UNIVERSITY OF SOUTHAMPTON

FACULTY OF BUSINESS AND LAW

School of Management

Modelling Patient Length of Stay in Public Hospitals in Mexico

by

Maria de Lourdes Guzman Castillo MSc

Thesis for the degree of Doctor of Philosophy

May 2012

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF BUSINESS AND LAW

SCHOOL OF MANAGEMENT

Doctor of Philosophy

Modelling Patient Length of Stay in Public Hospitals in Mexico

by María de Lourdes Guzmán Castillo

This thesis is concerned with the modelling of patient length of stay in public hospitals in Mexico. Patient length of stay is the most commonly worldwide employed outcome measure for hospital resource consumption and performance monitoring. Most of the hospitals around the world use average length of stay as starting point for resource planning. However average estimates frequently gives non-accurate results due to the high variability of the length of stay data. The reason for such high variability may be attributable to the diversity in the patient population and the environment where the patient is treated.

Through a systematic review of the literature on methods and models in the field of calculating and predicting patient length of stay, this research highlights the areas of opportunity and research gap from previous studies and practices, and proposes the use of finite mixture models to approximate the distribution of length of stay. Also, these models are proposed as the foundation of more sophisticated models designed to include the internal and external factors associated with LoS. In this context, the thesis proposes three different approaches to explore such factors: individual-based approach, group-based approach and multilevel group-based approach. These interrelated approaches allow a better understanding of the diversity in the patient population and enable length of stay predictions for individual patients, and for cohorts of patients within and between hospitals. In addition, this research is built and evaluated using data from all types of patients treated at two public hospitals operating in Mexico. It is the consideration of the full case-mix of these healthcare facilities that gives this research its unique nature.

List of contents

List of figures

List of tables

Declaration of authorship

I, Maria de Lourdes Guzman Castillo

declare that the thesis entitled

Modelling patient length of stay in public hospitals in Mexico

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed: ……..

Date:…….

Acknowledgements

I would like to thanks my supervisors Professor Sally Brailsford and Dr. Honora Smith for the assistance and advice they have given throughout the course of my PhD study. Especially for giving all the understanding and support that was needed through the critical moments of this journey.

Many thanks to Dr. Michelle Luke for the technical assistance, patience and valuable feedback.

Many people in Centro Medico ISSEMyM gave me much assistance. Especially I would like to thank Dr. Humberto Alegria, Dr. Jorge Munoz and Lic. Elia Enriquez for opening the doors at their hospital and providing the resources to conduct this research.

Especial thanks to Dra. Patricia Castaneda and Ing. Andres Gonzalez from the Maximiliano Ruiz Casteneda General Hospital for their trust and enthusiasm in this research.

An especial acknowledgment to my first lecturer in Operational Research: Dr. Maurice Levy Matarasso. His enthusiam and passion for OR were the triggers which initiated this journey.

I am also infinitely grateful to Consejo National de Ciencia y Tecnologia (CONACyT) who funded me through this research. Without its financial support this research would not have been possible.

My parents, Maria de Lourdes Castillo Arredondo and Juan Jose Guzman Ramirez, have supported me in every achievement throughout my life; thanks for their patience, support and unlimited love to endure my absence from home during these years.

Many thanks to my amazing, beautiful and smart sister Berenice Guzman Castillo, who is my engine, motivation and best friend.

A big thank-you to my beloved grandparents Ruben Castillo Morelos (†) and Guadalupe Arredondo Ortiz for their unconditional love and support. *"Ya ves abuelita? Si me viste"*

Thanks to my uncle Ruben Castillo Arredondo for his love and encouragement, not just during this journey but through all my life.

Thanks to all my friends here in Southampton and Mexico. *Friendship is unnecessary, like philosophy, like art... it has no survival value; rather is one of those things that give value to survival (C.S Lewis).*

Last but not least, I would very much like to thank Ali El Dirani for his support, encouragement and advice. Thanks to him, for making the years of my PhD studies exciting and unforgettable.

شكرا لانك اريتني الطريق الى الله

List of abbreviations

AIC	Akaike information criterion
ALoS	Average length of stay
AMI	Acute myocardial infarction
ANOVA	Analysis of variance
AQTM	Approximated quartile based truncated mean
ATAR	Average trimmed absolute residual
BIC	Bayesian information criterion
BOMPS	Bed-occupancy, management, and planning software
C4.5	Commercial version 4.5
CART	Classification and regression tree
CDF	Cumulative density function
CHAID	Chi-squared automatic interaction detection
CRM	Continuation ratio model
CV	Cross-validation
DCS	Discrete conditional survival
DRG	Diagnosed related group
DT	Decision tree
EBP	Error based pruning
EDA	Exploratory data analysis
EDF	Empirical distribution function
EM	Expectation-Maximization
FORF	First order random forest
GDP	Gross domestic product
GLM	Generalised linear models
GOLM	Generalised ordered logit model
HIV	Human immunodeficiency virus
ICD	International classification of disease
ICU	Intensive care unit
IIA	Independence of irrelevant alternatives
IMSS	Instituto Mexicano del Seguro Social
ISSEMyM	Instituto Mexicano de Seguridad Social del Estado de Mexico

LMT	Logistic model trees
LoS	Length of stay
LR	Likelihood ratio
MNLM	Multinomial logit model
MRC	Maximiliano Ruiz Castaneda
MSM	Multi-stage models
NBtree	Naive Bayes Tree
NGO	non-governmental organization
NHS	National Health System
OLS	Ordinary least square
ORM	Ordinal regression model
PH	Phase type distribution
PPOM	Partial proportional odds model
PROMPT	Patient resource operational management planning tool
QUEST	Quick, Unbiased, Efficient Statistical Trees
REP	Reduced error pruning
SA	Secretaria de Salud
SORM	Stereotype ordered regression model
TAN	Tree augmented naive bayes network
VIF	Variance inflation factor
VPC	Variance partition coefficient
VPN	Variance partition coefficient
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

