Contributions from Computational Intelligence to healthcare data processing

By

Asier Garmendia Mujika

Submitted to the department of Computer Science and Artificial Intelligence in partial fulfillment of the requirements for the degree of Doctor of Philosophy

PhD Advisors:

Prof. Manuel Graña Romay
Dr. Jose Manuel López Guede

at The University of the Basque Country (UPV/EHU)

Euskal Herriko Unibertsitatea
Universidad del País Vasco
Donostia - San Sebastián
2016

(c)2016 ASIER GARMENDIA MUJICA

Acknowledments

El camino hasta aquí parecía largo, y así ha sido. Y todo camino, por lo general, y como en la vida misma, tiene tramos fáciles y difíciles, y momentos más y menos agradables. De todo se aprende, y con ello me quedo. Pero quiero hacer un reconocimiento a quienes me han ayudado a hacer este trayecto posible y más llevadero.

Por ello, quiero agradecer, para empezar, a Manuel Graña y José Manuel López por la paciencia que han tenido conmigo, y por su ayuda tanto desde el punto de vista meramente técnico como desde el humano.

Mi agradecimiento también a Julián Estévez, sobre todo por su aportación determinante en un momento crucial del trayecto. Y, desde un punto de vista más cercano en lo cotidiano, a las personas que me han escuchado, aconsejado y, en definitiva, ayudado, sobre todo en los momentos de dificultades, como, por ejemplo, cuando los nubarrones parecía que presagiaban un final del camino muy distinto al deseado. Sin ellos tampoco hubiera podido conseguirlo. Quiero agradecérselo a mis amigos y familiares, pero en especial a mi hermano Mikel, y a mis amigos Xabier y Joseba.

Y llegado hasta aquí………… el camino sigue.

Asier Garmendia Mujika

Contributions from Computational Intelligence to healthcare data processing

by
Asier Garmendia Mujjka

Submitted to the Department of Computer Science and Artificial Intelligence on November 21th, 2016, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

The increasing ability to gather, store and process health care information, through the electronic health records and improved communication methods opens the door for new applications intended to improve health care in many different ways. Crucial to this evolution is the development of new computational intelligence tools, related to machine learning and statistics. In this thesis we have dealt with two case studies involving health data. The first is the monitoring of children with respiratory diseases in the pediatric intensive care unit of a hospital. The alarm detection is stated as a classification problem predicting the triage selected by the nurse or medical doctor. The second is the prediction of readmissions leading to hospitalization in an emergency department of a hospital. Both problems have great impact in economic and personal well being. We have tackled them with a rigorous methodological approach, obtaining results that may lead to a real life implementation. We have taken special care in the treatment of the data imbalance. Finally we make propositions to bring these techniques to the clinical environment.

Contents

1. Introduction **1**

1.1 Motivation ……………………………………………….. 1

1.1.1 Triage prediction in respiratory pediatric intensive care units ……………………………………. 2

1.1.2 ED Readmission prediction ………………….. 4

1.2 Objectives …………………………………………… 5

1.3 Contributions ……………………………………….. 6

1.4 Contents of the thesis ………………………………. 7

1.5 Publications achieved ………………………………. 8

1.5.1 Main publications in the work of the thesis ….. 8

1.5.2 Publications achieved as a collaboration indirectly related to the Thesis ……………………... 9

2. Literature review **11**

2.1 Data processing and Machine Learning ……………. 11

2.2. Children respiratory disease monitoring …………… 14

2.3. Readmission risk prediction approaches …………… 15

2.4. Figures ……………………………………………… 20

3. Prediction of respiratory crisis **21**

3.1 Introduction ………………………………………… 22

3.2. Description of the dataset ………………………….. 24

CONTENTS

3.2.1. Descriptive statistics of the dataset ………….. 25
3.2.2 Data preprocessing …………………………… 26
3.2.3. Approaches to the normalization of the age dependent variables ………………………………... 26

3.3. Experimental design ………………………………. 27

3.4. Results ……………………………………………...

28

3.5. Conclusions ………………………………………… 29

3.6. Tables and figures …………………………………. 31

4. Prediction of hospitalization after readmission 37

4.1 Introduction ………………………………………… 38

4.2 Dataset and classification features …………………. 39

4.3 Experimental design ……………………………….. 42

4.4 Computational results ……………………………… 43

4.5 Conclusions ………………………………………… 44

4.6 Tables and figures ………………………………….. 46

5. Conclusions and ideas for future work 59

5.1Processing health care data ………………………… 59

5.2 Monitoring of pediatric respiratory crisis …………. 61

5.3 Prediction of readmission and hospitalization …….. 61

Appendix A. Classifier training methods 63

A.1 Methods …………………………………………... 63

Appendix B. Class balancing techniques 65

B.1 Class Balancing …………………………………… 65

References 67

List of Figures

2.1 Deaths from chronic diseases in Chile in the period 2000 - 2012 [DEPT] …………………………………… 20

2.2 Deaths from respiratory diseases in Chile in the period 2000 to 2011[DEPT] …………………………………… 20

3.1 Frequency distribution of the variable age of patients … 34

3.2 Scatter plots of the Heart Rate versus Age for each triage value …………………………………………………….. 35

3.3 Scatter plots of the Respiratory Frequency versus Age for each triage value ………………………………………… 35

3.4 Scatter plots of the Systolic Blood Pressure versus Age for each triage value …………………………………….. 36

3.5 Scatter plots of the Diastolic Blood Pressure versus Age for each triage value …………………………………….. 36

4.1 Percentage of patients suffering readmission, and percentage of readmission events relative to the number of patients and events, respectively ……………………… 55

4.2 Decomposition of the trends for patients (fig. 4.2a) and events (fig. 4.2b) of the readmission of adults and pediatric according to the readmission threshold ………… 56

LIST OF FIGURES

4.3 Distribution of the number of readmission leading to hospitalization according to the threshold for readmission.. 56

4.4 Distribution of the number of readmission leading to hospitalization according to the threshold for readmission. (a) adult population, (b) pediatric population …………… 57

4.5 Distribution of the dataset into training and test datasets in order to avoid circular effects and biasing of test results by training data misuse …………………………... 58

List of Tables

3.1 Feature selection experimental design ……………… 31

3.2 Average accuracy achieved by each feature
Independently ………………………………………. 32

3.3 Average accuracy for each of the feature sets in
Table 3.1. ………………………………………….. 33

4.1 Distribution of the most salient motives for readmission
to the ED, for various readmission thresholds ……… 46

4.2 Distribution of the most salient motives for readmission
of adults to the ED, for various readmission thresholds . 47

4.3 Distribution of the most salient motives for readmission
of pediatrics to the ED, for various readmission thresholds 48

4.4 Distribution of the most salient motives for readmission
to the ED that lead to hospitalization, for various
readmission thresholds ……………………………… 49

4.5 Distribution of the most salient motives for readmission
of adults to the ED that lead to hospitalization, for various
readmission thresholds ……………………………… 50

4.6 Distribution of the most salient motives for readmission of pediatric patients to the ED that lead to hospitalization, for various readmission thresholds ………………… 51

4.7 Average performance of the data balancing methods across classifier training and readmission threshold ………. 52

4.8 Average performance of the classification training methods across learning methods and readmission threshold ……………………………………………. 53

4.9 Average performance of the readmission threshold across learning methods and classifiers …………………… 54

Chapter 1
Introduction

This Chapter provides the introduction to the Thesis giving its motivation, objectives and actual contributions. The Chapter is intended for a quick assessment of the Thesis content and the merits of the doctoral candidate, so it contains additionally the list of publications achieved.

1.1 Motivation

The present Thesis deals with the use of machine learning tools towards the improvement of health care by predicting specific conditions. The increasing facility to gather data and to analyze it, with solidly established methodologies and techniques has fostered this kind of approaches. However, this work does not deal with big data per se, because the data sets that we have exploited, though bigger than academic toy problems, are far from being considered big data. Another trend that we are not following is the cloud based processing of the data [ABU12] [SUL14] [STA14]. In essence, cloud computing poses many issues regarding data protection that cannot be risked when dealing with health data [ROD13], though it can be seen as an empowerment in isolated rural areas [LIN14] achieving ubiquitous healthcare [HE13], or an easy way to collaborate [LAI12]. The fact is that intelligent computing is becoming increasingly used to create predictive models in many areas of medicine and health-care [PAN09]. Data mining as a specific aspect of intelligent processing bordering statistics is very relevant [HAN11]. In this Thesis we are concerned with the quantitative measures gathered in electronic health records, so we do not need to deal with the already difficult problem of extracting information from free style written reports [NEU14], requiring text processing and natural language techniques.

1.1.1 Triage prediction in respiratory pediatric intensive care units

This manuscript details a study based on data obtained monitoring children hospitalized in the pediatric intensive care unit because of respiratory complications in the city of Santiago de Chile. Respiratory diseases have an increasing prevalence in large urban concentrations of the world, due to the apparently unstoppable increase of air pollution originated from a diversity of sources. Children are specially a fragile part of the population suffering these conditions. Improved monitoring of critical patients by means of automatized data gathering and processing, i.e. alarm raising, aims to alleviate the risks of critical patients. Pediatric respiratory critical care has not received much attention in the literature; most of it is devoted to adult patients suffering specific degenerative conditions. However, children care has specific conditions, such as the strong dependence of some physiological signals, i.e. heart rate, on the patient age. We approach the problem as triage prediction problem, formulated as multi-class classification problem, taking special care in the age normalization of physiological variables. Available data for use as classification features is scarce, in the sense that only a few variables are available, and that much of the qualitative information used by the medical doctors is not available. In this Thesis we report the experimental results obtained on a data sample covering patients of two years provided by a local pediatric hospital. The results conclude that it is possible to some extent to predict the triage that the medical doctors will assign the critical patients. However, we have also detected that medical doctors follow very conservative policies, i.e. taking into consideration the previous state allows almost perfect prediction. We consider these results preliminary steps towards a more comprehensive automation of the pediatric intensive care units, especially regarding the automatic raising of some flags when critical situation appears. Timeliness of the alert may be a life or death difference. Working with actual local hospital data in the design of the alerts is quite important, because studies made in countries like the USA may be biased by local customs, from food to exercise habits. Also, equipment available in one country may not be available in another, or in a different level hospital budget. Finally, local environmental conditions may invalidate conclusions taken from another environment. For instance, the local features of the pollution in Santiago, chemical contaminants and kind of particles, may be quite different from other cities. Even more, the specific epidemic conditions in Santiago may be irreproducible in other countries/towns, i.e. there is no high prevalence of respiratory diseases in children of the Basque Country.

1.1.2 ED Readmission prediction

Emergency Departments (ED) suffer heavy overload due to lack of primary attention service. Increasingly geriatric admissions pose specific problems contributing to this overload. A consequence is the increase of patient returning short time after discharge, i.e. readmissions, sometimes requiring hospitalization. In this latter case the patient problem was not solved in the first admission and its condition has aggravated. Therefore, the economic issues involved in the administration of the readmission event are worsened by the unsolved or aggravated health problem. Readmission has been tackled in many aspects by health care systems, and for various specific populations, each having specific features that impede porting solutions from one domain to another. Take for example the definition of the readmission threshold, i.e. the specific time interval within which the next visit of a patient will be considered a readmission. The insurance companies in the USA have set a threshold of 30 days for economical reason for the general hospitalization of patients. However, 30 days is a long term horizon in ED, where critical situations must be solved in a matter of minutes. For this reason, some countries, i.e. Chile, have adopted a threshold of 3 days to declare ED readmissions.

According to these readmission threshold variations, we have considered several such thresholds in our prediction experiments. Prediction of hospitalization following ED readmission is posed over a heavily imbalanced class distribution, so we have considered several approaches to deal with class imbalanced datasets and several base classifiers, as well as performance measures that enhance the critical comparison between approaches. Experimental works are carried out on real data from a university hospital in Santiago, Chile, corresponding to the period between January 2013 and April 2016, including pediatric and adult admissions to the ED. We achieve results that encourage the development of real life application of the data balancing and classification approach for prediction of hospitalization after readmission.

1.2 Objectives

The general objective of the Thesis is to tackle health care data problems from a Computational Intelligence perspective, building predictors that solve the problems stated in a machine learning framework. Mostly we formulate the problems as classification problems. We do not deal with regression problems for forecasts.

The general approach is data driven, i.e. a problem is defined by the available data, and models are induced from the data, instead of extracting knowledge rules from the expertise of the medical doctors.

The specific objectives are related to the specific problems attacked. We formulate each problem as a classifier learning problem; hence the classification performance is the measure of success. We assess the classification performance via cross-validation, obtaining estimations of the expected performance.

For the respiratory crisis monitoring, we look for the maximal accuracy predicting the case of transition to the worst state, equivalent to raising an alarm to notify the nurse that some critical situation is about to happen.

For the readmission prediction, we try to predict the event of the hospitalization of a readmitted patient, because it is a worst case scenario, where the medical condition has worsened dramatically, and it was not completely solved in the first visit.

From the operational point of view, to carry out the experiments we needed to set the working environment, which was based on the R language and packages available, and preprocess the data in order to be able to use them for training and testing. Data preprocess was different in each case, filling missing values, correcting manually some variable values, or discarding records with too many errors. Also we ensured that the data was fully anonymized.

In the readmission case, the data set was not small, so managing needed to carefully tap on the resources of the R environment and the available computing power. Therefore, managing data consistently was an additional operational objective.

Data classes were strongly imbalanced, especially in the case of the readmission prediction. Therefore, we had to deal with this imbalance treating the datasets before training, and ensuring that the manipulation of the training data did not contribute any bias to the testing phase.

The goal from the real life application point of view is to make recommendations about the actual classifier learning technique as well as which preprocessing (i.e. class balancing) is more appropriate for the data at hand.

1.3 Contributions

1. We have tackled two non-trivial realistic problems in the domain of health data processing, using real life datasets provided by hospitals in Santiago, Chile. Pending permission from the hospitals, we plan to publish the anonymized data to allow reproducibility of our results, or development of advanced solutions.

2. We provide a literature review and state of the art of each of the topics covered by the thesis.
3. We report results on the performance of the tested classifiers on the respiratory intensive care unit, identifying the most appropriate classifier training strategy. We provide recommendations for a real life implementation.
4. We report results on the performance of both the tested classifiers and the class balancing method tested to improve sensitivity of the classifiers. We provide recommendations on the class balancing method, and the classifier to be used for real life implementations.
5. Some ideas and discussion for future work are provided.

1.4 Contents of the thesis

- Chapter 2 contains a review of the literature concerning the main aspects of the Thesis, i.e. classification validation methodology, data imbalance preprocessing, pediatric respiratory crisis monitoring, and patient readmission.
- Chapter 3 contains our contribution regarding the triage prediction of children with respiratory crisis. State of the art classifier learning methods are applied successfully to the available dataset. Recommendations for real life techniques are proposed.
- Chapter 4 contains our contribution regarding prediction of hospitalization after readmission. We have to deal with imbalanced datasets, so the carefully designed data processing pipeline is described, and results of the entire data balancing and classification performance were provided. Recommendations on the most appropriate classifiers and data balancing methods are provided

- Chapter 5 contains some discussion of the paper contributions, conclusions and ideas for future work.
- Appendix A recalls definitions of the classifier learning methods.
- Appendix B recalls definitions of class data balancing methods.

