUs are at greater risk of being early, median and late readmitted. Discharge diagnoses that are significantly associated to greater risk of readmission include rheumatic diseases and trombosis. While the ones that are associated to a significantly lower risk of readmission include major surgery post-op and pregnancy.

Figure 6: Relative importance in the boosted trees model for readmissions

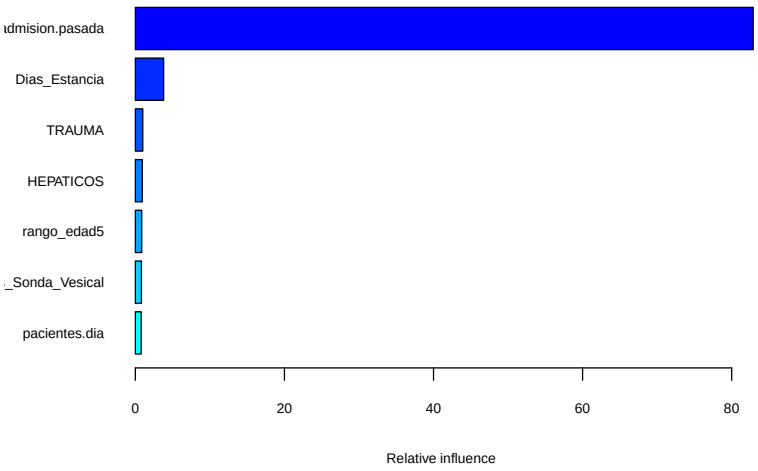
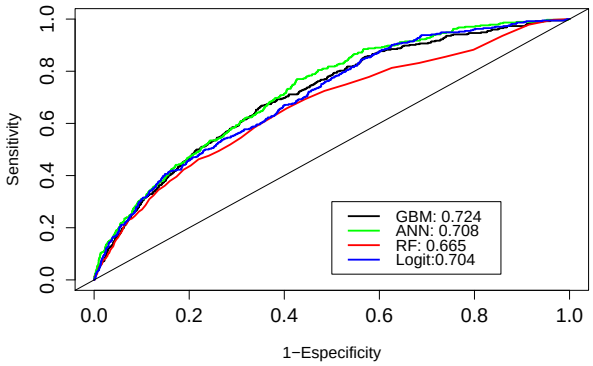


Figure (6) illustrates the seven features with the highest relative importance in the boosted tree model for readmissions. The indicator of past admissions has the highest relative importance followed by the length-of-stay, trauma and hepatic discharge diagnoses, being 19-44 years old, days of bladder catheterization, and daily patient flow.

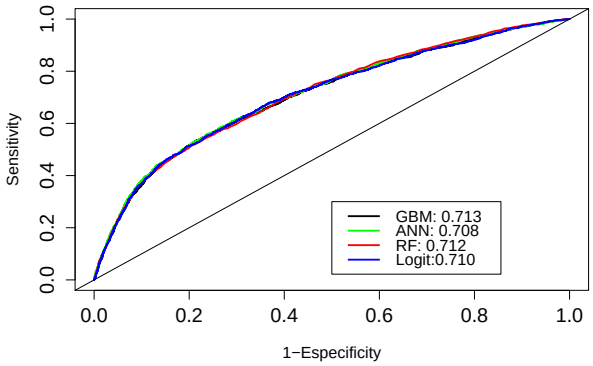
Figure (7) shows the out-of-sample ROC curves of the bagging of ANNs, GBM, RF and multinomial logit models for each type of readmission. To build the curves, first we obtain the matrix of probabilities of dimensions $N \times 4$ predicted by each model (for each type of readmission) and then we compare these probabilities to the observed outcome. The box in each figure shows the area under the curve. In the case of early readmissions there are greater differences between models than for median and late readmissions. The boosting outperforms the bagging of ANNs by 3 percentage point and the RF by 6 percentage points. Moreover, the logit model achieves a greater AUC than the random forest, which suggests there are no significant gains in predictive power when assuming a non parametric joint input variable distribution. This finding holds for the other type of readmissions.

Figure 7: ROC curves for each type of readmission

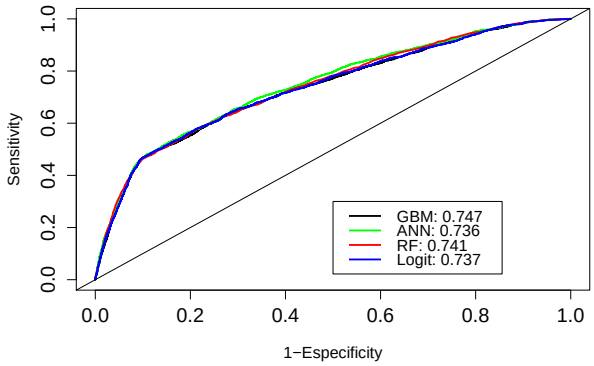
(a) Early readmissions



(b) Median readmissions



(c) Late readmissions



Predictions of late readmissions are more accurate overall compared to predictions of median readmissions (AUC of 75% in late readmissions versus 71% in median readmissions). This probably has to do with the fact we observe more late than median readmissions in the data, therefore the models have more information from the treatment group to obtain relevant patterns.