1 INTRODUCTION

1.1. General Importance

Patient length of stay (LoS) is one of the most commonly worldwide employed outcome measures for hospital resource consumption and performance monitoring. It also provides a better understanding of the flow of patients through a healthcare system which is essential for understanding both the operational and clinical functions of such a system (Adeyemi and Chaussalet, 2009)

In both domains, LoS estimations has countless applications such as: assessing future bed usage, estimating forthcoming demands on various hospital resources, defining the case-mix, helping to understand the course of the patient disease and recovery, delineating health insurance plans and reimbursement systems in the private sector, planning discharges for elderly patients, dependent patients or any patient with especial needs and as a crucial variable for the quality of life of the patients and families (Ramakrishnan, 2012).

In addition, LoS is a frequent point of comparison between patients, hospitals and countries, where there is always constant pressure from external authorities to decrease LoS, given that reduction of LoS is believed to help the hospital administration. Hence, hospitals can improve their quality of service to the patients with the available resources which in turn reduces costs (Ramakrishnan, 2012) and improves the quality of life of patients.

Therefore the importance of getting accurate estimation of patient LoS is a crucial factor on the healthcare scene.

1.2. The Problem

Most of the hospitals in Mexico and around the world use average length of stay (ALoS) as starting point for resource planning. However an average estimate frequently gives non-accurate results that may lead to undesirable outcomes: The medical staff of one of the hospitals under study recognized that surgeries are frequently rescheduled or cancelled due to the lack of available beds. In addition, everyday a proportion of patients have to be admitted to temporal beds because of lack of beds in the correct ward. Moreover misleading information about the discharge date is frequently given to patient and families.

The common practice when calculating ALoS is to select a specific cohort of patients: Patients within the same ward, same type of diagnosis or other grouping criteria. The most common approach in Mexico is to calculate ALoS per diagnosis, specifically using the International Classification of Disease (ICD) coding generated by the World Health Organization (WHO). For example, the ALoS for patients with umbilical hernia without obstruction or gangrene is two days. However looking at the histogram of LoS distribution (Figure 1.1), one can notice that most of the patients (i.e. around 65%) stay only one day at hospital, while an important proportion of them stay more than two days. Furthermore the positive difference between the mean and the median (1 day) indicates that the data are skewed in nature with a long tail of the distribution to the right (i.e. positively skewed), which can be confirmed by visual inspection of the histogram. In addition the coefficient of variation of 142%, computed from the standard deviation (2.84 days) divided by the mean, indicates a very high variability of the data. As it was just stated, average estimations rarely capture the real nature of data structures such as LoS, especially in cases with such variability. In this context, Shahani (1981) demonstrated mathematically that when high variability is present, this can lead to a large error in an average.

The reason why the variability is so high, although the data comes from patients within the same diagnosis, may be attributable to other patients differences within the cohort such as severity of illness, medical complications, speed of recovery, discharge destination or social and demographic circumstances (El-Darzi et al., 2009). This diversity in the patient population is frequently referred as the heterogeneity problem which coupled with the uncertainty inherent within health care systems makes it complicated to plan for effective resource use (Harper, 2002).

In the last couple of years, The Instituto Mexicano del Seguro Social (IMSS) and Secretaria de Salud (SA), the biggest two healthcare providers in Mexico, have adopted the Diagnosed Related Groups (DRGs) methodology, which allows to classify patients with similar expected resource use, measured by LoS but incorporating patients characteristics as age, presence of comorbidities and complications. Unfortunately this methodology is based as well on averages

estimates; plus it is expensive to implement and it is usually recalled for hospital performance and financial reports. Moreover, as far as the researcher could observe during the initial interviews with the staff of the hospitals under study, there is no evidence that this methodology is used for day by day operations.