1.5 Publications achieved

In this section we gather the publications that have been achieved during the works ending up in this Thesis. Some are directly related to the actual content of the Thesis, others are indirectly related in the sense that are the fruit of collaborations that have in some way involved methods also used in the actual Thesis work

1.5.1 Main publications in the work of the thesis

Asier Garmendia, Sebastian A. Rios, Jose M Lopez-Guede, Manuel Graña, “Triage prediction in pediatric patients with respiratory problems”. Neurocomputing (accepted)

Asier Garmendia, Manuel Graña, Jose Manuel López-Guede, Sebastián Ríos. “Predicting patient hospitalization after emergency readmission”. Cybernetics and Systems (accepted)

1.5.2 Publications achieved as a collaboration indirectly related to the Thesis

Jose Manuel Lopez-Guede, Jose Antonio Ramos-Hernanz, Jesus Maria Larrañaga, Asier Garmendia, Valeriu Manuel Ionescu., (2015), "Study on the influence of Lambda parameter on several performance indexes in Dynamics Matrix Control". Journal of Electrical Engineering, Electronics, Control and Computer Science, Vol 1, No 2, pp 1-8, ISSN 2457-7812

Eneko Solaberrieta, Aritza Brizuela, Cristina Fraile, Rikardo Minguez, Asier Garmendia, Olatz Etxaniz, Guillermo Pradies., (2015), "Integration of Reverse Engineering and Dental Mandibular Dynamics". DYNA, Vol. 90 Issue 6, pp 643-646, ISSN 0012-7361

Eneko Solaberrieta, Asier Garmendia, Rikardo Minguez, Aritza Brizuela, Guillermo Pradies., (2015), "Virtual facebow technique". The Journal of Prosthetic Dentistry, Volume 114, Issue 6, pp 751-755, ISSN 0022-3913

Eneko Solaberrieta, Asier Garmendia, Aritza Brizuela, Jose Ramon Otegi, Guillermo Pradies, Andras Szentpétery.(2016), "Intraoral Digital Impressions for Virtual Occlusal Records: Section Quantity and Dimensions". BioMed Research International, Volume 2016, pp 1-7, ISSN 2314-6141

Chapter 2
Literature review

In this Chapter we provide a review of the literature regarding the two main topics covered in this Thesis in Chapters 3 and 4, namely prediction of respiratory crisis in children, and prediction of hospital readmission.

The Chapter starts with some recall of Machine Learning role in modern health care data processing in Section 2.1. Section 2.2 reviews literature related to children respiratory disease monitoring. Section 2.3 reviews the literature related to readmission prediction.

2.1 Data processing and Machine Learning

The use and analysis of large amounts of data is increasingly frequent in the health industry with the aim of improving the quality of healthcare. The volume of data and other properties such as the velocity of production and the need for real time response put the health care data into the realm of big data [BAT14][SUT15]. The promise of tackling with these large quantities of data is improvement on the quality of care and reduction of costs. The big data approach promises the ability to pose questions at multiple levels and multiple scales, from the nano to the macroscopic, from the molecular interaction to the level of large human populations [HER14].

However, the quantities of data processed in this paper are far below any threshold for big data. For the analysis of such data, Machine Learning approaches have been suggested, and caution about the statistical value of conclusions have been also raised [CRO15]. Methodological rigor is critical for the appropriate validation of the models, using cross-validation approaches that clearly separate training from testing data, thus avoiding any circularity or double dipping effect.

Machine Learning approaches encompass supervised and unsupervised learning methods. The latter better represented by the clustering approaches, such as the k-means [HAR15], that do not use any gold standard information to drive the learning. Clustering methods are exploratory in nature, proposing alternative representations of the data that may ease visualization and interpretation. On the other hand, supervised methods require some gold standard labeling of the data in order to estimate the error committed by the model, so training is guided by the minimization of such error. Supervised methods are predictive, either of the class label in classification problems, or of the given continuous index in regression problems. The distinction between regression and classification can be not so easy in cases where a continuous dependent variable is quantized in some intervals, and the predicted variable becomes a set of class labels.

The frontier between Machine Learning and traditional statistics may also become blurred as far as any of the two basic tasks are carried out. In [SON04] a comparison between linear and non linear learning methods for survival prediction has been carried out, showing consistent superiority of the artificial neural networks. However, the greater differences in prediction performance were related to the differences in population, from acute short term to long term predictions. Hinting that the quality of the data may be more critical than the learning algorithm. Machine Learning have been also used to tackle the problem of generating health care recommendations, much like in Amazon or Netflix, in the form of medication and other treatment orders [CHE16] after mining the existing electronic health records. Machine learning techniques widely used are artificial neural networks [AMA13], decision trees [POD02], k-nearest neighbor (kNN) classification [KHA14], k-means clustering [HAR15], and logistic regression [BAG01].

An important aspect of many health care datasets is the existence of big class imbalances, i.e. one of the classes is much more represented in the population than the other [LOP13]. Such kind of populations presents a big problem for most Machine Learning techniques which are naturally biased towards the majority class. In essence, the a priori probability of the class weights too much in most algorithms. The problem can be tackled in two basic ways:

- Manipulating the dataset in order to achieve a more balanced representation of the classes. For the fair evaluation it is required that the testing phase is carried out over an imbalanced sample of the data, even if the actual training has been carried out on an artificially balanced dataset
- Manipulating the learning algorithm so that it minimizes a version of the error function that takes into account the differences in the cost of misclassification, often much higher for the minority class.

2.2. Children respiratory disease monitoring

The prevalence of respiratory diseases in the world is increasing alarmingly [CRU07]. This situation is particularly severe in the city of Santiago, capital of Chile [GAR14]. Children are one of the most affected groups by these diseases in Santiago [PRI07], as happens in other major world cities, such as Sao Paulo (Brazil) [SOB89]. Children respiratory diseases are a pandemic in the developing countries, and in regions of developed countries suffering high air pollution levels.

There are a host of diseases related with respiratory problems [FIS65], although it is true that the particularities of the specific region clearly determine the greater or lesser incidence of these diseases. In the city of Santiago, besides the pollution caused by reasons inherent to a big city, the particular topography of the city, which is surrounded by the Andes, hinders the removal of pollutant particulates. The Environment Ministry of Chile provides through an internet portal daily data about air quality in the metropolitan region [MAP16]. Respiratory diseases often become chronic, passing from childhood to adulthood. According to the World Health Organization, chronic diseases kill more than 36 million people each year. If we analyze the situation in Chile, as illustrated in Figure 2.1, deaths caused by most chronic diseases show a clear downward trend [WHO16], however not all chronic diseases follow the same pattern. While cardiovascular diseases show a clear decrease in the mortality rate in recent years, chronic respiratory diseases remain steady (shown in more detail in Figure 2.2). One reason for decrease in mortality is better patient care. Therefore, care improvements for respiratory disease patient are strongly needed. Data mining techniques [FAY96] have been acknowledged as potential source of improvements for patient treatment and follow up. Specifically, in this Thesis we are dealing with the prediction of the risk level as specified by the triage classification performed by the nurse or the doctor.

Machine learning has already produced predictive models in the pediatric area, such as PRISM (Pediatric Risk of Mortality), a system that delivers an index that determines the probability of death of a patient admitted in the unit of critical hospitalization, up to mechanisms that attempt to identify abnormal behaviors in vital signs in order to prevent diseases using as a basis statistical studies of data [HAN10]. Other predictive systems seek to predict the likelihood of mortality in patients based on a traditional regression approach. For example, the Predictive Index of Mortality (PIM) system performs a measurement of variables as vital signs, fan pressure, among others. These are measured in the first 24 hours after admission of a patient in the PICU (Pediatric Intensive Care Unit) [BRA06]. Other authors consider the impact of variables using a scoring system based on the values of the variables [RAD14]. Other studies discover patterns of deviation of vital signs by analyzing its percentiles, so that signal deviations from expected ranges of can be identified on hospitalized children [BON13] [SUB11].

In this Thesis we apply Machine Learning techniques to predict the Triage variable, which is revised by the nurses and doctors each hour while the patient is under intensive care. The

2.3. Readmission risk prediction approaches

Nurse practice and culture [FLE09] [KOV08] [SCO05] is evolving thanks to the advent of new technologies that allow the use of data and accumulated evidence [ZIM14]. Though the introduction of electronic health records meets some reluctance by the practitioners because they claim that non justified additional effort must invested filling the them, the accumulated information allows for a posteriori processing and extracting conclusions from more extensive and accurate data collection. One not minor task is to assess from the gathered information conditions that can have health or economic impact, such as the likelihood of the readmission event. In Spain, 12.4% of the internal medicine admissions in the period 2006-2007 were rehospitalized within 30 days [ZAP12]. This prevalence of readmissions is widespread in the health care systems, with a huge economic and personal health cost. In this Thesis we focus on the readmissions in emergency departments (ED), which suffer additional stress due to the kind of patients and the fact that many non emergency patients resort to this service as a way to obtain quicker response due to the malfunction of other services. Therefore, readmissions can be especially cumbersome for these patients because they imply that somebody that was supposed to have an acute urgent condition was returned home without appropriate care.

The issue of deciding if a patient revisiting the service is a readmission or not is not trivial. Many papers in the literature assume that 30 days is the threshold for readmission [ZAP12], mainly because this timeline was set in USA by the medical insurance companies to enforce economic conditions on hospitals. However, 30 days may be too large interval for an ED readmission, because the repetition of a condition after so many days implies that it is more a long term disease than an emergency. Therefore some countries, i.e. Chile, have set a shorter interval for the categorization of a revisit as a readmission, i.e. 3 days.

Prediction of readmission risk has been approached for specific populations, as discussed in a recent review [KAN11]. Often indices are limited to specific subpopulations, such as people suffering from Chronic Obstructive Pulmonary Disease [NGU14] or Acute Myocardial Infarction [ZAI13], or services, such as internal medicine [ZAP12]. The modeling approaches often combine several administrative, demographic, biochemical measures, and psychological tests to compute a risk index, which can be estimated by some regression approach based on actual data. Differences between centers in electronic medical records and recorded patient information lead to the need to develop institution specific models, i.e. predictive models trained with institution specific data [YU15], or specific healthcare networks [HAO15].

Aging population poses new problems to the health care systems [PAR07], and it is a growing global concern [BLO10] [REC09]. This situation is also felt in the ED suffering of additional workload caused by geriatric patients. Prediction works focused on a stratified sample of older patients reported in [BES15] developed specific predictors for each kind of patient. Patient strata were the following: diabetes mellitus, heart failures, chronic respiratory problems, and general care problems. Feature predictive value for each population strata was modeled by the Gini index associated to the variable in a decision tree built over the dataset, following the method suggested in [BRE01]. The results showed wide differences in the set of most discriminant variables between population strata, suggesting that specific population strata predictors are justified.

The highest rates of ED readmission, the longest stays, and greatest resources invested in ancillary tests correspond to adults above 75 years of age [PER15] [SIL15]. Despite this intense use of resources, these patients often leave the ED unsatisfied, with poorer clinical outcomes, and higher rates of misdiagnosis and medication errors compared to younger patients.

Additionally, they have a higher risk of ED readmission, hospitalization after readmission, death and institutionalization [CAR11]. Readmission risk prediction is therefore critical for this kind of patients [DES15].

A widely accepted approach is the LACE readmission index [VAN10] developed from data of a network of Canadian hospitals. It has defined on the base of logistic regression analysis of physiological and demographical variables of a sample of near 50000 patients. Though not small dataset is smaller than our own database in this paper. The LACE+ [VAN12] makes use of administrative data to improve the risk assessment. Closely related to LACE, HOMR (Hospital patient One year Mortality Risk) by [VAN15] is a model for predicting death within one year after hospital admission. According to the authors the goal is to predict long-term survival after admission to hospital. The variables used are included in the following categories: Demographics (age, sex, etc.), Health status (Charlson comorbidity index, number of visits to hospital emergency etc.), and Acuity disease (emergency admissions, direct intensive care unit admissions etc.). The dataset used for the development and validation of this model consists of more than three million instances obtained from several hospitals in the areas of Ontario, Alberta and Boston.

Most modeling approaches use logistic regression as an established statistical technique that allows also assessing the significance of specific variables [ZAP12] [VAN10]. An approach based on classification by support vector machines (SVM) [YU15], allows tuning the prediction model to particular, avoiding to apply a risk index developed on a specific population to potentially very different populations. In this Thesis we follow this approach to tune the model to pediatric and adult subpopulations of a specific hospital.

Variables taken into account vary from population to population. Most works deal with physiological, biochemical and administrative/demographic information. The most predictive variables can be selected following various methods related to the machine learning approach used. For instance, logistic regression allows to assess the most sensitive variables analytically. Other approaches use the classifier building information, such as the Gini index used in [BES15]. Specific traits, such as medication regime are less predictive than expected, but clustering patients have been found to improve prediction [OLS16]. For instance, in the stratified study [BES15] the five most discriminant variables for the case management population are: Use of antipsychotic drugs, availability of caregiver, age, use of insulin, and mental disorder diagnosis. For the heart failure population, the five most discriminant variables are: diagnosis of vascular disease, renal disease, memory complaint, living alone, and use of inhalers. For the diabetes mellitus population, these five variables are: diagnosis of sensing impairment, use of heparin, comorbidity index, and arrhythmia and medication reconciliation. Therefore, wide variations in the collection of most significant variables must be expected, which hinders seriously the practical use of predictive models for large heterogeneous populations.

2.4. Figures

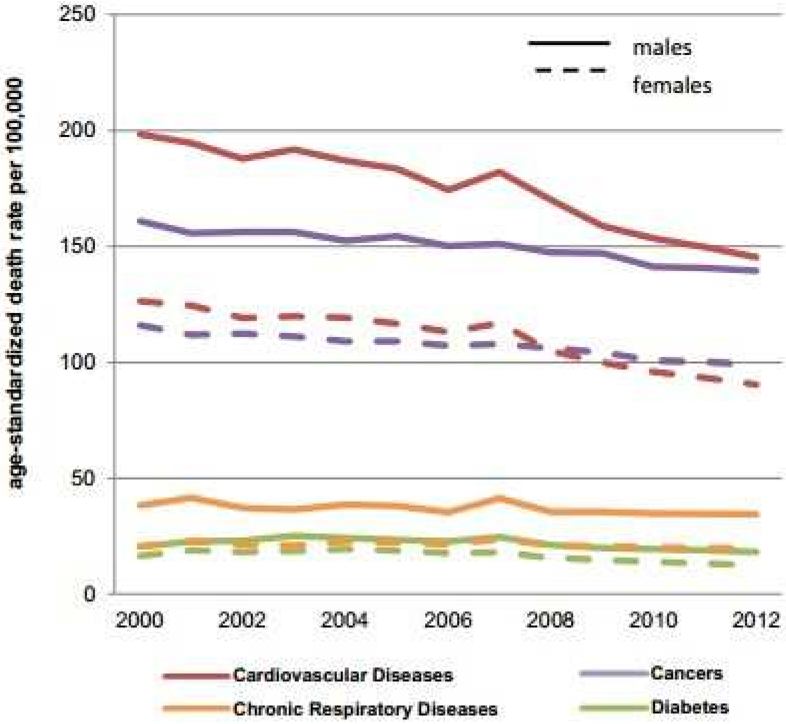


Figure 2.1: Deaths from chronic diseases in Chile in the period 2000 - 2012 [DEPT].