4.2 Mortality

For the exercise of estimating a model that is comparable to the APACHE II in terms of the features it uses (only those that are measured at the moment of admission), table (6) shows the odds ratios of the logit model and the 95% confidence interval.

Table 6: Odds ratio of the logit model for mortality comparable with APACHE II

	Odds ratio	2.50%	97.50%
Demographics			
Male	1.030	0.951	1.115
Age 18-44	0.855	0.752	0.973
Age 50-54	0.994	0.833	1.182
Age 55-59	1.116	0.950	1.310
Age 65-69	1.101	0.944	1.282
Age 70-74	1.084	0.927	1.267
Age more than 75	1.273	1.121	1.446
Medical/Hospital			
Patient inflow	0.892	0.847	0.939
Adm. Cause surgical	0.788	0.132	15.347
Adm. Cause medical	0.961	0.161	18.679
Surgery origin	0.594	0.465	0.762
Same floor origin	1.343	1.080	1.678
ER origin	1.045	0.867	1.269
Second floor hospitalization	2.198	0.384	42.431
Medical ICU hospitalization	0.319	0.051	6.375
Diagnoses			
Shock	4.683	4.002	5.487
Cardiac	0.518	0.421	0.637
Neurologic	1.807	1.553	2.104
Trauma	2.305	1.907	2.784
Multiple organ failure	5.006	3.542	7.014
Major POP	0.721	0.581	0.895
Infections	1.765	1.504	2.072
Respiratory	1.713	1.451	2.022
Chronic cardiac risk	0.571	0.313	0.959
Hepatic	2.900	2.086	3.977
Pathology of the aorta	2.136	1.541	2.912
Electrolyte imbalance	0.738	0.533	1.002
Cancer	2.202	1.582	3.018
Pregnancy	0.264	0.065	0.706
Others			
Admission day Monday	0.987	0.857	1.138
Admission day Tuesday	0.853	0.740	0.984
Admission day Thursday	0.870	0.753	1.004
Admission day Friday	0.907	0.786	1.047
Admission day Saturday	0.975	0.844	1.125
Admission day Sunday	1.019	0.881	1.179
Admission month January	1.372	1.125	1.674
Admission month February	1.476	1.216	1.790
Admission month March	1.277	1.058	1.542
Admission month April	1.264	1.048	1.524
Admission month May	0.962	0.797	1.160
Admission month June	0.691	0.559	0.850
Admission month July	0.695	0.567	0.851
Admission month August	0.918	0.761	1.107
Admission month October	1.018	0.850	1.218
Admission month November	1.228	1.031	1.463
Admission month December	1.664	1.349	2.050

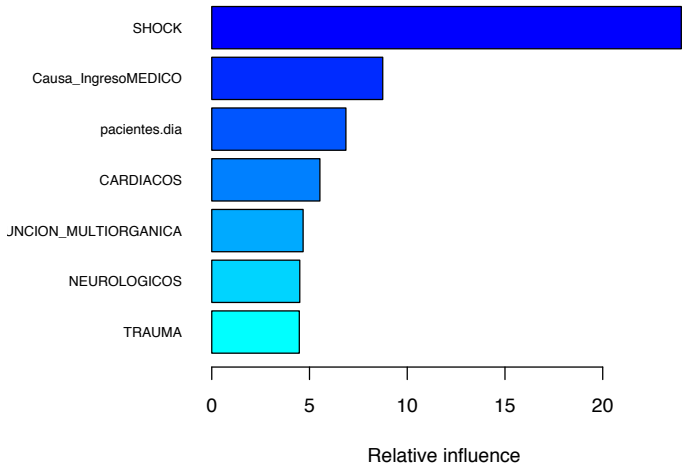
Note: The odds ratios of subsidized regime, contributory regime, capital cities, and department of residence Cauca are not reported.

A patient that is admitted to the ICU with a diagnose of shock has a risk of dying that is four times greater than the risk of patients who are admitted with other diagnoses. Patients with cancer, neurologic diseases, pathologies of the aorta, infections, and respiratory diseases at the moment of admission are also at greater risk of dying at the ICU. On the other hand, pregnancy, major surgery post-op, and chronic cardiac risk at admission are associated to a significantly lower risk of dying.