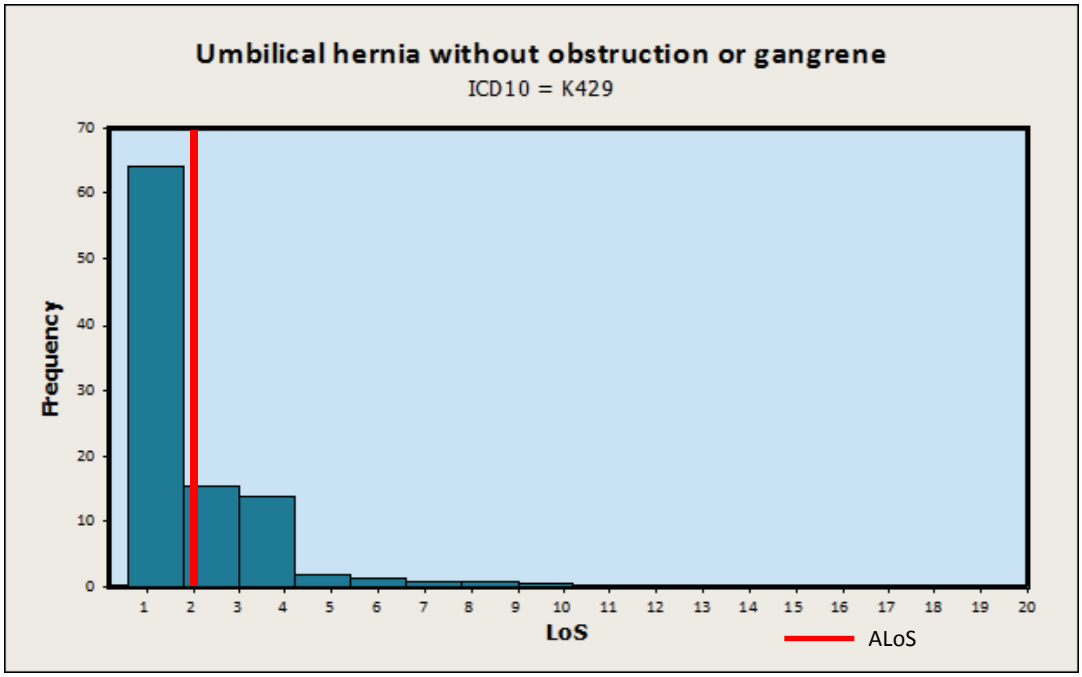

Figure 1.1: LoS for patient with umbilical hernia at hospital in Mexico City

In addition, the problem of heterogeneity goes far beyond the hospital scope: It is believed that LoS is influenced by institutional and national context (Rotter et al., 2010). According to the National Health Plan 2007 – 2012, the main problem of the Mexican national healthcare system in terms of quality and efficiency is this huge heterogeneity that exists among the main service providers. It is found that the ALoS for identical surgical procedures or diagnoses varies considerably across different healthcare providers: For example, the ALoS for appendectomies at the IMSS hospitals is 6.5 days against 3 days in hospitals in the State Services. The ALoS for inguinal hernioplasty in IMSS hospitals with over 120 beds is 1.5 days against 2.6 days in the State Services hospitals and the Secretariat of Health hospitals (Secretaria de Salud, 2007).

Independently, which is the cause, size and scope of the LoS variability, it seems clear that under such complexity a crude estimation of LoS in the form of an “average” finds lot of limitations to represent the nature of the problem.

1.3. Research Objectives and Questions

Taking into account the importance of accurately estimating LoS and the problems faced by healthcare decision makers in all levels of the Mexican healthcare system nowadays, there is a need for more complex and sophisticated tools to understand length of stay to help into the development of a common action plan to improve quality and efficiency of the healthcare services.

This research is geared towards developing a statistical model to predict patient length of stay (LoS) in Mexican public hospitals that:

A. Captures the variability of the LoS distribution
B. Recognises and addresses the heterogeneity problem
C. Supplies LoS predictions for individual patients and cohort of patients (within and between hospitals)
D. Demonstrates a solid application into the decision-making process

In view of the above, the following research questions are addressed:

1. Can a statistical model approximate to the underlying LoS distribution?
2. Is it possible to use the same model for other hospitals or does each hospital needs its own customised model?
3. Which are the internal and external factors that affect LoS distribution and what it is the nature of this influence?
4. Can a statistical model be clinically and/or operationally meaningful?
5. What type of information can be derived from the model that can be incorporated in a decision-making process?
6. Can this model derived from routinely collected data be accurate in predicting LoS?

1.4. Country Context

This research was based on data from Mexico, the country from where the researcher is originated. Mexico or officially known as the United Mexican States is a representative democratic republic with a population close to 110,000,000 of habitants. Mexico is characterised as a country on demographic transition, with a complex epidemiological profile delineated by the growth of chronic diseases, accident rates and unhealthy lifestyle behaviours (Consejo Nacional de Poblacion, 2009).

According to the World Health Organisation (2007), this demographic transition has been defined by three events that have changed the profile of the Mexican population in the last decades: the increase of the life expectancy from 49.6 years in 1950´s to 75.7 years in 2010's, the decrease in the rate of birth from 6.8 children per women in 1970´s to 2.2 in 2010's and the decrease in mortality from 16 deaths per 1000 habitants in 1950´s compared to 4.4 deaths per 1000 habitants in 2010's[1].These three phenomena have boosted a process called "ageing of the population"; today the proportion of adult people is higher. With more adults living longer, chronic diseases (e.g. ischemic heart disease, cerebrovascular accidents, chronic obstructive pulmonary disease and diabetes) have replaced transmittable as major causes of death. In fact, the complications derived from diabetes are the leading cause of death in women and the second in men and affects more than five million of habitants of the country.

Risky behaviours and risk factors such as being overweight and obesity have increased in all groups of society, mainly in urban areas, affecting 51.8% of women between the ages of 12 and 49 and 5.5% of children under five. In 2002, 26.4% of the urban population aged 12-65 (14.3% rural) were smokers; approximately 32 million people aged 12-65 years consumed alcohol (Instituto de Salud Publica, 2006).

In addition, the marked historical structural inequities and income concentration that have led to inequities in access to basic services, opportunities, and social participation, such as: lower life expectancy in the indigenous communities(i.e. 51 years for women and 49 years for men), malnourished children (i.e. more than 1.2 millions of children), lack of access to water (i.e. more than five million people) and lack of access to healthcare or social security (i.e. around 51.4% of the population), continue to persist. The next figure highlights the most affected areas of the Mexican territory in terms of inequality and poverty (Secretaria de Salud, 2007).

[1]However the decline has been smaller among ethnic minorities and rural populations.