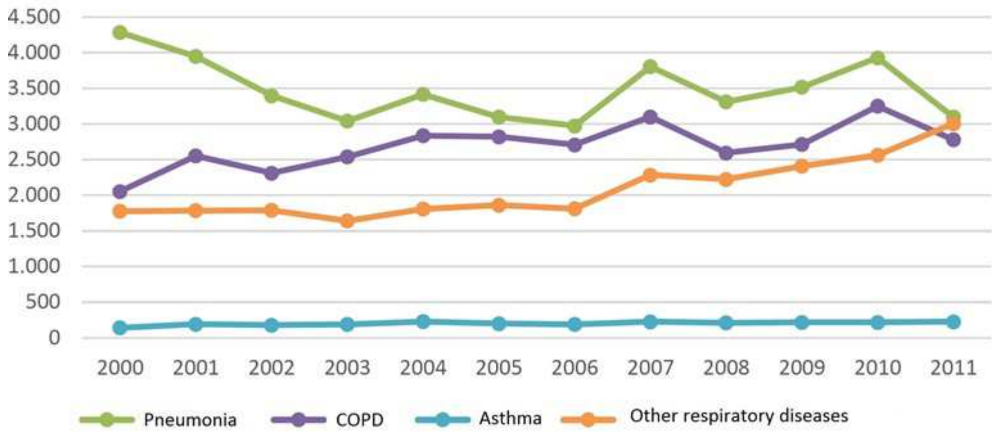


Figure 2.2: Deaths from respiratory diseases in Chile in the period 2000 to 2011[DEPT]

Chapter 3
Prediction of respiratory crisis

This Chapter is devoted to the first practical application of computational intelligence and predictive approaches to a health care problem. Specifically, we tackle the problem of respiratory crisis prediction in children, which is a rather critical situation, because children crisis evolve quite fast and dramatically, if not treated. Crisis prediction is considered as the prediction of the Triage value for the children in an hourly monitoring environment.

The structure of the Chapter is as follows. Section 3.1 provides a short introduction to the Chapter objectives, and a short review of the state of the art of data mining applications in health care, with the aim of setting the stage for applications in respiratory disease assistance. Section 3.2 provides the description of the actual dataset employed for the computational experiments, explaining the variables included in the study, and their preprocessing. Section 3.2.1 provides descriptive statistics of the data, which serve as an analysis previous to the application of machine learning approaches. Specifically, we look closely at the dependence of some variables with respect to the patient age. Section 3.2.2 discusses two ways of normalization of the dependent variables to make them uncorrelated with patient age that we have explored. Section 3.3 describes the computational experiments conducted seeking to predict the triage at each time instant. Section 3.4 presents the results of these experiments. Finally, Section 3.5 gives our conclusions.

3.1 Introduction

The prevalence of respiratory diseases is increasing in big populated areas which also attract industries and heavy traffic. Air pollution is the leading cause for children hospitalization in developing countries. Children are the most vulnerable sector of the population to air pollution for several reasons: they are still developing the respiratory system; they have greater income of volume of air per unit of bodyweight per breath than adults; they perform rapid and deep breaths; and they are more typically mouth-breathers [BAT07]. There are definitive evidences that areas with industries that are emitters of fine particles (PM2.5), sulfur dioxide (SO2) and nitrogen dioxide (NO2) have a greater prevalence of respiratory complications, increasing related hospital admissions [BRA16][MOO16]. A clear effect of diesel engine traffic on the increase of critical respiratory hospital admittances has been found [NIR15b]. However, pollution due to industry and transportation is not the unique cause. Regions naturally having high amounts of ambient dust in the air, like deserted zones [NIR15B], also suffer this kind of children respiratory pandemics.

Increasingly automated health monitoring systems are desirable, though they require advanced electronic equipment [URB15] that may not be widely available, such as the direct connection of continuous reading physiological sensors to decision support systems [CHA13][PER15b]. Already developed monitoring systems that use computational intelligence tools, such as Markov models [RAV14], may have not taken into consideration the specific characteristics of the children population, i.e. the strong dependence between physiological variable values in the healthy state and the patient age. This neglect ion may render useless for children the proposed automated solutions. In other words, it is not viable to apply successfully alarm detection systems developed for adults and aging subjects, over data extracted from them, to children. Monitoring and respiratory assistance has been received attention in the care of acute cases, such as children with muscular weakness [ABU15] or ventilated newborns [ACH16], but not very much for milder cases that can become acute when untreated, such as is often the case of low income children [STE16].

The main focus of the work reported in this Chapter is building predictive models of the risk levels of pediatric patients hospitalized for diseases related with respiratory problems. These risk levels are determined by a "triage" process [ROB06]. In the data considered in this paper, risk levels are coded by numbers from 1 to 4, with 1 being the mildest and 4 the most severe, risking death. Predictive models have been built by machine learning approaches [BAG01] [HAR15] [KHA14] [POD02], applying a strict cross-validation methodology to assess generalization of prediction results. The available data comes from health records previous to any electronic implementation so that they have been entered manually in excel spreadsheets. Therefore they are noisy and need some preprocessing to remove erroneous values and records with missing variable values. After preprocessing and data cleaning, we carry out a feature selection process, which can be exhaustive because of the small dimensionality of the data. We have not performed any transformation, such as Principal Component Analysis, in order to preserve the original meaning of the variables, and, therefore, to be able to argue with the experts about the value of findings. In medical applications it is essential to have clear explanation of the actual rules that lead to a conclusion; otherwise the medical doctors tend to dismiss conclusions not supported by medical explanation and intuition. Besides, it is not clear up to what point the original data space is continuous enough to allow sensible data transformations.

3.2. Description of the dataset

The experimental data for the work reported in this chapter is based on records obtained in pediatric units and CPU (Critical Patient Unit) of the Hospital Dr. Exequiel González Cortés (HEGC), which is a pediatric medical center belonging to the public health system of Chile, located in the municipality of San Miguel in Santiago de Chile. These records belong to patients who have different diagnoses, but all of them fall in the class of diseases considered as respiratory conditions. The initial number of records is 22025, corresponding to the hourly monitoring of 45 patients (29 boys and 16 girls) aged between a few months to 16 years, that were hospitalized in the intensity care unit sometime in the period between 2012 and 2014. We had access to anonymized physiological measurements and record annotations. Before carrying out the analysis of the starting data, there has been a preprocessing of the data, considering, on the one hand, the contained variables, and, on the other hand, the records. Regarding the variables, we have the following ones:

- Ancillary information: the patient Age, number of days hospitalized, time instant of variable measurement. The care protocol states that measurements are to be taken each hour of the day. There are two special times in the morning and afternoon when a more detailed medical visit is paid to the patient.

- Physiological variables measured: Temperature, Heart Rate, Respiratory Frequency, Systolic and Diastolic Blood Pressure.

- The Triage assigned by the physician. This variable indicates the severity of the patient condition. There are four triage levels, being classified as 1 the mildest, and ranked 4th most severe.

The objective is the prediction of triage. In other words, we aim to imitate the medical doctor decisions.

3.2.1. Descriptive statistics of the dataset

Considering the gender variable, from the 45 patients considered, 64.4% are male (29) and 35.6% women (16). Age is a quite important variable, because many normal physiological variable values are strongly dependent on it. For instance, heart rate is much quicker in small age children. The minimum, mean, and maximum age in the dataset are 0.08, 3.43, and 16.75 years, respectively. Figure 3.1 gives a more precise idea of the distribution of patient age variable (measured in years). Most of the patients are of tender age, with mean around three years and half. On the other hand, there is a tail in the age distribution corresponding to adolescent patients of fifteen and sixteen years, which can be considered as chronic patients passing to the next layer of care. The Triage distribution is also irregular: 7.3% with triage 1, 36.4% triage 2, 20.9% triage 3, and 35.4% triage 4.

To assess the dependence of the physiological variables on the age, we show the scatter plots and regression line of the Heart Rate, Respiratory Frequency, Systolic Blood Pressure, and Diastolic Blood Pressure versus Age in Figures 3.2, 3.3, 3.4, and 3.5, respectively, separating the plots by each Triage value. It can be appreciated that the younger patients have much higher Heart Rate and Respiratory Frequency than older children. The blood pressure variables do not show such trends. For this reason, it is necessary to normalize these variables removing the age dependent trend.

3.2.2 Data preprocessing

The respiratory monitoring data missing values were filled by using the previous value of the variable in the time sequence, because the natural behavior of the nurse is to write down only variables that have changed. The original data was written in paper sheets, the digitization was carried out manually to fill an excel spreadsheet with the data. Some variables were incorrectly transcribed in some registries giving inconsistent values that had to be corrected manually. When a register had too many errors it was removed.

3.2.3. Approaches to the normalization of the age dependent variables

Taking into account the age dependence of the physiological variables, before applying classification models, we have normalized the variables Heart Rate, Respiratory Frequency, Systolic Blood Pressure, and Diastolic Blood Pressure in order to achieve age independent classifiers. We have tried two normalization approaches:

1. The first approach, denoted *Norm1* in the results below, computes a linear regression model considering each of the dependent variables and the age variable. Then, the regression residuals are the classification feature new values corresponding to each of these

variables.

2. The second approach, denoted *Norm2*, takes into account the published normal values of these variables by age ranges in healthy subjects [FLE11] [NORM]. The values of each record of each patient have been interpolated to the same common age range, obtaining standardized values.

3.3. Experimental design

To build Triage predictors from the normalized physiological features we have applied four kinds of classification algorithms: Multilayer Perceptron (MLP), Decision Trees (DT), k-Nearest Neighbors (k-NN), and Naive Bayes (NB). We have used the Caret package [CARET] (short for Classification and Regression Training) in the R programming environment, which provides a set of functions to streamline the process for predictive model creation and validation. This package allows to process the dataset applying "Repeated k-fold Cross Validation" using the learning algorithm as a parameter. Some further details are given in Appendix A. For the experiments in this Chapter we have applied 10-fold Cross-Validation repeated 3 times. We have carried out a feature selection process as detailed in Table 3.1, and the exploration of the value of each feature independently.

Overall, 10 cross-validation experiments were carried out. Five experiments with the values of the variables that have some dependency on the age normalized by using the residuals of a linear regression on the age (*Norm1*). Other five experiments in an analogous manner, but with the values of the dependent variables of the age normalized by interpolation inside the range of normal values of healthy children (*Norm2*).

Notice that we have not used other conventional intelligent system tools such as case based reasoning [CHI93] because they are not easy to train, and they are not as robust to noise and uncertainty as the numerical methods referred that we apply. Specifically, case based reasoning has great difficulties to increase in a robust way the database, while neural network and probabilistic methods assume continuous representations that interpolate naturally between data samples. However, it can be the tool of choice in other applications where the cases are precisely described by crisp values, such as RFID information manipulation [CHO09] [POO09]. Some extensions using grey codes to achieve some kind of noise robustness have been proposed [XIN12] but need careful evaluation before being of use in our problems. Hybridizations of case based reasoning approaches may be of use in health care data, i.e. adolescent health care [WAN07], but are rare and have not been further exploited.

3.4. Results

The average accuracy obtained in cross-validation experiments where each input record contains only one feature, i.e. assessing the prediction power of each feature independently, is presented in Table **3.2**. It can be appreciated that the only feature that provides a result above 0.70 accuracy is the Respiratory Frequency after normalization by linear regression (*Norm1*). The normalization by linear regression is significantly better than the categorization in age intervals (*Norm2*), confirmed by a one-sided t-test ($p<0.01$) computed over all results obtained with each normalization. The results, still, are far from being satisfactory. Table **3.3** provides the results for the combinations of features specified in Table **3.1**. Significantly best results are achieved by feature set #5 which includes past Triage values in the features for classification. ($p<0.001$ in pairwise t-test between feature set #5 against the others). An unexpected result is that the Respiratory Frequency seems to have a more salient role in the Triage decision that the Blood Oxygen Saturation, which is often referred by clinicians as the preferred biomarker for the severity of the respiratory condition.

3.5. Conclusions

The aim of this work is to develop automatized systems to monitor children suffering from respiratory diseases in a pediatric intensive care unit. We proceed by trying to emulate the Triage decisions of the physicians as recorded in a dataset containing the physiological variable measurements and the Triage decision. The dataset original recordings are quite noisy, with many missing values and some inconsistent variable values. Direct recording of physiological sensors and storage without human intervention would improve this situation, but these technologies are not widely available yet. The actual experiment allows extracting several conclusions on the data and the clinical protocol followed by the clinicians. First, it is surprising that the Respiratory Frequency appears to be much more influential in setting the risk level than the Blood Oxygen Saturation. When asked, clinicians answer that BOS is the primary variable to take decisions of patient state. The second important conclusion is that the clinical practice is guided by conservative decision strategies, explaining the big increase of accuracy achieved when the feature vector includes recent past Triages, and the fact that the decision “do not change the Triage” achieves 85% prediction accuracy. Thirdly, the very low prediction rates using the current value feature vectors (i.e. without past triages) is an indication that there are unrecorded qualitative information that it is also very influential on the physician decisions.

The improvement of clinical care practices requires additional feedback on the value of the actual stream of decisions taken by the practitioners. New data gathered will be including continuous monitoring of physiological variables, as well as analysis of ancillary information, in order to extract additional clues. The eventual outcome of the hospitalization, as well as the information about readmissions will help improve the Triage predictive system and the patient care. The ultimate goal is to obtain a continuous monitoring of the children to raise alarms in advance. Future works may also address the integration of the free style written information put down by either nurses or doctors in order to fill gaps in the actual information gathered from the monitoring [KIM16]. Such hybrid systems may greatly improve the knowledge extraction from the past monitoring records.

3.6. Tables and figures

#	Features
1	T, HR, RF, BOS, SBP, DBP
2	*t* ,T, HR, RF, BOS, SBP, DBP
3	#H, T, HR, RF, BOS, SBP, DBP
4	*t* , #H, T, HR, RF, BOS, SBP, DBP
5	*t* , #H, T, HR, RF, BOS, SBP, DBP, TR_{t-i} i=1, … ,6

Table 3.1: Feature selection experimental design. Variables T= Temperature, HR= Heart Rate, RF= Respiratory Frequency, BOS= Blood Oxygen Saturation, SBP= Systolic Blood Pressure, DBP= Diastolic Blood Pressure, #H =Hospitalized days, TR_t= Triage at time *t*.