Figure (8) shows the most relevant variables for the boosted tree model. Similar to what was found

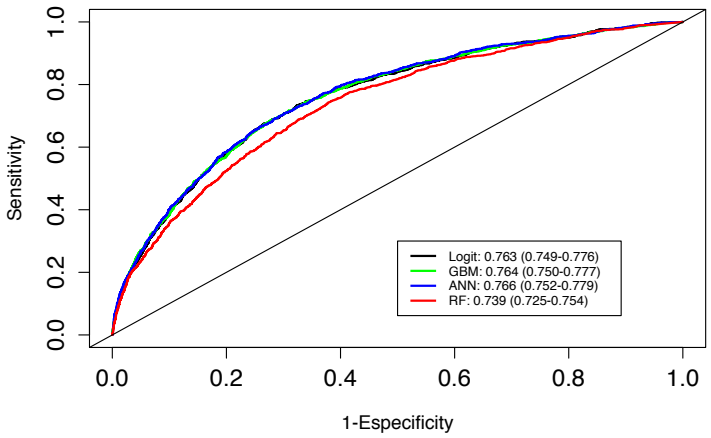
with the logit model, admission diagnoses of shock, neurologic diseases, multiple organ failure, cardiac diseases, and trauma have a relatively high importance. Patient flow is the third most important variable in the boosted tree.

Figure 8: Relative importance in the boosted trees model for mortality comparable to the APACHE II



The out-of-sample AUC of each model with its corresponding 95% confidence interval is reported in figure (9). The RF is out performed by the rest models by nearly 3 percentage points and there are no important differences between the predictive power of the logit, the bagging of ANNs, and the boosted tree. The AUC's upper limit for the bagging of ANNs is 78%. Choosing this model as the best predictor of mortality with variables that are measured at the moment of admission, we compare it with the APACHE II.

Figure 9: ROC curves of each model for the outcome of mortality comparable to the APACHE II



To recover the predicted mortality from the APACHE II in our data we follow figure (3) of the seminal article of Knaus et al. (1985). In this figure, the authors show a correlation between intervals of the APACHE II and the death rate in their original sample of patients. We are unable to use the methodology proposed by the authors to recover the death rate (reported in the appendix of Knaus et al. (1985)) because we lack information regarding the exact admission diagnosis of each patient. However, evidence from this particular hospital shows that computing the individual predicted mortality as suggested is a good baseline to measure quality of care at the ICU. Figure (10) shows the results of the comparison between the predicted mortality of APACHE II and our model. The bagging of ANNs has the same predictive power as the APACHE II, but it is trained with significantly less data. Our model uses information of admission diagnosis, admission date, municipality of residence, age, gender, cause of admission, and origin, while the APACHE II uses indicators of long-term or chronic diseases, emergency surgery or selective surgery, age, physiological variables such as body temperature, heart rate, and respiratory rate, and lab test results such as levels of sodium, potassium, creatinine, white blood cell count, and mean arterial pressure.

Figure 10: Best model vs. APACHE II

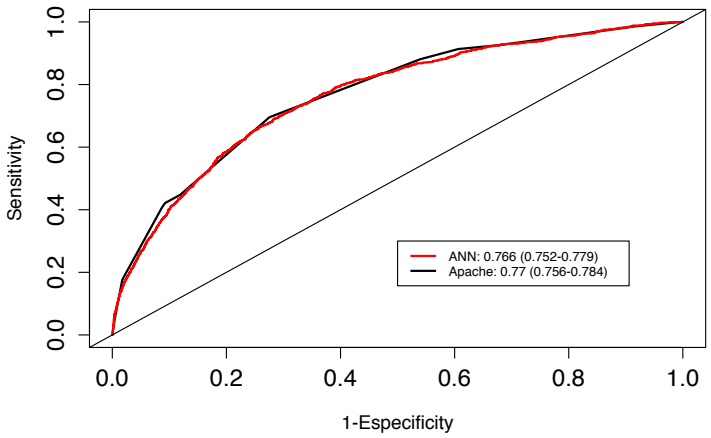


Figure (11) shows the hospital mortality predicted by our model versus the one predicted by the APACHE II and the observed death rate on a monthly basis from 1998 to 2015. According to the anecdotal evidence provided by doctors at the ICU, the tendency of the predicted mortality of APACHE II in relation to the observed mortality is as expected. The deviation of the observed mortality from the one predicted by the APACHE II reflects an improvement of the quality of care at the unit. Nonetheless, such deviations also suggest the APACHE II is not accurate for predicting an overall hospital death rate, while our model is.