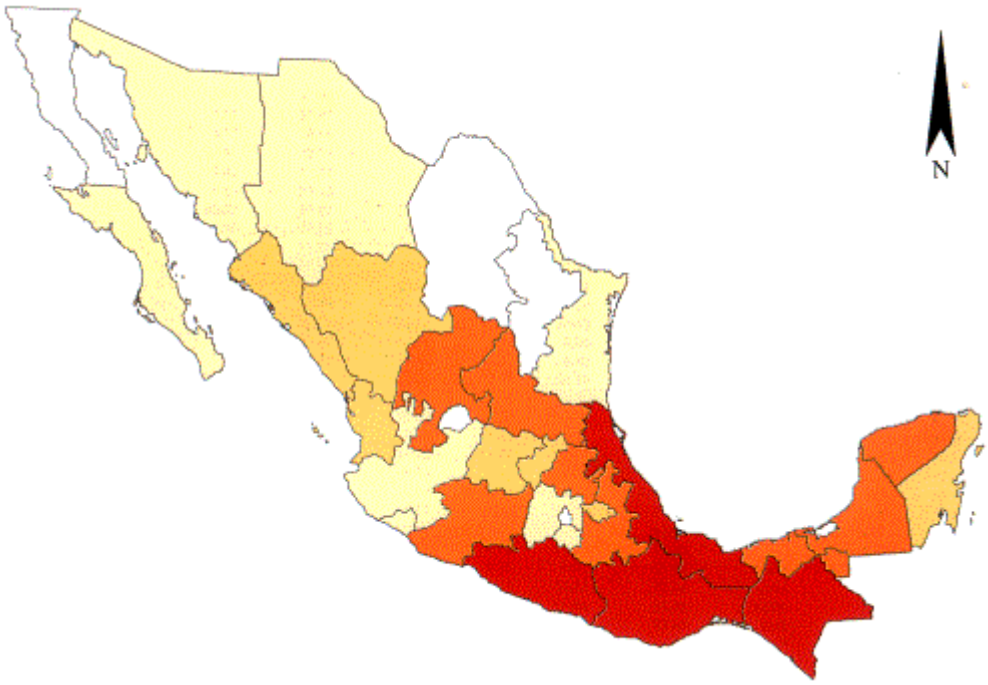

Figure 1.2: Level of marginality in Mexico: the poorer states located in the country's southern region have the highest concentration of rural and indigenous population groups, the highest disease prevalence and mortality rates for preventable causes.

Regarding to economic indicators, the total expenditure on health is 6.5% of the GDP, which is much less than other Latin-American countries with similar economies like Brazil or Argentina. Yet 95% of the private expenditure in health is made by families out of their pockets. In terms of workforce, the percentage of physicians and nurses (including midwifes) per 10,000 habitants is 28.9% and 39.8%, where the percentage of nurses and midwifes is a lot much lower than the regional average which is 61.5% (Pan American Health Organization, 2008).

With respect to the resource infrastructure, Mexico has 23269 healthcare units and 86.8% of these belong to the public sector. The country has 1.1 hospitals per 10,000 habitants and 0.79hospital beds per each 1000 habitants, which is lower than the WHO guideline of 1.0 bed per 1000 habitants.

The health system which has evolved since the second half of the last century is a complex body where the public resources finance two basic types of public institutions: social security institutions for employers and employees of the private and public formal sector; and health institutions for uninsured population: employees of the informal economy, self-employed and the unemployed. Figure 1.3shows the different population groups and the bodies who provide the services. Notice that in addition to the links described above, there are other connections among the different groups, represented by dotted lines. For example: social security subscribers may choose to be treated at Secretariat of Health because they prefer the care they

receive there or because they live far from the physician (or healthcare facility) assigned by the insurance (resulting in a cross-subsidy from the Secretariat of Health) or they have private insurance for major medical expenses (often provided as an additional work benefit). Furthermore, social security beneficiaries and the general public, at all economic levels, ultimately seek private medical care, paying for the services out-of-pocket.

Such a complex structure generates an enormous heterogeneity between the major services providers, making it even more challenging for the government to develop a common plan of action to improve the performance of the services. According to the government National Plan of Healthcare 2007-2010, the percentage of complications of vaginal births in the hospitals of the State Services and the Secretariat of Health (0.48%) is 2.6 times greater than the percentage in IMSS hospitals (0.18%). The percentage of appendicitis cases in State Services and Secretariat of Health hospitals is more than 6% against less than 2% in the hospitals of the IMSS.

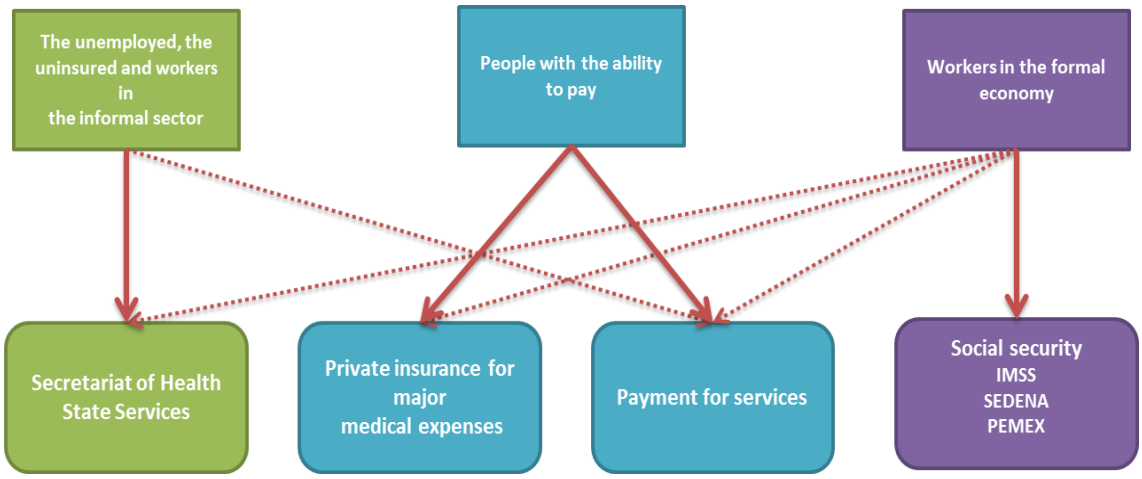

Figure 1.3: Health care system in Mexico (Adapted from Secretaria de Salud, 2007)

In addition to the technical quality problems there are efficiency problems: In the context of LoS, the average number of days of stay for the inguinal hernioplasty in IMSS hospitals (with over 120 beds) is 1.5 days against 2.6 days in the State Services hospitals and the Secretariat of Health hospitals. The average number of days of stay in a hospital for appendectomies with (60 beds or less) at the IMSS is 6.5 days against 3 days in hospitals in the State Services. This evidence suggests that (beyond patient characteristics and internal factors) LoS can vary from one hospital to other or from one healthcare provider to other.