		Classifier			
Feat.	Norm.	MLP	DT	k-NN	NB
HR	Norm1	0.3726	0.3741	0.4242	0.3733
	Norm2	0.3971	0.3965	0.3904	0.3959
RF	Norm1	0.5546	0.6987	0.7163	0.5555
	Norm2	0.5822	0.5919	0.5935	0.58
T	Norm1	0.4058	0.406	0.4109	0.4081
	Norm2	0.4058	0.406	0.4109	0.4081
BOS	Norm1	0.4474	0.4521	–	0.4573
	Norm2	0.4443	0.4511	–	–
SBP, DBP	Norm1	0.4652	0.5188	0.4263	0.4457
	Norm2	0.4815	0.4935	0.4496	0.4506

Table 3.2: Average accuracy achieved by each feature independently.

Classifier					
#Feat	Norm	MLP	DT	k-NN	NB
1	*Norm1*	0.5749	0.5981	0.5521	0.5608
	Norm2	0.5952	0.5969	0.5534	0.5941
2	*Norm1*	0.5739	0.6014	0.5608	0.5665
	Norm2	0.5937	0.6017	0.5427	0.5999
3	*Norm1*	0.5812	0.6336	0.6179	0.5905
	Norm2	0.5939	0.671	0.6219	0.6173
4	Norm1	0.5814	0.6378	0.6171	0.5946
	Norm2	0.5965	0.6728	0.6023	0.6228
5	Norm1	0.8962	0.969	0.685	0.9359
	Norm2	0.8779	0.9695	0.6961	0.9353

Table 3.3: Average accuracy for each of the feature sets in Table 3.1.

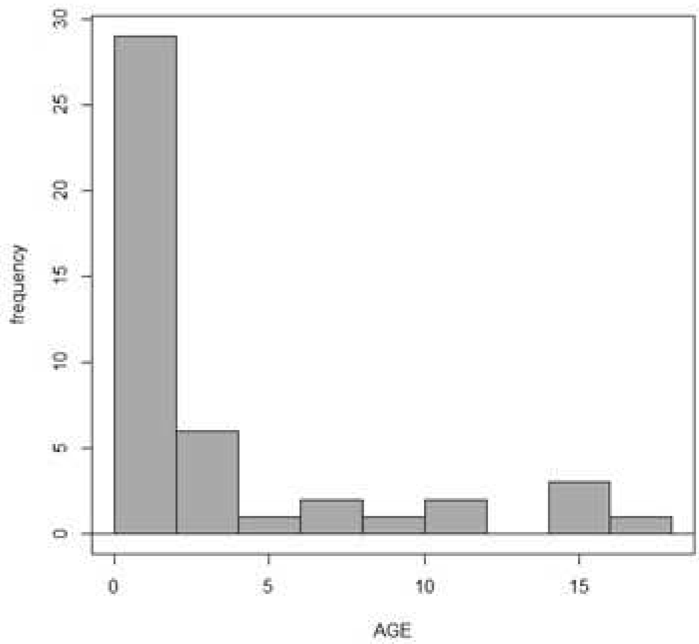


Figure 3.1: Frequency distribution of the variable age of patients.

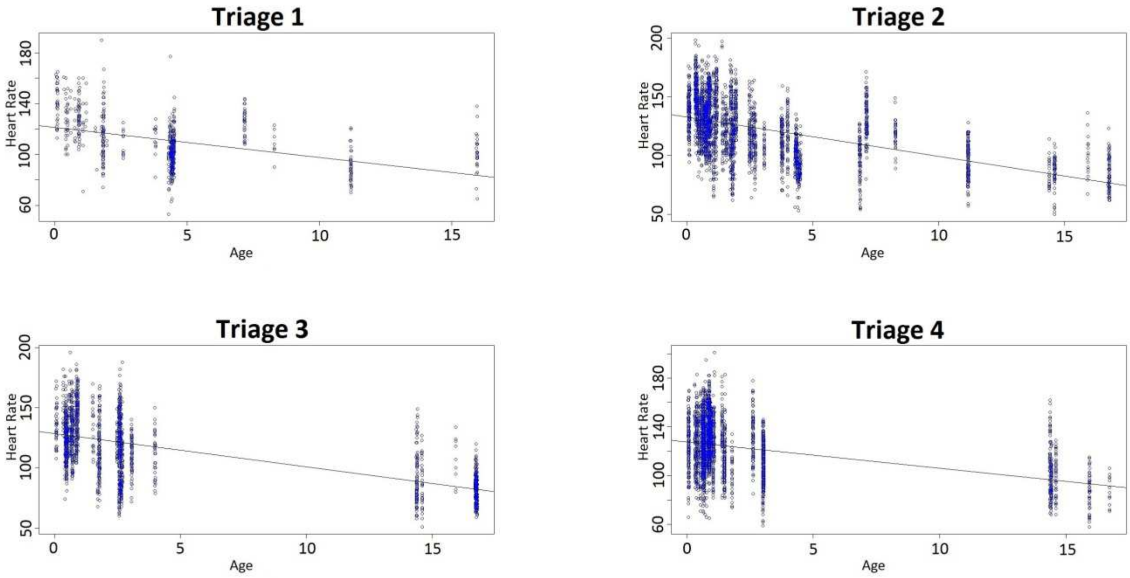


Figure 3.2: Scatter plots of the Heart Rate versus Age for each triage value.

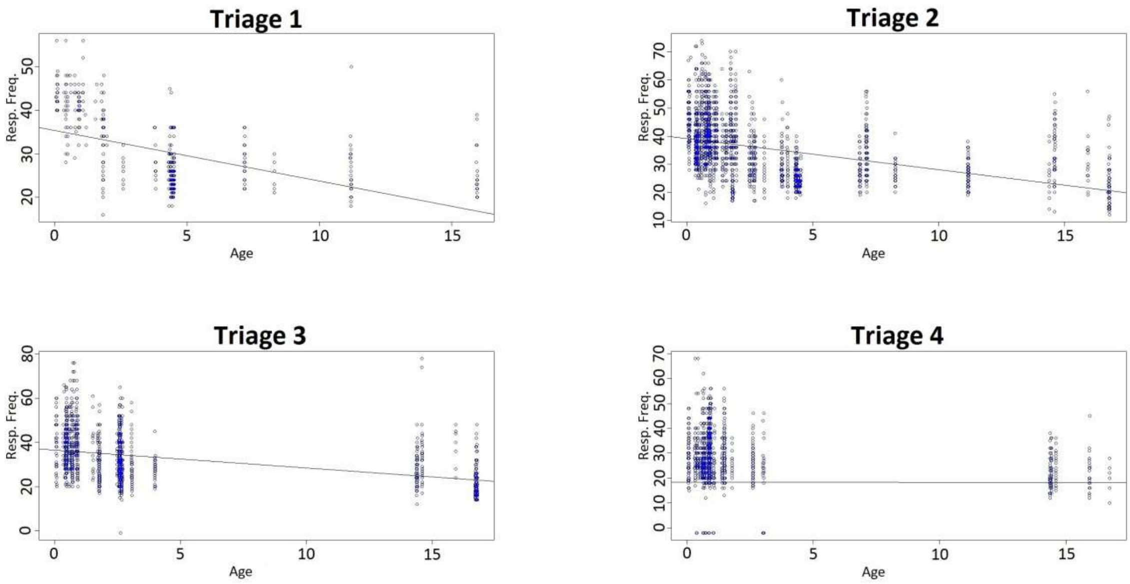


Figure 3.3: Scatter plots of the Respiratory Frequency versus Age for each triage value.

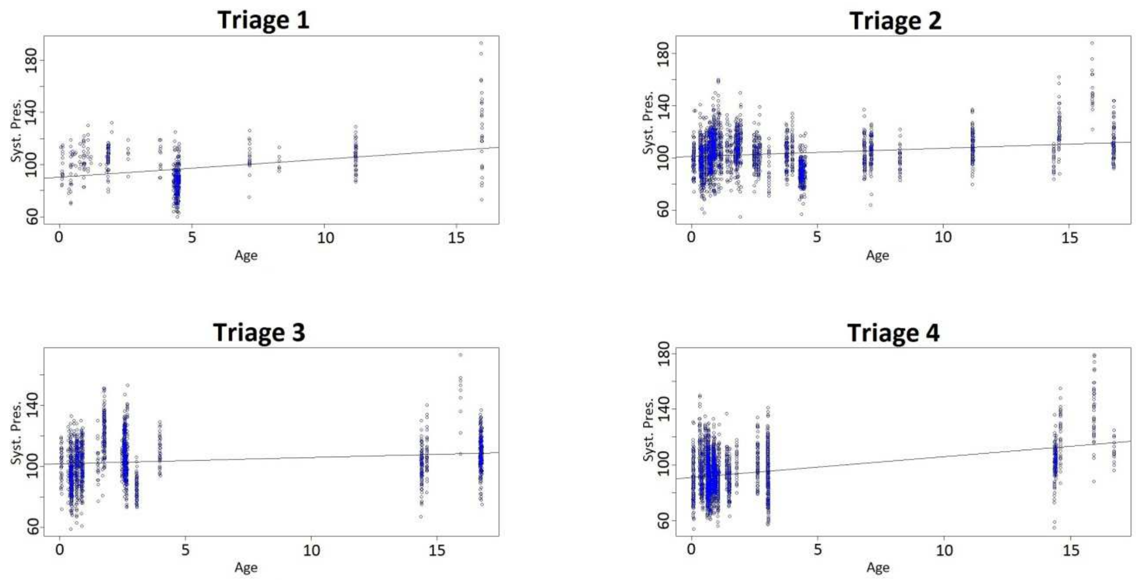


Figure 3.4: Scatter plots of the Systolic Blood Pressure versus Age for each triage value.

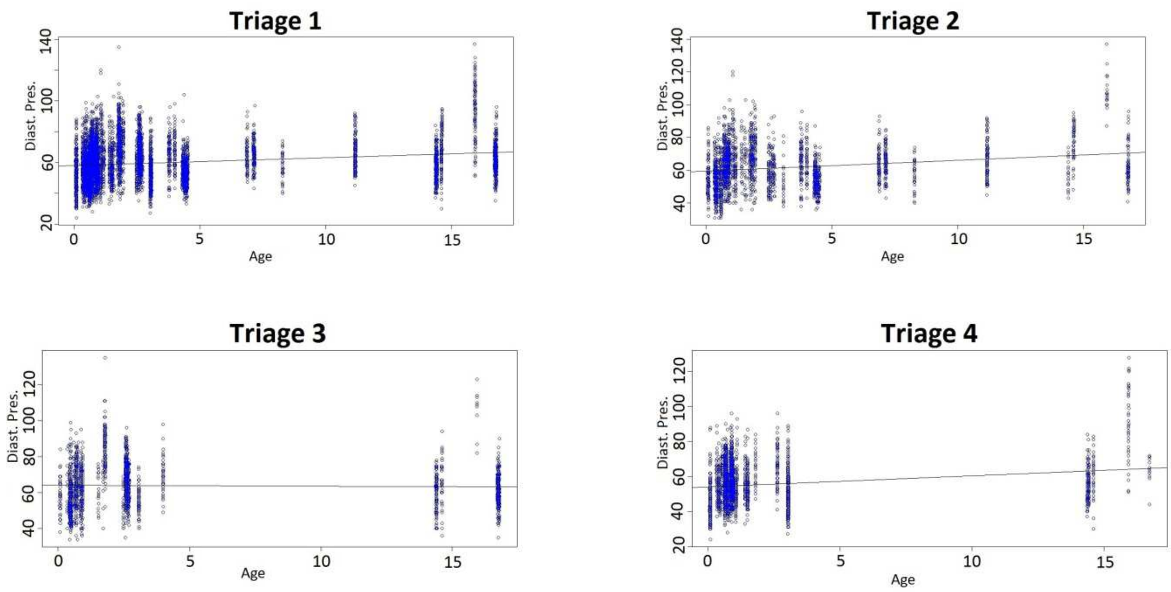


Figure 3.5: Scatter plots of the Diastolic Blood Pressure versus Age for each triage value.

Chapter 4
Prediction of hospitalization after readmission

This Chapter is devoted to another important healthcare problem: the prediction of the readmissions that end up in hospitalization of the patient. Such events are symptomatic of bad quality health care, and they have some relevant economical impact that can be prevented. The approach in this Chapter to tackle this issue is to build predictors of this event based on the information of the patient at the time of admission. The problem is heavily imbalanced, so that we test several class balancing procedures in order to improve sensitivity of the predictors. The computational experiments on the dataset provided by a university hospital allow recommending some specific class balance method and classifier training.

Structure of the Chapter is as follows. Section 4.1 gives some introductory remarks. Section 4.2 presents the dataset and the features for classification. Section 4.3 presents the experimental design. Section 4.4

gives the computational experiments results. Finally, Section 4.5 gives some conclusions.

4.1 Introduction

Emergency department (ED) readmissions within a short period of time after a previous patient discharge are indicative of either a bad quality of healthcare service or structural problems in the healthcare systems, such as chronic patients being attended in the ED for lack of a proper planning of their care. There is a growing need for sensitive predictive tools in order to improve planning and distribution of resources, as well as to provide a better healthcare experience to the patient. Some tools are specifically tailored to geriatric patients treated at ED [BES15], others are developed to address the needs of general healthcare services [HAO15], some are proposed as institution specific prediction models [YU15], finally some are focused on specific fragile populations [NGU14][OLS16][PER15].

In this paper we focus on the event of hospitalization after readmission prediction, which has received little attention in the literature. These events imply that the aggravation of patient condition since the last admission could have been prevented. We pose its prediction as a classification problem. Hospitalization is a rare event; therefore the class distribution in the dataset is forcefully very imbalanced, requiring the application of class balancing method before training the classifiers. We have carried out cross-validation experiments testing all combinations of class balancing method, classifier training techniques, and readmission thresholds. Readmission thresholds range from 3 days up to 30 days depending on political circumstances, so considering several corresponds to the a discrete survival time estimation problem. Figure **4.1** shows the percentage of readmission events and patients as the readmission threshold grows in the actual database used in the experiments. Notice that, as expected, the number of readmission grows with the time allowed to count as a follow up and undesired return of the patient. Hence in some administrations, the threshold is set in order to justify some predetermined quality criteria, e.g. 3 days to achieve a rate of readmission below 5%.

The anonymised dataset used in the computational experiments covers more than three years of the activity of the ED in a university hospital of Santiago in Chile, while the dataset explored in this paper includes adult and pediatric patients, which have quite different patterns of attention and readmission. Therefore, we have carried out separate experiments of hospitalization prediction for pediatric and adult populations. In this Chapter, Section 2 describes the dataset and the classification features. Appendices A and B describe the classification training and data balancing methods. Section 3 gives the experimental design ensuring that there is no bias in the results. Section 4 presents our results, and Section 5 ends with some conclusions and directions of future work.