Figure 11: Evolution of predicted and observed mortality

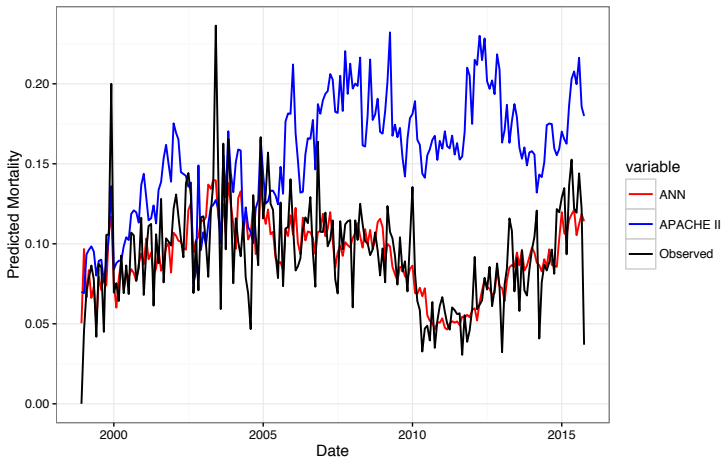


Table 7: Odds ratio of the logit model for the outcome of mortality as a quality indicator

	Odds ratio	2.50%	97.50%
Diagnoses			
Shock	4.007	3.444	4.667
Major POP	0.683	0.554	0.842
Cardiac	0.609	0.502	0.739
Neurologic	1.875	1.617	2.175
Multiple organ failure	4.646	3.305	6.481
Trauma	2.244	1.861	2.703
Infections	1.742	1.497	2.028
Hepatic	3.003	2.194	4.064
Respiratory	1.756	1.501	2.055
Cancer	2.537	1.867	3.407
Pathology of the aorta	2.047	1.477	2.791
Pregnancy	0.245	0.060	0.660
Electrolyte imbalance	0.748	0.543	1.009
Complications			
Infections	0.909	0.765	1.076
Respiratory	1.127	0.789	1.588
Catheter related	1.165	0.872	1.545
Immobilization	1.566	1.143	2.121
Procedures			
Dialysis and related	3.021	2.474	3.681
Nutrition	1.033	0.899	1.185
Respiratory invasive	0.826	0.679	1.000
Cardiac invasive	2.226	1.556	3.143
Monitoring			
Invasive monitoring	2.848	2.577	3.148
Other	1.913	1.332	2.716
Medical/Hospital			
Adm. cause medical	0.742	0.128	14.205
Adm. cause surgical	0.535	0.092	10.258
Adm. cause cardiology	0.503	0.086	9.658
Surgery origin	0.480	0.397	0.580
Same floor origin	1.061	0.907	1.241
ER origin	0.837	0.748	0.937
Medical ICU hospitalization	0.660	0.114	12.762
Second floor hospitalization	3.068	0.587	57.480
Patient inflow	0.873	0.830	0.918
Demographics			
Male	1.027	0.951	1.110
Age 19-44	0.802	0.717	0.896
Age 50-54	0.964	0.823	1.126
Age 60-64	0.958	0.831	1.101
Age 70-74	1.025	0.891	1.177
Age greater than 75	1.248	1.120	1.390

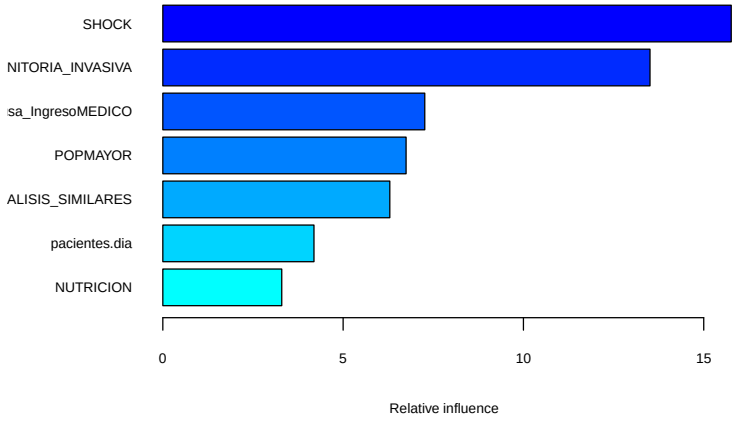
Note: The odds ratios of admission day of week and month, subsidized regime, contributory regime, enterprises, capital cities, and department of residence Cauca are not reported.

If we allow the models to use the subset of predictors known during the patient's stay we move towards a mortality score that serves as a hospital quality indicator, but it is no longer comparable to the APACHE II. Table (7) shows the odds ratios of the logit model with this new subset of variables. Several effects remain significant from the previous exercise. For example, admission diagnoses of shock, multiple organ failure, trauma, infections, neurologic diseases, cancer, respiratory diseases, and pathologies of the aorta increase the risk of dying at the ICU almost twice. In particular a patient that enters the ICU with shock is 4 times more likely to die than a patient that enters the ICU with a diagnose different from the ones reported. Patients who are immobilized at the ICU are 50% more likely to die and patients who receive dialysis or related procedures are 3 times more likely to die than patients who do not receive procedures. Cardiac invasive procedures and invasive medical monitoring such as intrabdominal pressure and intracranial pressure also increase the risk of

dying at the ICU in almost 120-180%.