1.4.1. Institutional Context

The process to select the hospitals to participate in this research was done by what in quantitative analysis is known as convenience sampling (Cramer, 2004). This sampling

technique selects a population (hospital) based on the opportunity. The data was simply ready, available and convenient for the research purposes of this thesis. The main drawback of this technique is that it is not possible to make generalisations about the total population of hospitals from the sample because it is not very representative of the population. In Chapter 7 the researcher deals with this issue.

The hospitals under study are the Instituto de Seguridad Social del Estado de México (ISSEMyM) Medical Centre and the Maximiliano Ruiz Castaneda (MRC) General Hospital. Both hospitals are located in State of Mexico, which is the most populous and densest state of the country.

The MRC general hospital is located in the heart of an urban area, it belongs to the Secretariat of Health and it is open to the general population which makes it the preferable option for people who cannot afford private medical services or who are not affiliated to another healthcare provider. It is a 148 bed second level hospital, which means it offers outpatient walk-in clinics and hospitalisations for basic medical specialties such as adult medicine, paediatrics, obstetrics and gynaecology, and general surgery. Hospitals that correspond to this level of care have operating rooms and equipment suitable for performing surgery of low and medium level of complexity.

On the other hand, the ISSEMyM Medical Centre (see Figure 1.4) belongs to the Social Security Institution for the State of Mexico and municipalities. The hospital treats employees of State of Mexico government and their families. ISSEMyM is a modern hospital with 330 beds and it is classified as a third level hospital which means it provides outpatient and hospitalisation services for a wider range of medical subspecialties such as gastroenterology, endocrinology, geriatrics, urology, angiology, haematology, nephrology, infectious diseases, oncology and neurology. In addition third level hospitals are designed to perform more complex surgeries for the specialties and subspecialties of the second level. Hospitals from this level also provide support services, diagnosis and therapy, which require a high technological degree of specialisation.

Figure 1.4: ISSEMyM Medical Centre (Mexico City), one of the hospitals under study

1.5. Thesis Structure

This thesis is structured in the following way:

Chapter 1 describes the general importance of the issue discussed in this thesis and sets the research objectives. Then it presents an overview of country and institutional context in which the research was conducted.

In Chapter 2 provides the literature review of the topic, which is divided in four parts: the arithmetic methods, statistical methods, case-mix based models and the multi-stage models.

Chapter 3 builds on Chapter 2 by identifying the primary characteristics of the model that will help to overcome the limitations and the areas of opportunity found in the literature review. Then the chapter outlines the formal methodology to build the model that satisfies the research objectives and answers the research questions and it finishes with a brief preliminary analysis of the data

Chapter 4 explores the role of finite mixture models in modelling LoS, aiming to approximate to the real distribution of LoS and to capture the variability of the data.

Chapter 5 explores the internal factors associated with LoS. This chapter is divided into main parts: the first part is devoted to the variable selection process and the second part explains how the previously selected variables are incorporated to the finite mixture model developed in Chapter 4.

Chapter 6 defines the internal factors associated with LoS-homogenous groups of patients, where patient attributes or variables will be used to predict the LoS category to which the patient is likely to belong.

Chapter 7 provides an extension of the models built in previous sections, in order to understand the environment on which the patient is treated and how this affects it length of stay.

Chapter 8 highlights the application of the models developed in Chapters 5, 6 and 7 to the decision-making process on healthcare.

Chapter 9 outlines the research limitations and further work on the topic. The thesis is finalised with a critical analysis of the main outcomes of this study and its novel contributions.

Finally, Appendix A contains additional models that where developed during an initial stage of this research when only data from MRC hospital was available. However, they are not described in the main body of this thesis because a broader perspective of the research problem suggested that the approach adopted was insufficient and limited. In addition, the analysis of new data confirmed that the model suggested originally were inappropriate.

2 LITERATURE REVIEW

The previous chapter defined the general objective of developing a statistical model to predict patient LoS in public hospitals in Mexico which approximates to the distribution of the LoS, recognises and addresses the heterogeneity population problem, supplies LoS predictions for individual patients and cohort of patients (within and between hospitals), and demonstrates a solid application into the decision-making process.

However in addition to the previous objective there is an underlying objective, that it is essential in any research, which is to identify the areas of opportunity and research gap from previous models and methods in order to provide a novel contribution to the field.

This chapter explores those relevant methods and research in the field of calculating and predicting patient LoS to date. Published papers included in this review were found mainly in journals for healthcare management, operational research applied to healthcare, medical decision analysis and in non-governmental organizations (NGO's) reports. For practicality, the relevant literature, after systematic research on the topic, was categorised into four groups as depicted in Figure. While the first group contains the arithmetic methods which are described in Section 2.1, the second group corresponds to the statistical methods which are explored in Section 2.2. The third group (described in Section 2.3) contains those methods which have a case-mix analysis base, and finally the fourth group corresponds to the multi-stage models, described in Section 2.4. All the methods are briefly explained with examples of their application in the context of the LoS and the objectives of this research. However more emphasis is placed on the third and fourth groups which are currently the most dominant methodologies utilised in the field.