4.2 Dataset and classification features

Our raw dataset is composed of ED admission events of N=101507 patients divided into 2 groups, namely adults NA=80508 (78.82%) and pediatrics NP=21269 (20.96%). Some pediatric patients have changed into adults, so they appear in both lists. The dataset contains 156120 admission cases recorded between January 1st, 2013 and April 30, 2016 in the electronic medical record system of the Hospital José Joaquín Aguirre of the Universidad de Chile. At admission time a set of 17 variables were collected, including physiological measures, such as temperature, breath rate, heart rate and blood pressure, biochemical measures such as glucose level, demographic variables such as gender, age, fragility index. There are five triage levels to distribute the patients. Besides, the nurse had to select a motive for the admission, which is a categorical value among five hundreds. If the time between visits to the ED falls below the readmission threshold, then it is a readmission event, otherwise it is an unrelated event. Readmission thresholds vary between countries for political or economical reasons. We have considered four possible thresholds (3, 7, 15 and 30 days). At discharge patients can go home (74.02 %), be hospitalized (12.72%), translated to another center (3.26%), or other situations (9.6%) including left without being seen (8.88%). We are concerned with the event of patient hospitalization as a result of readmission, because it is symptomatic of some lack of diagnosis or treatment leading to worsening of patient condition. Figure **4.1** shows the distribution of such events according to the readmission threshold. Though these events are rare compared to the total ED events (2% at most), they are significant relative to the readmission events (more than 20%). Hence they deserve some special attention. Table **4.1** provides the most salient ED visiting motives, those accounting for 1.5% of the cases or more, for patients that are labeled as readmitted under the different readmission thresholds. To assess the difference in the patient profile, Table **4.4** provides the most frequent admission motives for the patients that end up hospitalized after readmission, according to the readmission threshold. In all cases the non-informative category OTHER is the most frequent, pointing to excessive workload on the nurses or difficulty to assess precisely the predefined categories. There is little difference on the main causes of readmission, save for some minor changes in order. Besides, it can be appreciated that the distribution of motives is similar for all the readmission thresholds, with little variations in their ordering. More acute symptoms have greater prevalence in the shorted threshold (3 days) than in the larger one (30 days).

The feature set is composed of the numerical codification of the variables measured at admission time. This codification is trivial in all variables, but not for the variable motive of the admission. One way to encode the motive is to define a binary variable per value, ending up with a feature space of more than 500 dimensions very sparsely populated, which poses many difficulties for training predictors. In this paper we have encoded the motive in a single numerical feature whose value is computed in one of two ways: (a) the percentage of hospitalizations for a given motive relative to the number of readmissions for this motive, and (b) the position of the motive in the ranking of readmissions. Notice that this ranking, according to Table **4.1** is different for each readmission threshold, so that this feature encoding depends on that threshold.

Finally, we must consider the strong statistical differences that exist between adult and pediatric populations. Figure **4.2** shows the percentage of patients (figure **4.2a**) and of events (figure **4.2b**) that end up in a hospitalization, segregated into adults and pediatric populations. It can be appreciated a big difference in the number of readmissions between both populations. Figure **4.4** shows the distributions of the hospitalization events relative to the readmissions for the adult (fig. **4.4a**) and pediatric (fig **4.4b**) populations. It can be appreciated the adult population percentage of hospitalization is much higher for all the readmission thresholds (33%) than in the pediatric population (14%). There are many other differences that counsel to tackle each population separately, as will be done in our experiments in Section 4.4, such as the distribution of motives for the visit. Table **4.2** and Table **4.3** show the distribution of motives for the adult and pediatric patients readmission, respectively. Analogously, Table **4.5** and Table **4.6** show the distribution of motives for the adult and pediatric patients that end up hospitalized after readmission. It can be appreciated that some causes that have little impact in the adult population, i.e. cough, are very salient in the pediatric population, and vice versa.

4.3 Experimental design

Figure **4.5** shows the data selection process carried out at each cross-validation experiment repetition, in order to ensure that no bias is introduced in the classification evaluation. The process starts with the collection of all relevant data of readmission events (lower left corner) and proceeds by random splitting the data into 70/30 % subsets, which are used, respectively, for training and testing. Training includes the application of data balancing technique and the training of a classifier, which is applied to the test data. This process is repeated ten times for each combination of readmission threshold, data balancing method and classifier learning approach. That means that we carry out 5x4x4x10x2 cross-validation experiments, each featuring independent data balancing processes (5), classifier training (4), readmission threshold (4), and coding of the “motive” variable (2). The performance indices reported in the experiments are the following:

- Accuracy (A) computed as A=(TP+TN)/N

- Sensitivity (S) computed as S=TP/ (TP+FN) measures how much of the actual positive class we have discovered.
- Positive Predictive Value (PPV) PPV=TP/(TP+FP) measures how confident we can be of our positive predictions
- F= $(PPVxS)^{1/2}$ is a kind of f-score which combines S and PPV to assess the power of prediction of the positive class in the case of imbalanced data.

4.4 Computational results

We have summarized the cross-validation results in two aspects. First we have considered which of the data balancing methods can be recommended for further exploitation of the data. To this end, we compute the average performance metrics achieved with each balancing method for the separate populations of adults and pediatric patients. Average is computed over all readmission thresholds and learning methods. We compute pairwise one sided t-test on the cross-validation results achieved by each pair of data balancing methods, declaring as winner the method that has significant improvement ($p<0.001$) in all comparisons. Second, we compare the classifier approaches in the same way.

Table **4.7** shows the average results of the data balancing methods on the adult and pediatric populations, bold results per column signal the winner method. When ties occur both methods are highlighted. Notice the poor performance of the SMOTE approach. The winner method is the undersampling (UNDER) if we consider sensibility (S) and f-measure (F). Considering the positive predictive value (PPV), however, the winner is the TOMEK. Notice that the accuracy (A) is highly misleading, especially for the pediatric population which is much more imbalanced. Table **4.8** gives the average results of the learning methods tested. Attending to the S and F values, the multilayer perceptron (MLP) provides the best results; however Naive Bayes (NB) provides the best A and PPV values. Notice the difference in results between adult and pediatric populations due to their statistical properties. Specifically, the value of A is strongly influenced by results on the majority population. As a final recommendation for a practical hospitalization prediction we propose the data balancing by undersampling and the use of MLP as the classifier training method. One reason for the comparatively poor results of SMOTE is that the interpolation of the motive cause is meaningless, because minor alterations may change dramatically its meaning, i.e. it is not truly a continuous valued variable. Finally, Table **4.9** shows the results averaged for each value of the readmission threshold, from 30 down to 3 days. Consistently with the results of the Tables **4.7** and **4.8**, it can be appreciated that the results for the pediatric population are much worse than for the adult population. The average per readmission threshold is more biased towards the majority class as can be appreciated by the high accuracy and very low sensitivity and PPV values. The threshold of the 30 days allows more precise prediction, but thought may be statistically significant they are not big differences, so that we can assert that prediction would be insensitive to the readmission threshold.

4.5 Conclusions

Readmissions can be taken as a measure of the quality of service in healthcare systems. In some countries, i.e. USA, hospitals are penalized by readmissions under a time threshold, i.e. 30 days. The problem is not widely tackled in the case of emergency departments (ED), which may have stringent requirements, i.e. 3 days threshold in Chile. We have considered in this paper the case of hospitalizations as a result of readmission, which is a severe indication of the lack of medical solutions to the patient. We have approached the problem as a classification problem, with some success after testing several class balancing approaches and classifier learning methods. From our cross-validation experiments, we recommend the use of artificial neural networks as the classification training method, and majority class undersampling as the class balancing method. We found that the prediction results are relatively insensitive to the readmission threshold set.

Future works on this same database may involve trying other codification methods for the variable "motive", such as the use of orthogonal binary codifications. Other approaches may deal with independent one class classifiers trained for each motive value. The results in this Chapter may serve to enhance the information collected at admission time, in order to improve prediction and treatment.

4.6 Tables and figures

Readmission							
<3 days		<7 days		<15 days		<30 days	
Motive	%	Motive	%	Motive	%	Motive	%
OTHERS	30.13	OTHERS	27.25	OTHERS	25.36	OTHERS	23.37
GAP	8.20	GAP	8.02	GAP	7.54	GAP	7.36
1/3DF	5.40	COUGH	5.16	COUGH	5.68	COUGH	6.87
COUGH	4.28	1/3DF	4.91	24HF	5.21	24HF	6.68
24HF	4.10	24HF	4.27	1/3DF	4.64	1/3DF	4.79
HA	3.04	HA	3.11	HA	3.11	HA	3.01
D	2.59	D	2.90	D	2.82	T	2.91
T	2.43	T	2.63	T	2.62	D	2.79
EP	1.86	EP	1.72	LegP	1.87	LegP	1.80
LuP	1.51	LegP	1.72	LuP	1.69	AD	1.69
LegP	1.45	4/7DF	1.71	EP	1.61	LuP	1.65
4/7DF	1.44	LuP	1.61	AD	1.56	EP	1.49
IFPr	1.27	AD	1.32	4/7DF	1.40	GD	1.47
RFPr	1.25	RFPr	1.31	GD	1.33	NAUSEA/T	1.38
NAUSEA/T	1.22	GD	1.28	NAUSEA/T	1.31	DYSURIA	1.25

Table 4.1: Distribution of the most salient motives for readmission to the ED, for various readmission thresholds. Motive codes: GAP: general abdominal pain, 24HF: fever <24 hours, HA: headache, 1/3DF: fever between 1 and 3 days, GD: general discomfort, EP: epigastric pain, T: throwing up, D: diarrhea, LegP: leg pain, LuP: lumbar pain, AD: acute disnea, IFPr: pain in the right iliac fossa, RFPl: pain in the left renal fossa, RFPr: pain in the right renal fossa.

Readmission_adults							
<3 days		<7 days		<15 days		<30 days	
Motive	%	Motive	%	Motive	%	Motive	%
OTHERS	16.57	OTHERS	16.70	OTHERS	17.16	OTHERS	17.01
GAP	11.08	GAP	10.38	GAP	9.62	GAP	9.59
HA	4.94	HA	4.95	HA	4.75	HA	4.64
EP	3.44	LegP	3.09	LegP	3.27	LegP	3.24
LuP	2.81	EP	3.05	LuP	2.96	LuP	2.94
LegP	2.73	LuP	2.88	EP	2.80	EP	2.66
RFPr	2.32	RFPr	2.36	GD	2.35	GD	2.66
RFPl	2.27	GD	2.30	AD	2.21	AD	2.52
IFPr	2.18	RFPl	2.09	RFPr	1.97	24HF	2.14
GD	2.13	HYPr	1.95	24HF	1.94	RFPr	1.81
HYPr	2.10	AD	1.87	HYPr	1.82	D	1.80
D	1.88	IFPr	1.80	RFPl	1.77	HYPr	1.73
24HF	1.75	24HF	1.80	D	1.75	RFPl	1.59
AD	1.67	D	1.76	KP	1.55	KP	1.50
NAUSEA/T	1.37	NAUSEA/T	1.47	AP	1.54	NAUSEA/T	1.49

Table 4.2: Distribution of the most salient motives for readmission of adults to the ED, for various readmission thresholds. Motive codes: GAP: general abdominal pain, 24HF: fever <24 hours, HA: headache, 1/3DF: fever between 1 and 3 days, GD: general discomfort, EP: epigastric pain, T: throwing up, D: diarrhea, LegP: leg pain, LuP: lumbar pain, AD: acute disnea, IFPr: pain in the right iliac fossa, RFPl: pain in the left renal fossa, RFPr: pain in the right renal fossa, HYPr:right hypochondrium pain, D:diarrhea, KP: knee pain, AP: arm pain

Readmission_pediatrics							
<3 days		<7 days		<15 days		<30 days	
Motive	%	Motive	%	Motive	%	Motive	%
OTHERS	45.60	OTHERS	40.23	OTHERS	35.95	OTHERS	31.12
1/3DF	10.36	COUGH	9.73	COUGH	11.28	COUGH	13.56
COUGH	7.59	1/3DF	9.68	24HF	9.45	24HF	12.20
24HF	6.78	24HF	7.30	1/3DF	9.43	1/3DF	9.42
GAP	4.92	T	5.52	T	5.58	T	6.03
T	4.89	GAP	5.12	GAP	4.86	GAP	4.65
D	3.39	D	4.31	D	4.20	D	4.01
4/7DF	2.49	4/7DF	3.16	4/7DF	2.67	4/7DF	2.26
FP3	1.06	EARACHE	1.28	EARACHE	1.63	EARACHE	1.81
NAUSEA/T	1.06	EXANTH	1.09	NAUSEA/T	1.12	NAUSEA/T	1.25
EXANTH	1.00	FP3	1.09	DYSURIA	1.10	DYSURIA	1.16
EARACHE	0.90	NAUSEA/T	1.05	FP3	1.06	HA	1.02
HA	0.87	HA	0.86	EXANTH	1.04	FP3	1.01
AD	0.68	DYSURIA	0.81	HA	0.98	EXANTH	0.90
DYSURIA	0.62	AD	0.64	AD	0.72	CRYING	0.71

Table 4.3: Distribution of the most salient motives for readmission of pediatrics to the ED, for various readmission thresholds. Motive codes: GAP: general abdominal pain, 24HF: fever <24 hours, HA: headache, 1/3DF: fever between 1 and 3 days, GD: general discomfort, EP: epigastric pain, T: throwing up, D: diarrhea, LegP: leg pain, LuP: lumbar pain, AD: acute disnea, IFPr: pain in the right iliac fossa, RFPl: pain in the left renal fossa, RFPr: pain in the right renal fossa, FP3: pediatric fever 3 years, 4/7DF: fever between 4 and 7 days

Hospitalization (after readmission in X days)							
<3 days		<7 days		<15 days		<30 days	
Motive	%	Motive	%	Motive	%	Motive	%
OTHERS	21.57	OTHERS	20.49	OTHERS	19.67	OTHERS	19.49
GAP	12.74	GAP	12.68	GAP	12.68	GAP	12.33
HA	3.57	HA	3.44	AD	3.77	AD	4.40
1/3DF	3.25	1/3DF	3.29	24HF	3.49	24HF	4.25
HYPr	3.12	AD	3.00	HA	3.22	COUGH	3.04
EP	3.05	24HF	3.00	1/3DF	3.10	1/3DF	2.97
AD	2.79	HYPr	2.95	COUGH	2.87	HA	2.91
24HF	2.79	COUGH	2.95	D	2.79	D	2.57
COUGH	2.73	EP	2.80	EP	2.79	HYPr	2.57
D	2.53	D	2.75	HYPr	2.75	EP	2.51
IFPr	2.47	RFPr	2.75	RFPr	2.32	GD	2.42
RFPr	2.34	IFPr	2.11	LegP	2.28	LegP	2.32
T	2.27	T	2.06	T	1.85	RFPr	2.14
RFPl	1.75	RFPl	1.97	IFPr	1.81	LuP	1.61
LegP	1.69	LegP	1.87	GD	1.73	IFPr	1.58