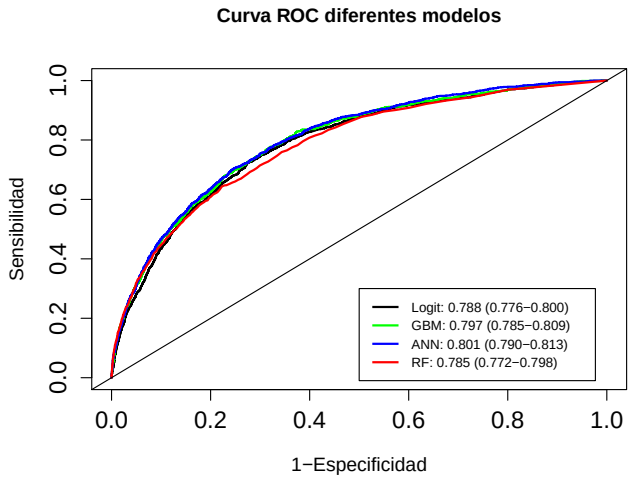
Figure (12) shows the seven most important predictors in the boosted tree model. Shock diagnosis and invasive monitoring are the two features that reduce the most the loss function, followed by the indicator of medical admission cause, major surgery post-op, receiving dialysis and related procedures, daily patient flow, and receiving enteral or parenteral nutrition.

Figure 12: Relative importance in the boosted trees model for mortality as an indicator of hospital quality



The out-of-sample classification power of each model measured as the AUC and its 95% confidence interval is illustrated in figure (13). There are no differences between the models since their confidence intervals overlap, but notice the AUC in this exercise is greater than the AUC in the previous exercise, meaning feature creation and inclusion improves predictions. The best model for predicting mortality at the ICU is the bagging of ANNs, which reaches an AUC of 80% with a standard deviation of 1 percentage point.

Figure 13: ROC curves of each model for the outcome of mortality as a hospital quality indicator



4.3 Catheter infections

This subsection shows the results of predicting catheter infections at the ICU. This exercise can be considered as one of predicting complications or endangering states at the ICU. Table (4.3) shows the odds ratios of the logit model. In this case, variables that increase significantly the risk of getting a catheter infection include: catheter days, number of Swan Ganz catheters, APACHE II score, number of complications and procedures, and being more than 60 years old.

Additional risk factors stand out from the relative importance in the boosted tree model (figure 14). These include: hours of invasive ventilation, length of stay, days of bladder catheterization, and daily patient flow.

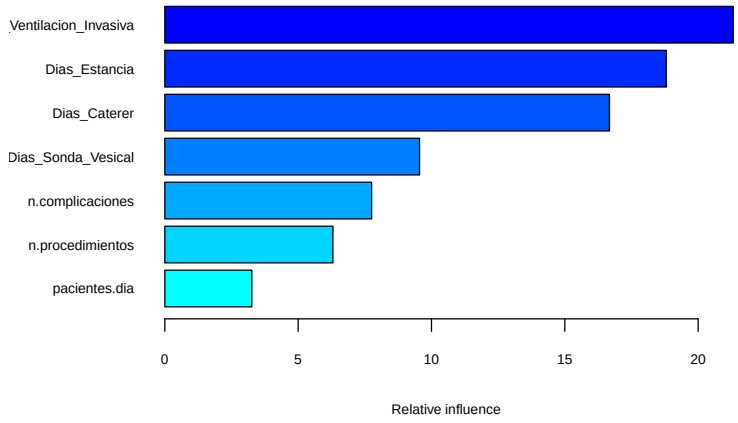
Table 8: Odds ratios of the logit model for catheter infections

	Odds ratio	2.50%	97.50%
Diagnoses			
Infections	1.421	0.971	2.043
Complications			
Respiratory	0.544	0.258	1.078
Immobilization	0.828	0.473	1.418
Procedures			
Respiratory invasive	0.864	0.514	1.455
Intubation	0.428	0.224	0.775
Nutrition	1.184	0.667	2.095
Dialysis and related	0.688	0.382	1.220
Medical/Hospital			
Length of stay	1.029	0.981	1.054
Catheter days	1.180	1.063	1.309
Horas ventilación invasiva	1.000	0.927	1.071
Days of bladder catheterization	0.918	0.814	1.031
# of Swan Ganz	1.049	1.007	1.112
Apache	1.208	1.079	1.315
# central catheter	1.007	0.910	1.052
# arterial lines	1.017	0.946	1.077