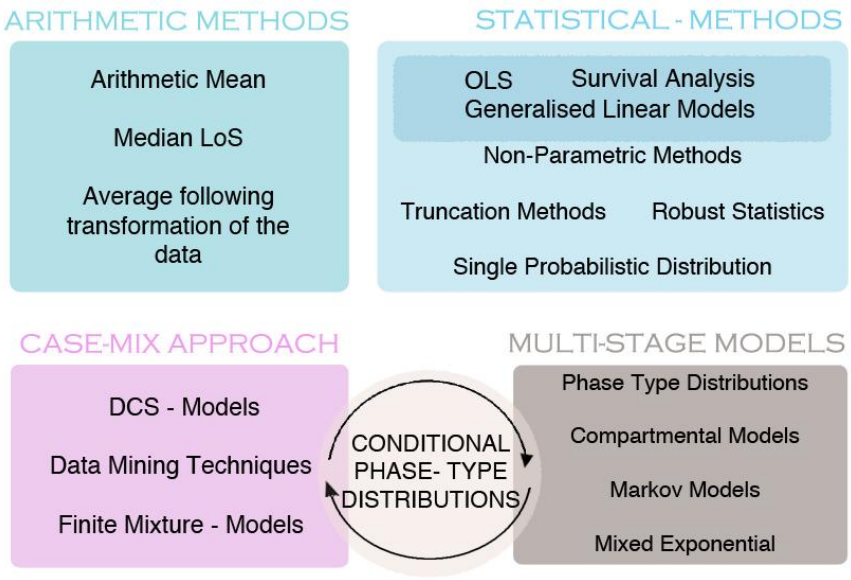

Figure 2.1: Literature review classification

2.1. Arithmetic Methods

Despite the considerable amount of research that has been conducted in relation to patient LoS, and which will be discussed in this chapter, arithmetic methods are still the most common methodology used on daily basis routine at hospitals (Millard, 1994), especially in developing countries as Mexico. Arithmetic methods usually compute the average length of stay (ALoS):

$$ALoS_1 = \frac{\sum_{i=1}^{N} patient_i LoS}{Number\ of\ discharges} \tag{2.1}$$

Or

$$ALoS_2 = \frac{Average\ number\ of\ occupied\ beds \times 365}{Number\ of\ discharges} \tag{2.2}$$

Other methods have been proposed to deal with the nature of the LoS distribution (i.e. asymmetric distribution, multimodal, with long and heavy tails to the right), which adapt better to the skewed nature of LoS (See Equation 2.3 and 2.4):

$$ALoS_3 \cong Median(Patient\ LoS) \tag{2.3}$$

And

$$ALoS_4 = Antilog\left(\frac{\sum_{i=1}^{N}\log(patient_i LoS)}{Number\ of\ discharges}\right) \quad (2.4)$$

Millard (1994) commented on the flaws of Equations (2.1) and (2.2): Equation (2.1) gives an overestimate of the LoS if there are clearly defined short-stay and long-stay group of patients (i.e. leading to high variability and therefore high error on estimated average, see Section 1.2), whereas Equation (2.2) can lead to wrong estimates if beds are blocked.

Ramakrishnan (2012) compares methods derived from Equations (2.1), (2.3) and (2.4) on Intensive Care Unit (ICU) LoS data from a 600-bed corporate hospital in India. The results indicated that the methods for the calculation of the arithmetic mean very often overestimate the ALoS, whereas the method of log-transformation of the data seemed to underestimate ALoS. Moreover, a full description of the drawbacks of these methods was provided in Section 1.2 which includes the inadequacy to represent the inherent variability of LoS and the fact that they ignore the heterogeneity population problem: arithmetic methods assume that all the patients (i.e. all the patients from the same cohort under study) will have identical LoS regardless of their personal characteristics.

2.2. Statistical Methods

Modelling LoS has a well-established statistical base where most of the traditional models have already been employed in one way or another to estimation of LoS. Among the statistical methods described in this section, Figure 2.1highlights a special subgroup of the statistical methods which includes the analysis of covariates. Covariates are defined in the context of LoS as the patient's characteristics and external factors which possibly predict LoS (i.e. medical condition, patient age, patient gender, pathological history, etc.). Within this subgroup are found linear regression, generalised linear regression and the proportional hazard model, which is a special case of survival models.

2.2.1. Methods Based on Truncation of Data

As mentioned before, the LoS distribution tends to have very long right tails and some very extreme scores, named atypical points or outliers. One common practice for handling these outliers is to truncate them (i.e. remove them), and this can be done either through visual inspection of the data using box plots or through other more formal procedures named truncation rules. These truncation rules determine upper and lower boundaries based on three measurements: a measure of position, a measure of scale and a factor (Ruffieux et al., 2000). The boundaries are set at a certain distance from the position, the distance being product of the

scale times the factor. Then a truncated LoS mean is calculated as the arithmetic mean of the data contained between the two boundaries (see Equation 2.5):

$$ALoS = \frac{\sum_{i=1}^{N} patient_i LoS \ (within\ boundaries)}{Number\ of\ discharges\ (t_1 < LoS < t_2)} \tag{2.5}$$

wheret_1 and t_2 are the lower and upper boundaries respectively.

Ruffieux et al. (2000) explored five different truncation rules and compare them with another method called the approximated quartile based truncated mean (AQTM) which takes into account the shape of the data distribution, on a database containing 4,758,347 LoS from five European countries. An interesting feature of AQTM is that it assumes that the sample distribution is the mixture of a regular distribution (e.g. Gamma, Lognormal or Weibull) and a "contaminating distribution" that describes the irregular, exceptional and unexpected stays. Table 2.1Table 2.2 display the notation and formulas to calculate the boundaries for each truncation rule and AQTM.

Truncation rule	Transformation	Position	Scale
Tiqr	Identity	Quartiles	Interquartile range
TLmr	Logarithmic	Median	Interquartile range
TLqr	Logarithmic	Quartiles	Interquartile range
TLmm	Logarithmic	Median	Median absolute deviation
Tlas	Logarithmic	Mean	Standard error

Table 2.1: Notations for truncation rules. The abbreviations stand for the type of transformation and the measures of position and scale.