Table 4.4: Distribution of the most salient motives for readmission to the ED that lead to hospitalization, for various readmission thresholds. Motive codes: GAP: general abdominal pain, 24HF: fever <24 hours, HA: headache, 1/3DF: fever between 1 and 3 days, GD: general discomfort, EP: epigastric pain, T: throwing up, D: diarrhea, LegP: leg pain, LuP: lumbar pain, AD: acute disnea, IFPr: pain in the right iliac fossa, RFPl: pain in the left renal fossa, RFPr: pain in the right renal fossa, HYPr:right hypochondrium pain

Hospitalization_adults (after readmission in X days)							
<3 days		<7 days		<15 days		<30 days	
Motive	%	Motive	%	Motive	%	Motive	%
OTHERS	15.76	OTHERS	15.37	OTHERS	15.48	OTHERS	16.05
GAP	14.47	GAP	14.03	GAP	13.85	GAP	13.47
HA	4.48	HA	4.25	HA	3.89	AD	4.55
HYPr	4.13	HYPr	3.81	AD	3.84	24HF	3.91
EP	3.96	EP	3.56	EP	3.45	HA	3.45
IFPr	3.10	RFPr	3.56	HYPr	3.45	HYPr	3.15
RFPr	3.10	AD	2.98	24HF	3.15	EP	3.03
AD	2.50	IFPr	2.60	RFPr	2.91	GD	2.96
RFPl	2.33	24HF	2.60	LegP	2.86	LegP	2.85
LegP	2.24	RFPl	2.54	IFPr	2.17	RFPr	2.62
24HF	2.15	LegP	2.41	GD	2.17	D	2.09
LuP	1.89	D	1.90	D	2.12	LuP	1.97
D	1.72	GD	1.84	RFPl	2.12	RFPl	1.90
GD	1.72	LuP	1.78	LuP	1.87	IFPr	1.82
1/3DF	1.46	1/3DF	1.78	1/3DF	1.72	1/3DF	1.82

Table 4.5: Distribution of the most salient motives for readmission of adults to the ED that lead to hospitalization, for various readmission thresholds. Motive codes: GAP: general abdominal pain, 24HF: fever <24 hours, HA: headache, 1/3DF: fever between 1 and 3 days, GD: general discomfort, EP: epigastric pain, T: throwing up, D: diarrhea, LegP: leg pain, LuP: lumbar pain, AD: acute disnea, IFPr: pain in the right iliac fossa, RFPl: pain in the left renal fossa, RFPr: pain in the right renal fossa, HYPr:right hypochondrium pain

Hospitalization_pediatrics (after readmission in X days)							
<3 days		<7 days		<15 days		<30 days	
Motive	%	Motive	%	Motive	%	Motive	%
OTHERS	39.42	OTHERS	38.04	OTHERS	36.10	OTHERS	34.86
COUGH	9.52	COUGH	10.87	COUGH	11.58	COUGH	13.54
1/3DF	8.73	1/3DF	8.48	1/3DF	8.49	1/3DF	8.12
GAP	7.41	GAP	8.04	GAP	8.11	GAP	7.28
T	7.14	T	7.17	T	6.95	T	6.09
D	5.03	D	5.65	D	5.41	24HF	5.75
24HF	4.76	24HF	4.35	24HF	4.83	D	4.74
AD	3.70	4/7DF	3.48	AD	3.47	AD	3.72
4/7DF	2.65	AD	3.04	4/7DF	3.28	4/7DF	3.55
NAUSEA/T	1.59	NAUSEA/T	1.30	CC	1.16	CC	1.18
CC	1.06	CC	1.09	NAUSEA/T	1.16	NAUSEA/T	1.02
HA	0.79	HA	0.65	F>1W	0.97	F>1W	0.85
AAP	0.53	F>1W	0.65	HA	0.58	FP3	0.68
IFPr	0.53	FP3	0.65	FP3	0.58	JAUNDICE	0.68
F>1W	0.53	AAP	0.43	JAUNDICE	0.58	HA	0.51

Table 4.6: Distribution of the most salient motives for readmission of pediatric patients to the ED that lead to hospitalization, for various readmission thresholds. Motive codes: GAP: general abdominal pain, 24HF: fever <24 hours, HA: headache, 1/3DF: fever between 1 and 3 days, GD: general discomfort, EP: epigastric pain, T: throwing up, D: diarrhea, AD: acute disnea, IFPr: pain in the right iliac fossa, RFPl: pain in the left renal fossa, RFPr: pain in the right renal fossa, HYPr:right hypochondrium pain; F>1W: fever > 1 week; 4/7DF: fever between 4 and 7 days; FP3: pediatric fever 3 years; CC: Convulsive crisis

Balan.	Adult				Pediatric			
	A	S	PPV	F	A	S	PPV	F
UNDER	0.64	**0.61**	0.43	**0.51**	0.66	**0.60**	0.15	**0.29**
OVER	0.65	0.48	0.44	0.45	0.73	0.49	0.20	**0.30**
SMOTE	0.63	0.58	0.42	0.49	0.80	0.35	0.25	0.28
OSS	**0.69**	0.42	**0.49**	0.45	**0.90**	0.10	0.47	0.21
TOMEK	**0.69**	0.42	**0.50**	0.45	**0.90**	0.10	**0.50**	0.21

Table 4.7: Average performance of the data balancing methods across classifier training methods and readmission threshold. Undersampling majority class (UNDER), Oversampling minority class (OVER), random interpolation (SMOTE), selective removal (TOMEK), one side selection (OSS). Bold indicates that the approach difference is significant ($p<0.001$) in all pairs of t-test one side comparisons.

Classif.	Adult				Pediatric			
	A	S	PPV	F	A	S	PPV	F
MLP	0.67	**0.57**	0.47	**0.51**	0.79	**0.37**	0.21	0.21
DT	0.66	0.49	0.45	0.46	0.81	0.31	0.37	0.28
kNN	0.62	0.51	0.40	0.45	0.75	0.34	0.26	0.24
NB	**0.69**	0.44	**0.51**	0.47	**0.84**	0.29	**0.42**	**0.30**

Table 4.8: Average performance of the classifier training methods across data balancing methods and readmission threshold. Multilayer perceptron (MLP), Decision Tree (DT), k nearest neighbors (kNN), Naive Bayes (NB). Bold indicates that the approach difference is significant ($p<0.001$) in all pairs of t-test one side comparisons.

Readm. Thresh.	Adult				Pediatric			
	A	S	PPV	F	A	S	PPV	F
30 days	**0.68**	**0.51**	**0.47**	**0.49**	**0.82**	**0.34**	0.32	0.25
15 days	0.67	**0.51**	0.46	0.48	0.81	0.32	0.30	0.24
7 days	0.66	**0.51**	0.45	0.47	0.80	0.33	**0.35**	**0.28**
3 days	0.65	0.50	0.46	0.47	0.78	**0.34**	0.31	**0.28**

Table 4.9: Average performance of the readmission threshold across classifier training and data balancing methods. Bold indicates that the approach difference is significant ($p<0.001$) in all pairs of t-test one side comparisons.

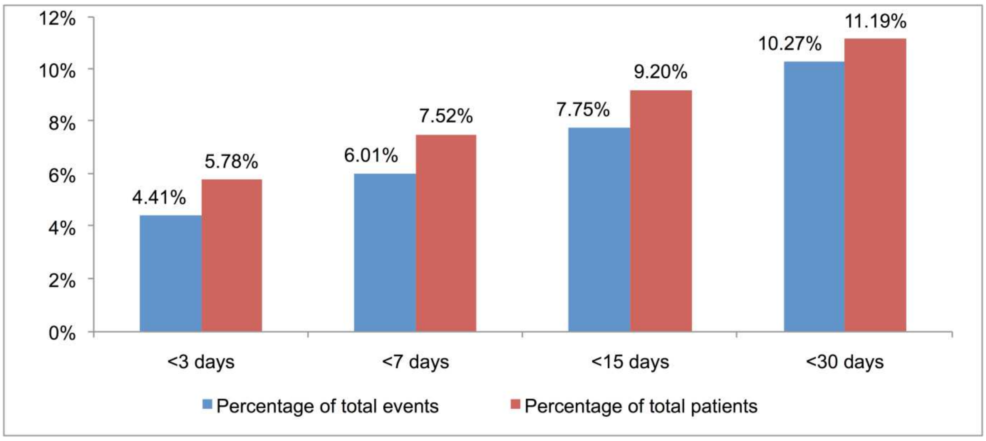
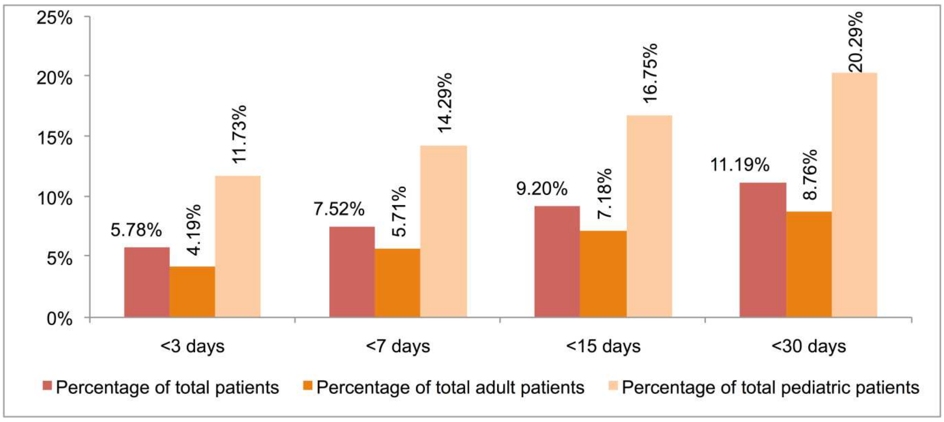
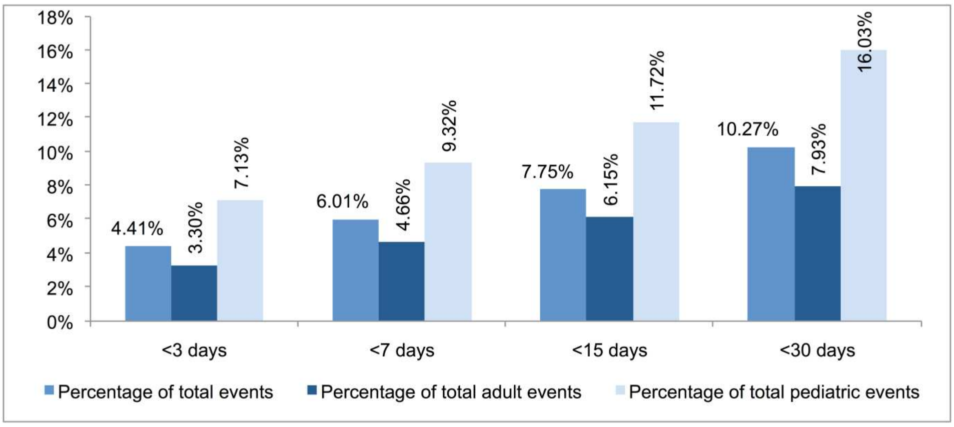


Figure 4.1: Percentage of patients suffering readmission, and percentage of readmission events relative to the number of patients and events, respectively



(a)



(b)

Figure 4.2: Decomposition of the trends for patients (fig. 4.2a) and events (fig. 4.2b) of the readmission of adults and pediatric according to the readmission threshold.

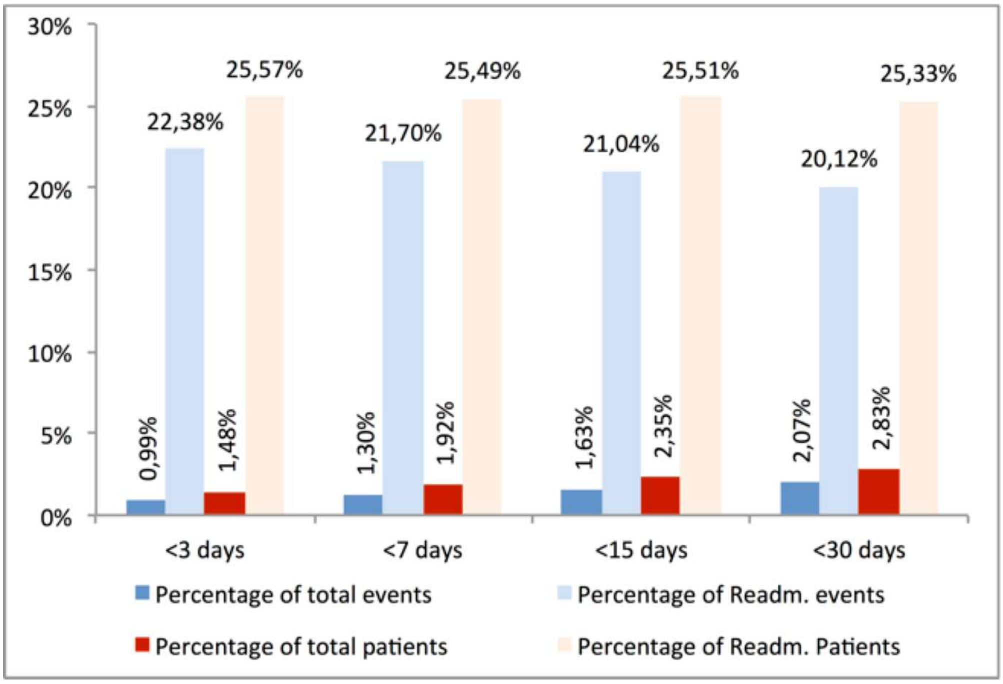
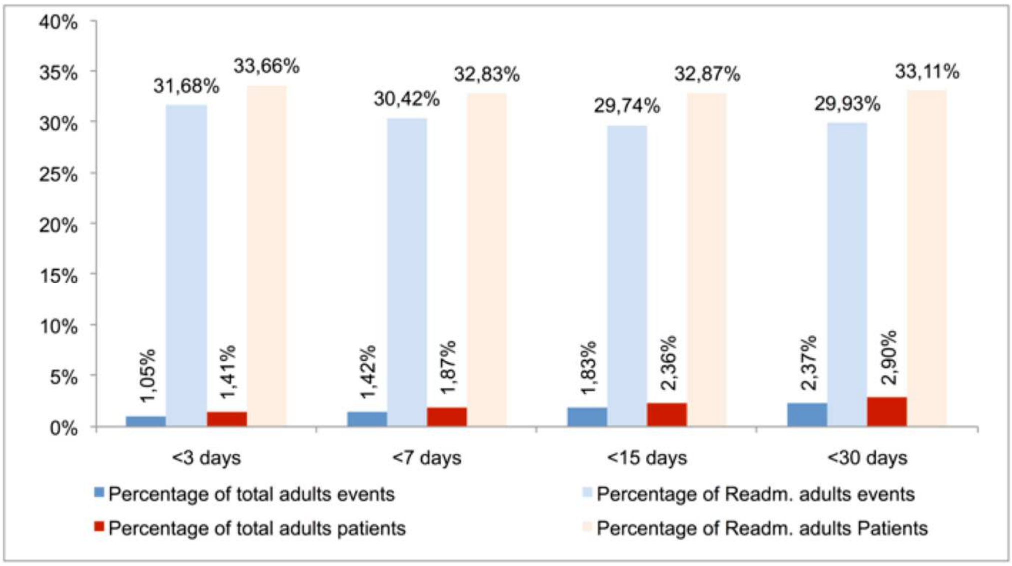
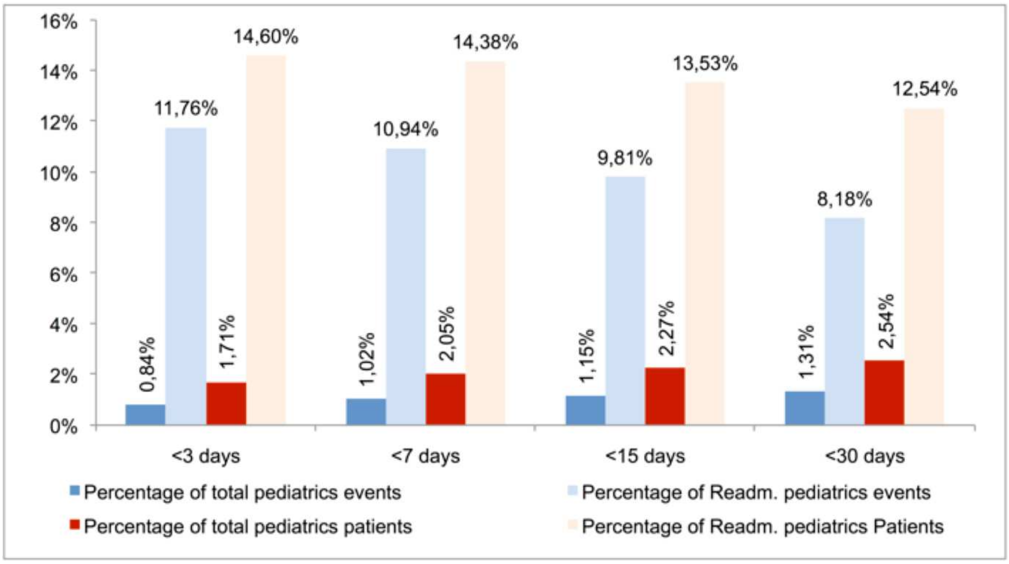