Catheter days > perc(95)	1.303	0.788	2.142
Length of stay > perc(95)	7.049	4.516	10.950
Patient inflow	0.919	0.770	1.091
# of complications	1.201	1.117	1.291
# of procedures	1.283	1.090	1.503
# of monitoring	1.004	0.869	1.158
Adm. cause surgical	0.710	0.506	0.991
Adm. cause cardiology	0.888	0.569	1.356
Other ICU origin	1.450	0.936	2.186
Demographics			
Age 19-44	1.128	0.800	1.575
Age 45-49	1.361	0.780	2.255
Age 60-64	1.641	1.067	2.457

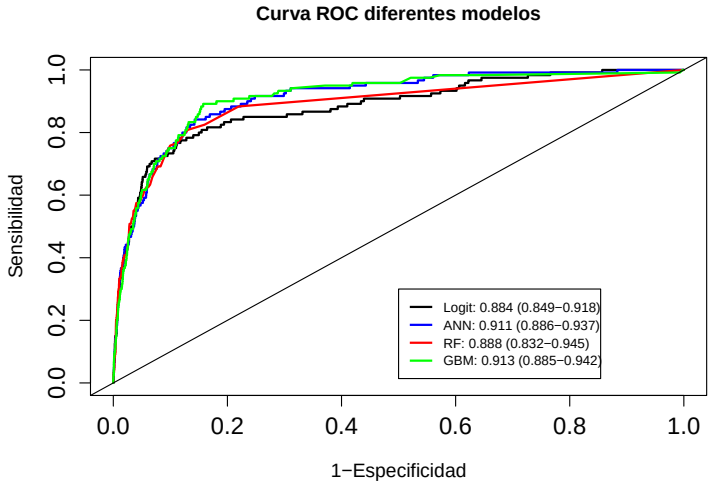
Note: The odds ratios of admission day of week and month, and departament of residence Caldas, are not reported.

Figure 14: Relative importance in the boosted trees model for catheter infections



Finally, figure (15) presents the out-of-sample ROC curves of the models plus the bagging of ANNs and the RF. Once again there are no important differences in the predictive power of each model, but compared to the other exercises in this application, in this case we reach an AUC of 91% with the boosted tree model, with a standard deviation of 3 percentage points.

Figure 15: ROC curves of each model for the outcome of catheter infection



5 Comparison with the existing literature

The two main goals of this paper are to provide health management tools that aid at patient care at the ICU and to find risk factors for our three outcomes that are significant from a medical point of view. To prove that the machine learning methods used in this study accurately predict the probability of being readmitted, dying, or getting a catheter infection at the ICU, and therefore are potential tools to be implemented at the hospital from where our data comes from and similar hospitals, we compare our results to the related literature.

Table (9) shows the AUC of some of the studies that predict certain types of readmission. The majority of them use the logit model with predefined variables, while Goulart et al. (2015) add stepwise backward selection in the logit to choose the features that closely correlate to the event of readmission. In the field of machine learning and data mining, Fiahlo et al. (2012) use forward selection in decision trees to choose the relevant features and then fuzzy modelling (association rules) to predict ICU readmission. The highest AUC is reported for Ferreira et al. (2014) and Jo et al. (2015), 76% each, followed by Goulart et al. (2015) and Ouanes et al. (2012), 74% each. These studies use information related to lab test variables such as white and red blood cell count, creatinine levels, etc. and physiological variables such as weight, size, heart rate, respiratory rate, and body temperature, all of which we lack. However we have a larger sample size compared to all of these studies. Despite having less information regarding the patient's morbidity, the machine learning methods of our study equal the classification power in the related literature, but compared to the studies that use a similar set of predictors (Campbell et al. (2008)) we outperform their results.

Table 9: 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

Table (10) shows the AUC of several studies related to death or mortality at the ICU. Sujin et al. (2011) estimate decision trees, ANNs, and support vector machines, the first being the best predictor. Also in the line of data mining algorithms, Ramon et al. (2007) use decision trees, random forests, naive bayes, and tree augmented naive bayes to predict mortality, the latter being the best predictor. On the other hand, the rest of studies showed in the table use logit models with predefined features to analyze the risk of death at the ICU.

Compared to our application of the mortality score as a quality indicator, our best model (Bagging of ANNs with an AUC of 80%) is outperformed by three of the reported studies (Badawi and Breslow (2012), Sujin et al. (2011), Ramon et al. (2007)). Nonetheless, as mentioned previously, these authors had more information regarding the patient's health status. To name a few, Badawi and Breslow (2012) use the following predictors in their model: body mass index, acidosis, alkalosis, creatinine levels, white blood cell count, serum glucose levels, mean arterial pressure, respiratoty rate, heart rate, etc.