Truncation rule	Lower boundary t_1	Upper boundary t_2
Tiqr	$t_1 = q_1 - 1.7[q_3 - q_1]$	$t_2 = q_1 + 1.7[q_3 - q_1]$
TLmr	$\ln(t_1) = \ln(q_2) - 1.5[\ln(q_3 - \ln(q_1)]$	$\ln(t_2) = \ln(q_2) + 1.5[\ln(q_3 - \ln(q_1)]$
TLqr	$\ln(t_1) = \ln(q_1) - 1.15[\ln(q_3 - \ln(q_1)]$	$\ln(t_2) = \ln(q_1) + 1.15[\ln(q_3 - \ln(q_1)]$
TLmm	$\ln(t_1) = \ln(q_2) - 3mad\{\ln(x)\}$	$\ln(t_2) = \ln(q_2) + 3mad\{\ln(x)\}$
Tlas	$\ln(t_1) = ave\{\ln(x)\} - 3sd\{\ln(x)\}$	$\ln(t_2) = ave\{\ln(x)\} + 3sd\{\ln(x)\}$
AQTM	$\ln(t_1) = \ln(q_2) - k_1[\ln(q_3 - \ln(q_1)]$	$\ln(t_2) = \ln(q_2) + k_2[\ln(q_3 - \ln(q_1)]$
Weibull	$k_1 = 3.26 - 1.36s + 0.20s^2$	$k_2 = 1.20$
Gamma	$k_1 = 1.718 + 0.167s - 0.153s^2$	$k_2 = 1.710 - 0.437s - 0.071s^2$
Lognormal	$k_1 = 1.72 - 0.55s$	$k_2 = 1.725$
	$s = \ln(q_3) - \ln(q_1)$	

Table 2.2: Five truncation rules and AQTM. Whereq_1, q_2 and q_3 denote the first, second and the third quartiles, $mad\{.\}$ is the median absolute deviation, SD is the standard deviation and x a particular LoS.

Later, Cots et al. (2003) compared four different truncation rules on data including LoS and costs from 35,262 patients. The results pointed the method referred as GM2 (which stands for 2 standard deviations from the geometric mean) as the most satisfactory method to detect outliers (see Table 2.3.)

Truncation rule	Upper boundary
GM2	$gm + 2 \cdot sd$
GM3	$gm + 3 \cdot sd$
IQ15	$q_3 + 1.5(q_3 - q_1)$
IQ20	$q_3 + 2(q_3 - q_1)$

Table 2.3: Four truncations rules used by Cots et al. (2003) on LoS data

2.2.2. Robust Statistics

Other methods to handle outliers are proposed by Marazzi et al. (1998) and Ramakrishnan (2012), where, unlike the truncated methods, the outliers are not eliminated but substituted by the lower and upper boundaries t_1 and t_2. Marazzi et al. (1998) defines $t_1 = 5^{th}\ percentile$ and $t_2 = 95^{th}\ percentile$. On the other hand, Ramakrishnan (2012) defines the boundaries using the same truncation rule as Tlas (see Table 2.2) with the difference that the observations

below the lower boundary t_1 and above the upper boundary t_2 are treated as t_1 and t_2 respectively. Therefore, ALoS is calculated using Equation 2.1 but with the outliers substituted by the new values.

2.2.3. Probability Distributions for Skewed Outcomes

Rather than a crude prediction of the LoS, it might be of interest to decision makers to try to define the underlying probabilistic distribution of LoS. The most common distributions used in the literature, given the skewed nature, are Gamma, Lognormal and Weibull.

Faddy (2009) compared Gamma and Lognormal distributions fitted to a dataset from two hospitals in Australia. The results gave a very poor fitting of the Gamma to the data and to a lesser extent the Lognormal models.

Marazzi et al. (1998) compared Gamma, Lognormal and Weibull distributions using two methods: the Cox test and the average trimmed absolute residual criterion (ATAR). Using3279 hospital samples from five European countries in three statistical years for 417 diagnosis related groups, the results indicated that the Lognormal model was the model which fits with the LoS distribution of the majority of the samples. The Gamma model was quite similar in terms of performance to the Weibull model; the latter is preferred because of its advantage of being computationally simpler than the Gamma model. On other hand, some countries found Lognormal distribution was the best fit, while others got better results with Weibull. This phenomenon might reflect a difference in hospital practices among the studied regions. However almost 36% of the samples could not be associated with any model, some of the explanations for such problem might be: early peaks (LoS = 1 day) combined with a strong concentration for a few consecutive days; multimodality of the distribution and large samples (more than 1500 observations).

2.2.4. Survival Analysis

Any of the distributions mentioned in 2.2.3 for continuous non-negative variables can serve as a survival distribution, which is the complement of the cumulative density function, and gives the probability that an event of interest has not occurred by duration *t* (see Equation 2.6). In terms of LoS, the survival analysis computes the probabilities that the patient will still be retained in hospital after *t* days.

$$S(t) = 1 - F(t) = \int_t^{\infty} f(t)dt \tag{2.6}$$

For example, length of hospital stay after Staphylococcus aureus bacteraemia (i.e. presence of bacteria in the bloodstream) for patients who did not die was analysed using a semi-parametric Weibull survival analysis model in Cosgrove et al. (2005).

However, the former type of models assumes a homogeneous population, where the lifetimes of all objects (or subjects) are governed by the same survival function. The next types of models introduce the presence of covariates that may affect the survival time. The most frequently used model for adjusting survival function for the effects of covariates is the Cox proportional hazard model (Cox, 1972).