Figure 4.3: Distribution of the number of readmission leading to hospitalization according to the threshold for readmission.



(a)



(b)

Figure 4.4: Distribution of the number of readmission leading to hospitalization according to the threshold for readmission. (a) adult population, (b) pediatric population.

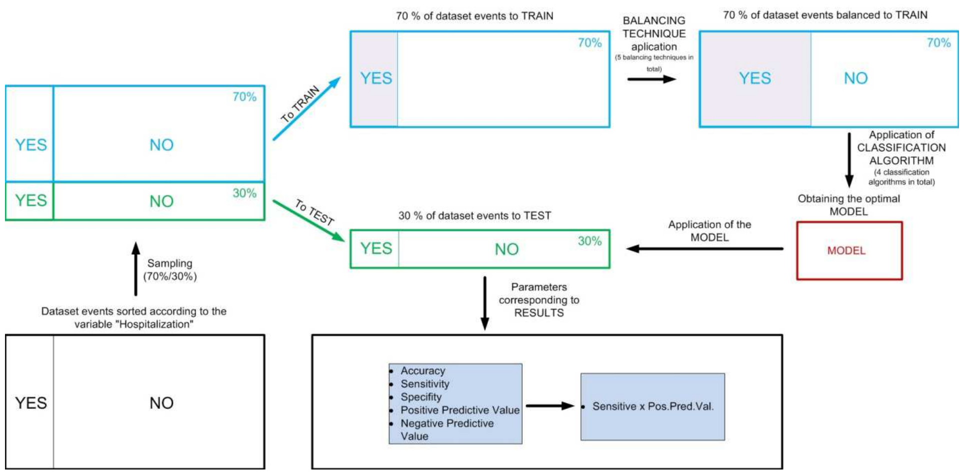


Figure 4.5: Distribution of the dataset into training and test datasets in order to avoid circular effects and biasing of test results by training data misuse.

Chapter 5
Conclusions and ideas for future work

This chapter concludes the Thesis giving some conclusions and ideas for future work. We discuss some aspects of processing health care data. We tackle independently the main applications.

5.1Processing health care data

The processing of health care data poses several challenges. The first and paramount is the availability of data. Legal and economical issues make quite tricky to obtain data to carry out the computational experiments. In the best case it requires a good deal of work to capture the data systematically. Once the data has been obtained, the intent of the original data capture may be different to the actual research work carried out, so that critical information may be missing, while redundant information may be overflowing. Besides, data can be very noisy even with the use of electronic health records. For instance, in the readmission data we have detected an extraordinary abundance of “OTHER” as the motive of the visit. This fact has two interpretations: (a) the nurse was overwhelmed by the work and the interface of the system was cumbersome, (b) the actual motives considered were not comprehensive enough. Either interpretation poses a lot of questions for the design and the working environment. We must have in mind that the systems are embedded in a clinical stressful scenario.

The noise in the data includes missing values, as well as erroneous values, so that missing value imputation may be a promising field of work for future research efforts. Intelligent techniques guessing the data values may include fuzzy system approaches as well as bayesian estimation based on Markov random field models. Appropriate methods have a great value and may improve performance of ensuing classification methods.
Furthermore, data distributions in health care tend to be uncomfortable, i.e. with some properties that make them difficult to treat. One such properties is the imbalance of classes in the data, where often the emphasis is in the minority class. Though there is a large body of literature on the topic of class balancing, most is devoted to continuous data that allow interpolation with good results. However, in the health care domain many variables are qualitative, so that numerical interpolation is meaningless for them. Dealing with qualitative imbalanced datasets is an open question of great interest for health care data.

Finally, data in the health care environment is becoming huge, so it may fall in the realm of Big Data, techniques and methods developed to deal with big data may have straightforward application in health care.

5.2 Monitoring of pediatric respiratory crisis

We have dealt with the monitoring and alarm detection in pediatric intensive care unit as a classification problem, trying to predict the triage set by the nurse and doctor. Data were records of a small number of patients taken each hour. We have successfully tackled the noise and the missing values, achieving high predictive performance. The work in this Thesis sets the stage for real life clinical applications that may help to save many lives. However, data was manually captured, so there is a need to develop and implement electronic devices for the capture of the data, possibly interfacing between equipments. Such devices may be mobile, like the tablets and iPad, with easy interfaces, and may implement the alarms built from the classifiers tested in this Thesis. They can embed some life-long learning procedure, so that they can adapt to changing conditions and populations.

5.3 Prediction of readmission and hospitalization

Hospitalization after readmission is a critical event with strong economical and health costs. The actual dataset is big and very imbalanced; therefore we had to focus on the readmitted people population, predicting the hospitalization event in such environment. We have explored state of the art data balancing methods and classifier learning methods, with some relative success. So we were able to make some recommendation for potential real life prediction products. A note on the readmission threshold, which is set based on economical and political reasons: we have tested several definitions (72hours, 7, 15, 30 days) finding that our approach is rather robust against the precise setting of the readmission threshold. We attribute the limitation of the results to the coding of the Motive variable. New codifications of this variable, such as a collection of orthogonal binary variables one per motive, may help to improve results. But this kind of state variables definition introduces some problems on the data interpolation carried out for class balancing. Research in this line of work may contribute greatly to the literature on class data balancing. Further statistical work may be done based on survival theory, which is a branch of stochastic processes that has been exploited in areas such as insurance risk assessment, and cancer survival.

Appendix A
Classifier training methods

A.1 Methods

We tackle the hospitalization prediction as a classification problem into two classes. The class distribution in the dataset is strongly unbalanced. Moreover, the statistical and qualitative differences between adult and pediatric populations counsel to treat them separately. The computational tools have been borrowed from the R project, including the methods for cross-validation (package caret), and the data balancing methods for unbalanced datasets (package unbalanced) introduced by [DAL2013].

To build hospitalization event predictors from the physiological and demographic features we have tried four kinds of classification algorithms: Multilayer Perceptron (MLP), Decision Trees (DT), k-Nearest Neighbors (k-NN), and Naive Bayes (NB). We have applied the default settings of these algorithms in the caret package: the same number of hidden units as inputs in the MLP, one neighbor in k-NN, and a maximum of 10 levels in DT. Details of the algorithms are well known and can be found elsewhere (Graña et al. 2015, Haykin 1998)

Appendix B
Class balancing techniques

B.1 Class Balancing

Real life classification problems are often imbalanced, that is, the number of samples of one class is much greater than the other. Most classifier building approaches are biased towards the majority class, so that they achieve high classification accuracy but low sensitivity on the minority class, which is often the interesting one. Therefore, accuracy is less relevant as a measure of performance than other measures focused on the positive minority class prediction. The case at hand fits into this picture, because the target hospitalization event is much scarcer than the readmission, and it has specific economic value. There are two basic ways of dealing with the issue of class imbalance. One manipulates the cost function weighting differently training examples, so the errors committed on the minority class are most costly. The other pre-processes the original dataset, either by over- sampling the minority class and/or under-sampling the majority class. In this paper we apply five methods following the later approach:

- Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al. 2002) consists in the random linear interpolation between nearest neighboring samples of the minority class. Notice that SMOTE may “fill the gaps” in data distributions that show disperse connected regions.
- The under-sampling (UNDER) consisting in randomly removing samples of the majority class until the desired balance is reached.
- The oversampling (OVER) of the minority class, consisting in the random repetition of some of the samples.
- The method proposed by (Tomek 1976) (TOMEK) that consists in the removal of samples that do not affect the performance of 1-NN classifiers.
- The one sided selection (Kubat 1997) (OSS) focus on the positive class, retaining all samples and removing all redundant samples, which are far from the decision boundary.

References

[ABU12] AbuKhousa, E., Mohamed, N., & Al-Jaroodi, J. (2012). e-Health cloud: opportunities and challenges. Future Internet, 4(3), 621-645.

[ABU15] R. Abusamra and R. R. Russell, “Management of respiratory disease in children with muscular weakness,” Paediatrics and Child Health, vol. 25, no. 11, pp. 515 – 521, 2015.

[ACH16] G. Schmalisch, “Basic principles of respiratory function monitoring in ventilated newborns: A review,” Paediatric Respiratory Reviews, pp. –, 2016.

[AMA13] Filippo Amato, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, Aleš Hampl, Josef Havel “Artificial neural networks in medical diagnosis,” Journal of Applied Biomedicine, vol. 11, no. 2, pp. 47 – 58, 2013.

[BAG01] S. C. Bagley, H. White, and B. A. Golomb, “Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain,” Journal of Clinical Epidemiology, vol. 54, no. 10, pp. 979 – 985, 2001.

[BAT07] T. F. Bateson and J. Schwartz, “Children’s response to air pollutants,” Journal of Toxicology and Environmental Health, Part A, vol. 71, no. 3, pp. 238–243, 2007.

[BAT14] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: Using analytics to identify and manage high-risk and high-cost patients,” Health affairs, 2014.

[BES15] Besga, A., B. Ayerdi, G. Alcalde, A. Manzano, P. Lopetegui, M. Graña, A. Gonzalez-Pinto "Risk factors for emergency department short time readmission in stratified population", *BioMed Research International* 2015 (2015) Article ID 685067.

[BLO10] Bloom, D. E., Canning, D., & Fink, G. (2010). Implications of population ageing for economic growth, Oxford Review of Economic Policy, 26(4), 583-612.

[BON13] C. P. Bonafide, P. W. Brady, R. Keren, P. H. Conway, K. Marsolo, and C. Daymont, “Development of heart and respiratory rate percentile curves for hospitalized children,” Pediatrics, 2013.

[BRA06] A. R. Brady, D. Harrison, S. Black, S. Jones, K. Rowan, G. Pearson, J. Ratcliffe, and G. J. Parry, “Assessment and optimization of mortality prediction tools for admissions to pediatric intensive care in the united kingdom,” Pediatrics, vol. 117, no. 4, pp. e733–e742, 2006.

[BRA16] A. Brand, K. E. McLean, S. B. Henderson, M. Fournier, L. Liu, T. Kosatsky, and A. Smargiassi, “Respiratory hospital admissions in young children living near metal smelters, pulp mills and oil refineries in two canadian provinces,” Environment International, vol. 94, pp. 24 – 32, 2016.

[BRE01] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[CAR11] Carpenter, C., K. Heard, S. Wilber, A. Ginde, K. Stiffler, L. Gerson, et al., "Research priorities for high-quality geriatric emergency care: medication management, screening, and pre- vention and functional assessment", *Acad Emerg Med* 18, no 6, (2011):644–5.

[CARET] Caret package http://topepo.github.io/caret/index.html

[CHA02] Chawla, N.V., K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer. "Smote: Synthetic minority over-sampling technique". *Journal of Artificial Intelligent Research*, 16 (2002):321–357.

[CHA13] A. M. Chan, N. Selvaraj, N. Ferdosi, and R. Narasimhan, “Wireless patch sensor for remote monitoring of heart rate, respiration, activity, and falls,” in Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, July 2013, pp. 6115–6118.

[CHE16] J. H. Chen, T. Podchiyska, and R. B. Altman, “Orderrex: clinical order decision support and outcome predictions by data-mining electronic medical records,” Journal of the American Medical Informatics Association, vol. 23, no. 2, pp. 339–348, 2016.

[CHI93] Chi, R. T., Chen, M., & Kiang, M. Y. (1993). Generalized case-based reasoning system for portfolio management. Expert Systems with Applications, 6(1), 67-76

[CHO09] Choy, K. L., Chow, K. H., Moon, K. L., Zeng, X., Lau, H. C. W., Chan, F. T. S., & Ho, G. T. S. (2009). A RFID-case-based sample management system for fashion product development. Engineering Applications of Artificial Intelligence, 22(6), 882-896

[CRO15] W. H. Crown, “Potential application of machine learning in health outcomes research and some statistical cautions,” Value in Health, vol. 18, no. 2, pp. 137 – 140, 2015.

[CRU07] A. Cruz and et al., Global surveillance, prevention and control of chronic respiratory diseases A comprehensive approach, J. Bousquet and N. Khal- taev, Eds. WHO, 2007.