Table 10: Classification in the literature of mortality

Autor	Outcome	AUC
Badawi and Breslow (2012)	Death after discharge	0.92
Campbell et al. (2008)	Death after discharge	0.74
Sujin et al. (2011)	Death after discharge	0.89
Ramon et al. (2007)	Survival after discharge	0.88
Ouanes et al. (2011)	Death within 7 days	0.74

Although there is very few literature concerned with predicting catheter infections at the ICU, this outcome could be related to the literature of developing endangering states. Table (11) presents the AUC of several papers dealing with the prediction of sepsis and inflammations at the ICU. All of them expect for Ramon et al. (2007) use the logit model to predict the outcome. Most studies find that the levels of procalcitonin and C-Reactive Protein (CRP) are highly predictive of endangering states. For example, Balcl et al. (2003) have an AUC of 97% using procalcitonin as a predictor of sepsis and Miller et al. (1999) reach an AUC of 86% using CRP as a predictor of systemic inflammatory response syndrome. Even though we have no information regarding procalcitonin or CRP, our best predictor of catheter infections at the ICU (Boosted tree) has an AUC of 91% out of sample.

Table 11: Classification in the literature of endangering states

Autor	Outcome	AUC
Ramon et al. (2007)	Severe inflamation	0.84
Ramon et al. (2007)	Shock by inflamation	0.93
Moreno et al. (2008)	Death by sepsis	0.77
Balcl et al. (2003)	Sepsis (due to procalcitonin)	0.97
Ugarte et al. (1999)	Sepsis (due to CRP)	0.78
Miller et al. (1999)	Inflammatory response syndrome (due to CRP)	0.86

6 Conclusions and further research

This article borrows from the techniques and methods of machine learning to predict different types of readmissions, mortality, and catheter infections at the adult intensive care unit of a high complexity hospital in Colombia. To the best of our knowledge, it is the first time machine learning is applied in the field of health care in this country. This article serves two purposes: one is estimating models that can aid doctors in optimally assigning resources and efforts to the patients at the ICU, in other words, helping doctors manage health care; the other is to find risk factors for each outcome that are relevant from a medical point of view. To achieve the latter we include the logit in the menu of machine learning methods in order to obtain interpretable odds ratios for each predictor.

Bagging of artificial neural networks and boosted trees were the most predictive models for all outcomes. In the case of predicting early, median, and late readmissions we achieve an AUC of 72, 71 and 75%, respectively. In the case of predicting mortality we develop two exercises: one is intended to compete against the APACHE II scores and the other one is thought of as a hospital quality indicator. The difference between both relies in the subset of predictors in which they are trained. The first is restricted to variables that are measured at the moment of admission while the second is able to use information gathered during the patient's stay. Using far less information about the patient, we manage to equalize the predictive power of the APACHE II in our validation set (AUC of 76%). For the second exercise we reach an AUC of 80%. Although this measure is outperformed by most of the literature regarding mortality at the ICU, our set of predictors is much more limited in the sense we do not have the patient's physiological information nor lab tests results. Finally we estimate several models to predict catheter infections at the ICU. This exercise can be related to the literature of developing endangering states. In this case the boosted tree reaches an AUC of 92% out-of-sample.

In terms of the second objective, we find several risk factors that are transversal to all outcomes. Admission diagnoses such as shock, infections, pathologies of the aorta, cancer, neurologic diseases, and respiratory diseases increase the risk of readmission, death, and catheter infections at the ICU. Invasive procedures and monitoring such as dialysis, tracheostomy, and bronchoscopy are also positively correlated to the probability of observing these outcomes. Unlike most of the literature, gender and age had no significant correlation with the relative risk of being readmitted, dying, or getting a catheter infection in the context of this particular ICU in Colombia. Our results also confirm one of the findings in the literature of critical care, this is the fact that the level of occupation of the ICU is detrimental for patient health care.

In the case of catheter infections, results may be biased because developing such endangering state may depend on the attention the patient received before his admission to the ICU rather than on the

attention received exclusively at the ICU. Hence, risk factors associated to catheter infections may be capturing the quality of previous health care. Further research can be done in terms of predicting the development of endangering states. For example, nosocomial pneumonia is more proper to measure the quality of care at the ICU than perhaps catheter infections. Also, further research can be done in terms of the computation of out-of-sample AUC using cross-validation, or building the validation set from randomly selecting patients rather than randomly selecting admissions as done in this application.

References

Badawi, O. and Breslow, M. (2012). Readmissions and Death after ICU Dsicharge: Dvelopment and Validation of Two Predictive Models. *PLOS ONE Open access*, 7(11):1–15.

Balcl, C., Sungurtekin, H., Gurses, E., Sungurtekin, U., and Kaptanoglu, B. (2003). Usefulness of procalcitonin for diagnosis of sepsis in the intensive care unit. *Critical Care*, 7(1):85–89.

Bayati, M., Braverman, M., Gillam, M., Mack, K., Ruiz, G., Smith, M., and Horvits, E. (2014). Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study. *PLOS ONE Open access*, 9(10):1–9.

Benetis, R., Sirvinskas, E., Kumpaitiene, B., and Kinduris, S. (2013). A case-control study of readmission to the intensive care unit after cardiac surgery. *International medical journal of experimental and clinical research*, 19:148–152.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1983). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Brown, S. and Ratcliffe, S. (2012). The Epidemiology of Intensive Care Unit Readmissions in the United States. *American Journal of Respiratory and Critical Care Medicine*, 185(9):955–964.

Buchner, F., Wasem, J., and Schillo, S. (2015). Regression trees indentify relevant interactions: can this improve the predictive performance of risk adjustment? *Health Economics*.

Campbell, A., Cook, J., Adey, G., and Cuthbertson, B. (2008). Predicting death and readmission after intensive care discharge. *British Journal of Anaedthesia*, 100(5):656–662.

Ferreira, D., Kras, I., Forgiarini, L., and Rieder, M. (2014). Assessment of icu readmission risk with the Stability and Workload Index for Transfer Score. *Journal Brasileiro de Pneumologia*, 40(1):73–76.

Fiahlo, A., Cismondi, F., Vieira, S., Reti, S., Sousa, J., and Finkelstein, S. (2012). Data Mining Using Clinical Physiology At Discharge To Predict ICU Readmission. *Expert Systems with Applications*, 39(18):13158–13165.

Gajic, O., Malinchoc, M., Comfere, T., Harris, M., Achouiti, A., Yilmaz, M., Schultz, M., Hubmayr, R., Afessa, B., and Farmer, C. (2008). The Stability of Workload Index for Transfer Score predicts unplanned intensive care unit patient readmission: initial development and validation. *Critical Care Medicine*, 36(3):676–682.

Goulart, R., Roehrig, C., Pinheiro, R., Gasparetto, J. Pecanha, A., de Souza, P., Dexheimer, F., Balzano, P., and Teixeira, C. (2015). Comparison of Unplanned Intensive Care Unit Readmission Scores: A Prospective Cohort Study. *PLOS ONE Open access*.

Greene, W. (2012). *Econometric Analysis*, volume 1. Prentice Hall, 7 edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2012). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

Jo, Y., Lee, Y., Park, S., Yoon, H., Lee, J., Lee, C., and Cho, Y. (2015). Readmission to Medical Intensive Care Units: Risk Factors and Prediction. *Yonsei Medical Journal*, 56(2):543–549.

Knaus, W., Draper, E., Wagner, D., and Zimmerman, J. (1985). APACHE II: A severity of disease classification. *Critical Care Medicine*, 13(10):818–829.

Kogan, A., Cohen, J., Raanani, E., Sahar, G., Orlov, B., Singer, P., and Vidne, B. (2003). Readmission to the Intensive Care Unit After "Fast-Track" Cardiac Surgery: Risk Factor and Outcomes. *The Annals of Thoracic Surgery*, 76(2):503–307.

Kramer, A. and Higgins, T. (2012). Intensive care unit readmission in US hospitals: Patient characteristics, risk factors, and outcomes. *Critical Care Medicine*, 40(1):3–10.

Litmathe, J., Feindt, K., and Boeken, G. (2009). Predictors and Outcome of ICU Readmission after Cardiac Surgery. *The Thoracic and Cardiovascular Surgeon*, 57(7):391–394.

Miller, P., Munn, D., Meredith, J., Wayne, M., and Chang, M. (1999). Systemic Inflammatory Response Syndrome in the Trauma Intensive Care Unit: Who is Infected? *Journal of Trauma-Injury Infection and Critical Care*, 47(6):1004.

Moreno, R., Metnitz, B., Adler, L., Hoechtl, A., Bauer, P., and Metnitz, P. (2008). Sepsis mortality prediction based on predisposition, infection and response. *Intensive Care Medicine*, 34:496–504.

Ouanes, I., Schwebel, C., Francais, A., Bruel, C., Philippart, F., Vesin, A., Soufir, L., Adrie, C., Garrouste-Orgeas, M., Timsit, J., and Misset, B. (2011). A model to predict short-term death or readmission after intensive care unit discharge. *Journal of Critical Care*, 27(4):422.E1–9.

Ramon, J., Fierens, D., Guiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M., and Van Den Berghe, G. (2007). Mining data from intensive care patients. *Advanced Engineering Informatics*, 21(3):243–256.

Sujin, K., Woojae, K., and Rae, W. (2011). A comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. *Healthcare Informatics Research*, 17(4):232–243.

Ugarte, H., Silva, E., D., M., Mendoza, A., and Vincent, J. (1999). Procalcitonin used as a marker of infection in the intensive care unit. *Critical Care Medicine*, 27(2):498–453.