[DAL13] Dal Pozzolo, A, O. Caelen, S. Waterschoot, G. Bontempi "Racing for unbalanced methods selection", In Proceedings of Intelligent Data Engineering and Automated Learning - IDEAL'13., pages 24-31. Springer Verlag 2013

[DEPT] Departamento de estadísticas, Ministerio de Salud, Chile“http://www.deis.cl/.”

[DES15] Deschodt, M., E. Devriendt, M. Sabbe, D. Knockaert, P. Deboutte, S. Boonen, J. Flamaing, K. Milisen, "Characteristics of older adults admitted to the emergency department (ed) and their risk factors for ED readmission based on comprehensive geriatric assessment: a prospective cohort study", *BMC geriatrics* 15, no 1, (2015):54.

[FAY96] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “Advances in knowledge discovery and data mining,” U. M. Fayyad, G. Piatetsky- Shapiro, P. Smyth, and R. Uthurusamy, Eds. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996, ch. From Data Mining to Knowledge Discovery: An Overview, pp. 1–34.

[FIS65][22] A. Fisher, A synopsis of respiratory diseases. John Wright & Sons Ltd., Bristol,, 1965.

[FLE09] Fleischer, S., Berg, A., Zimmermann, M., Wüste, K., & Behrens, J. (2009). Nurse-patient interaction and communication: A systematic literature review. Journal of Public Health, 17(5), 339-353.

[FLE11] S. Fleming, M. Thompson, R. Stevens, C. Heneghan, A. Plüddemann, I. Maconochie, L. Tarassenko, and D. Mant, “Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies,” The Lancet, vol. 377, no. 9770, pp. 1011 – 1018, 2011.

[GAR14] P. A. Garcia-Chevesich, S. Alvarado, D. G. Neary, R. Valdes, J. Valdes, J. J. Aguirre, M. Mena, R. Pizarro, P. Jofré, M. Vera, and C. Olivares, “Respiratory disease and particulate air pollution in santiago chile: Contribution of erosion particles from fine sediments,” Environmental Pollution, vol. 187, pp. 202 – 205, 2014.

[GRA15] Graña, M., J.D. Nuñez-Gonzalez, L. Ozaeta, A. Kaminska-Chuchmała "Experiments of Trust Prediction in Social Networks by Artificial Neural Networks", *Cybernetics and Systems* 46, no 1-2, (2015):19-34

[HAN10] D. Han, M. Lee, and S. Park, “The-muss: Mobile u-health service system,” Computer Methods and Programs in Biomedicine, vol. 97, no. 2, pp. 178 – 188, 2010.

[HAN11] Han, J., Pei, J., & Kamber, M. (2011). Data mining: Concepts and techniques. Elsevier.

[HAO15] Hao, S., Y. Wang, B. Jin, A. Y. Shin, C. Zhu, M. Huang, L. Zheng, J. Luo, Z. Hu, C. Fu, D. Dai, Y. Wang, D. S. Culver, S. T. Alfreds, T. Rogow, F. Stearns, K. G. Sylvester, E. Widen, X. B. Ling, "Development, validation and deployment of a real time 30 day hospital readmission risk assessment tool in the Maine healthcare information exchange", PLoS ONE 10, no 10 (2015):1–15.

[HAR15] R. A. Haraty, M. Dimishkieh, and M. Masud, "An enhanced k-means clustering algorithm for pattern discovery in healthcare data," Int. J. Distrib. Sen. Netw., vol. 2015, pp. 7:7–7:7, Jan. 2015.

[HAY98] Haykin, S., *Neural Networks: a comprehensive foundation* (2nd edition). Prentice Hall, 1998

[HE13] He, C., Fan, X., & Li, Y. (2013). Toward ubiquitous healthcare services with a novel efficient cloud platform. IEEE Transactions on Biomedical Engineering, 60(1), 230-234.

[HER14] M. Herland, T. M. Khoshgoftaar, and R. Wald, "A review of data mining using big data in health informatics," Journal Of Big Data, vol. 1, no. 1, pp. 1–35, 2014.

[KAN11] Kansagara, D., H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, S. Kripalani, "Risk prediction models for hospital readmission: A systematic review", *JAM* 306 no 15, (2011) 1688-1698.

[KHA14] [16] H. S. Khamis, K. W. Cheruiyot, and S. Kimani, "Application of k-nearest neighbour classification in medical data mining," International Journal of Information, vol. 4, no. 4, 2014.

[KIM16] Kim, J., Han, M., Lee, Y., & Park, Y. (2016). Futuristic data-driven scenario building: Incorporating text mining and fuzzy association rule mining into fuzzy cognitive map. Expert Systems with Applications, 57, 311-323.

[KOV08] Kovach, C. R., Morgan, S., Noonan, P. E., & Brondino, M. (2008). Using principles of diffusion of innovation to improve nursing home care. Journal of Nursing Care Quality, 23(2), 132-139

[KUB97] Kubat, M., S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection". Paper presented at the Machine Learning-International Workshop, pages 179-186. Morgan Kaufmann Publishers, Inc., 1997.

[LAI12] Lai, I. K.W., Tam, S. K. T., & Chan, M. F. S. (2012). Knowledge cloud system for network collaboration: A case study in medical service industry in China. Expert Systems with Applications, 39(15), 12205-12212.

[LIN14] Lin, C. W., Abdul, S. S., Clinciu, D. L., Scholl, J., Jin, X., Lu, H., Chen, S. S., Iqbal, U., Heineck, M. J., & Li, Y. C. (2014). Empowering village doctors and enhancing rural healthcare using cloud computing in a rural area of mainland China. Computer Methods and Programs in Biomedicine, 113(2), 585-592.

[LOP13] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, Information Sciences 250 (2013) 113– 141.

[MAP16] Mapa De Calidad Del Aire De La Región Metropolitana, Santiago, Chile “http://portal.mma.gob.cl/pronostico-rm/.”

[MOO16] R. J. Moore and J. L. Hotchkiss, “The importance of toxicity in determining the impact of hazardous air pollutants on the respiratory health of children in tennessee,” Environmental Pollution, pp. –, 2016.

[NEU14] Neustein, A., Imambi, S. S., Rodrigues, M., Teixeira, A., & Ferreira, L. (2014). Application of text mining to biomedical knowledge extraction: analyzing clinical narratives and medical literature. Text Mining of Web-based Medical Content. De Gruyter, Berlin, 3-31.

[NGU14] Nguyen, H. Q., L. Chu, I.-L. Amy Liu, J. S. Lee, D. Suh, B. Korotzer, G. Yuen, S. Desai, K. J. Coleman, A. H. Xiang, M. K. Gould, "Associations between physical activity and 30-day readmission risk in chronic obstructive pulmonary disease", *Annals ATS* 11 (5) (2014) 695–705.

[NIR15] R. Nirel, N. Maimon, E. Fireman, S. Agami, A. Eyal, and A. Peretz, “Respiratory hospitalizations of children living near a hazardous industrial site adjusted for prevalent dust: A case– control study,” International Journal of Hygiene and Environmental Health, vol. 218, no. 2, pp. 273 – 279, 2015.

[NIR15b] R. Nirel, M. Schiff, and O. Paltiel, “Respiratory hospitalizations of children and residential exposure to traffic air pollution in jerusalem," International Journal of Hygiene and Environmental Health, vol. 218, no. 1, pp. 34 – 40, 2015.

[NORM][40] Normal blood pressure "http://www.md-health.com/normal-blood-pressure-for-children.html."

[OLS16] Olson, C.H., S. Dey, V. Kumar, K. A. Monsen, B. L. Westra, "Clustering of elderly patient subgroups to identify medication-related readmission risks", *International Journal of Medical Informatics* 85, no 1, (2016):43–52.

[PAN09] Pandey, B., & Mishra, R. B. (2009). Knowledge and intelligent computing system in medicine. Computers in Biology and Medicine, 39(3), 215-230.

[PAR07] Parker, M. G., & Thorslund, M. (2007). Health trends in the elderly population: getting better and getting worse. The Gerontologist, 47(2), 150-158.

[PER15] Pereira, L., C. Choquet, A. Perozziello, M. Wargon, G. Juillien, L. Colosi, R. Hellmann, M. Ranaivoson, E. Casalino, "Unscheduled-return-visits after an emergency department (ED) attendance and clinical link between both visits in patients aged 75 years and over: a prospective observational study", *PloS one* 10 no 4 (2015) :e0123803.

[PER15b] C. B. Pereira, X. Yu, V. Blazek, and S. Leonhardt, "Robust remote monitoring of breathing function by using infrared thermography," in Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, Aug 2015, pp. 4250–4253.

[POD02] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, “Decision trees: An overview and their use in medicine,” J. Med. Syst., vol. 26, no. 5, pp. 445–463, Oct. 2002.

[POO09] Poon, T. C., Choy, K. L., Chow, H. K. H., Lau, H. C. W., Chan, F. T. S., & Ho, K. C. (2009). A RFID case-based logistics resource management system for managing order-picking operations in warehouses. Expert Systems with Applications, 36(4), 8277-8301.

[PRI07] Prieto C MJ, Mancilla F P, Astudillo O P, Reyes P A, Román A O., “Excess respiratory diseases in children and elderly people in a community of santiago with high particulate air pollution,” Revista médica de Chile, 2007 Feb;135(2):221-8. Epub 2007 Mar 28.

[RAD14] H. I. Rady, S. A. Mohamed, N. A. Mohssen, and M. ElBaz, “Application of different scoring systems and their value in pediatric intensive care unit,” Egyptian Pediatric Association Gazette, vol. 62, no. 3 - 4, pp. 59 – 64, 2014.

[RAV14] H. Ravishankar, A. Saha, G. Swamy, and S. Genc, “An early respiratory distress detection method with markov models,” in Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, Aug 2014, pp. 3438–3441.

[REC09] Rechel, B., Doyle, Y., Grundy, E., & McKee, M. (2009). How can health sytems respond to population ageing?. European Observatory on Health Systems and Policies, Policy Brief 10.

[ROB06] I. Robertson-Steel, “Evolution of triage systems,” Emergency Medicine Journal, vol. 23, no. 2, pp. 154–5, 2006.

[ROD13] Rodrigues, J. J., de la Torre, I., Fernández, G., & López-Coronado, M. (2013). Analysis of the security and privacy requirements of cloud-based electronic health records systems. Journal of Medical Internet Research, 15(8), e186.

[SCO05] Scott-Cawiezell, J., Jones, K., Moore, L., & Vojir, C. (2005). Nursing home culture: a critical component in sustained improvement.Journal of Nursing Care Quality, 20(4), 341-348

[SIL08] Silverstein,M.D., H. Qin, S. Q. Mercer, J. Fong, Z. Haydar, "Risk factors for 30-day hospital readmission in patients> 65 years of age", in: Baylor University Medical Center. Proceedings, Vol. 21, Baylor University Medical Center, 2008, p. 363.

[SOB89][21] H. R. Sobral, “Special issue: Geography of health air pollution and respiratory diseases in children in são paulo, brazil,” Social Science & Medicine, vol. 29, no. 8, pp. 959 – 964, 1989.

[SON04] X. Song, A. Mitnitski, J. Cox, and K. Rockwood, “Comparison of machine learning techniques with classical statistical models in predicting health outcomes,” Studies in health technology and informatics, vol. 107, no. Pt 1, pp. 736–740, 2004.

[STE16] M. Stewart, J. Evans, N. Letourneau, J. Masuda, A. Almond, and J. Edey, “Low-income children, adolescents, and caregivers facing respiratory problems: Support needs and preferences,” Journal of Pediatric Nursing, vol. 31, no. 3, pp. 319 – 329, 2016.

[STA14] Stantchev, V., Colomo-Palacios, R., & Niedermayer, M. (2014). Cloud computing based systems for healthcare. The Scientific World Journal, 2(2014), 3-4

[SUB11] C. P. Subbe, “Centile-based early warning scores derived from statistical distributions of vital signs,” Resuscitation, vol. 82, no. 8, pp. 969 – 970, 2011.

[SUL14] Sultan, N. (2014). Making use of cloud computing for healthcare provision: Opportunities and challenges. International Journal of Information Management, 34(2), 177-184.

[SUT15] S. Suthaharan, Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, ser. Integrated Series in Information Systems. Springer US, 2015.

[TOM76] Tomek, I. "Two modifications of CNN" *IEEE Trans. Syst. Man Cybern.*, 6, (1976):769-772,.

[URB15] E. Urbankowska, K. Krenke, Ł. Drobczyński, P. Korczyński, T. Ur- bankowski, M. Krawiec, G. Kra j, M. Brzewski, and M. Ku- lus, “Lung ultrasound in the diagnosis and monitoring of community acquired pneumonia in children,” Respiratory Medicine, vol. 109, no. 9, pp. 1207 – 1212, 2015.

[VAN10] van Walraven, C., I. A. Dhalla, C. Bell, E. Etchells, I. G. Stiell, K. Zarnke, P. C. Austin, A. J. Forster, "Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community", *Canadian Medical Association Journal* 182, no 6, (2010):551–557.

[VAN12] van Walraven, C., J. Wong, A. Forster, "Lace+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data", *Open Medicine* 6 (2012) 80–89.

[VAN15] van Walraven, C., F. A. McAlister, J. A. Bakal, S. Hawken, J. Donzé, External validation of the hospital-patient one-year mortality risk (homr) model for predicting death within 1 year after hospital admission, Canadian Medical Association Journal 187 (10) (2015) 725–733.

[WAN07] Wang, W. M., Cheung, C. F., Lee, W. B., & Kwok, S. K. (2007). Knowledge□ based treatment planning for adolescent early intervention of mental healthcare: a hybrid case□ based reasoning approach. Expert Systems, 24(4), 232-251.

[WHO16] World Health Organization "http://www.who.int/en/."

[XIN12] Xing, G., Ding, J., Chai, T., Afshar, P., & Wang, H. (2012). Hybrid intelligent parameter estimation based on grey case-based reasoning for laminar cooling process. Engineering Applications of Artificial Intelligence, 25(2), 418-429

[YU15] Yu, S., F. Farooq, A. van Esbroeck, G. Fung, V. Anand, B. Krishnapuram, "Predicting readmission risk with institution-specific prediction models", *Artificial Intelligence in Medicine* 65 (2) (2015) 89–96 □

[ZAI13] Zai, A. H., J. G. Ronquillo, R. Nieves, H. C. Chueh, J. C. Kvedar, K. Jethwani, "Assessing hospital readmission risk factors in heart failure patients enrolled in a telemonitoring program", *Int. J. Telemedicine Appl.* (2013): 305819.

[ZAP12] Zapatero, Antonio and Barba, Raquel and Marco, Javier and Hinojosa, Juan and Plaza, Susana and Losa, Juan Emilio and Canora, Jesus, “Predictive model of readmission to internal medicine wards” European journal of internal medicine, 23(5) (2012):451--456

[ZIM14] Zimmerman, S., Shier, V., & Saliba, D., (2014). Transforming nursing home culture: evidence for practice and policy. The Gerontologist, 54 (Suppl 1), S1-S5