Newburger et al. (2003) used proportional hazard model to test the hypothesis that longer postoperative LoS after infant heart surgery is associated with worse later cognitive function. In contrast, Strate and Syngal (2003) used a proportional hazard regression to determine if time to colonoscopy impacts LoS in patients admitted for acute lower intestinal bleeding.

One of the most important features of survival models is that they can deal with censored data. Censoring occurs when the observation period finishes and some individuals have not presented the event of interest. For example, some individuals may die before the end of a clinical trial, or may drop out of a study for various reasons (i.e. transfer to hospital, voluntary discharge, etc.) other than death prior to its termination.

2.2.5. Linear Regression

Linear regression approaches such as Ordinary least square (OLS) are by far the most widely used modelling method. These approaches seek to predict an outcome variable (i.e. dependent variable) from several predictors (named covariates, explanatory variables or independent variables).

To mention some studies, Galski et al. (1993) used multiple regression to predict length of stay in rehabilitation for stroke patients. Knaus et al. (1993) developed a model to predict LoS in the ICU using the same technique. Classen et al. (1997) used linear regression to assess the effect on length of stay and cost of hospitalisations of patients who experienced an adverse drug event during hospitalisation. Whereas some other studies extended linear regression models to account for random effects, known as multilevel models or hierarchical models (Martin and Smith, 1996;Leung et al., 1998;Frick et al., 1999; Carey, 2002; Urbach and Austin, 2005 and Jong et al., 2006).

OLS is based on a number of assumptions. However the normality of the errors is the most relevant for the study of LoS. The reason is that the normality of the errors is seriously affected by the presence of extreme values and skewness, which are embedded characteristics of LoS.

The most common strategy to make non-normal data resemble normal data is by using a transformation. A considerable number of researchers have applied OLS following the transformation of the data. Fleischmann et al. (2003), for example, evaluated the effect of cardiac and non-cardiac complications on log-transformed LoS for patients undergoing non-cardiac surgery. Chertow et al. (2005) also used linear regression to evaluate the effect of acute kidney injury in log-transformed LoS, and Chen et al. (2005) used it to evaluate the impact of nosocomial infection on log-transformed LoS in ICU.

The use of a logarithmic or other transformation has been criticised due to the fact that transformed LoS has little meaning for decision-making process (Faddy et al., 2009). Moreover, models with logged dependent variable results are about geometric means, not arithmetic means, which can lead to biased estimates of the effects of independent variable (Manning, 1998). Additionally, there is the problem of retransformation. Although the common practice is to interpret the response to a particular variable as being the exponential of the coefficient of that variable in the model; this is only valid when the error term does not break the assumption of homoscedasticity (i.e. constant variance). For this reason, retransformation is usually named as homoscedastic retransformation. Manning and Mullahy (2001) found that when the log-scale error term was heteroscedastic the OLS estimates after homoscedastic retransformation can be appreciably biased.

However Manning and Mullahy (2001) suggests that the homoscedastic retransformation works better with heavy-tailed distributions (heavy-tailed on the log-scale) than any other of the alternatives they explored.

2.2.6. Generalised Linear Models (GLMs)

GLMs (Nelder and Wedderburn, 1972) are regression models where the dependent variable is specified to be distributed according to one of the members of the exponential family. Moreover the linear predictors (independent variables) are connected to the mean of such distribution by a link function. In fact, GLMs are usually presented as an alternative to transformation of the data for healthcare outcomes. Manning and Mullahy (2001); Manning et al. (2002) and Basu et al. (2004) suggest that the GLM models are more appropriate than an OLS on a logged dependent variable because they avoid the difficulty of the retransformation of the response to account for the heteroscedasticity of the error term (see Section 2.2.3).

Iglesias et al. (2006) compared models assuming different distribution functions (Gaussian, Gamma, inverse Gaussian) and link functions (identity, log) to compare the effect of alternating pressure mattresses with alternating pressure overlays on LoS for patients with pressure ulcer. In contrast, Sayers et al. (2007) used a Gamma GLM to measure the effect of psychiatric

comorbidities on length of stay. Similarly, Graves et al. (2007) used a Gamma GLM with log link function to estimate the independent effect of a single lower respiratory tract infection, urinary tract infection, or other healthcare acquired infection on patient LoS. Alternatively, Rauner et al. (2003) used a Quasi-Poisson model to evaluate the effect of day and month of admission, as well as different types of admission and discharge, on patient LoS at Austrian hospitals. The Quasi-Poisson model is an extension of the Poisson GLM with log link function to account specifically for the high variance of LoS.

2.2.7. Non-parametric Methods

Most of the methods described before make a hypothesis about the distribution form of the data, and identifying such form can be a challenging and uncertain task. One possible way to overcome this is to use non-parametric procedures, which eliminate the need to specify the form of the distribution in advance. Non-parametric methods, commonly named as distribution-free methods or parameter-free methods do not rely on assumptions that the data are drawn from a given probability distribution.

One of the most common non-parametric techniques is bootstrapping (Efron, 1979). In this technique, independent random sampling with replacement from the original dataset is carried out, firstly to create *n* simulated samples of the same sample size as the original dataset. Then, the statistic of interest (i.e. average LoS) is recalculated for each new sample (Dodd et al., 2006). The distribution of these recalculated statistics sample will tend to be the underlying true cumulative distribution function of the statistic, as a means to calculate confidence intervals for the statistic of interest. Ramakrishnan (2012) used the bootstrap method to estimate average LoS and confidence intervals on ICU LoS data from a 600-hundred-bed corporate hospital in India. However, bootstrapping should be used to estimate the sample distribution of the statistic of interest rather than the statistic itself, which will be biased if the original statistic is biased (Efron and Gong, 1983)

On the other hand, in the field of survival analysis the mostly used non-parametric counterpart is Kaplan-Meier survival curves (Kaplan and Meier, 1958), which allows the estimation of survival times on censored data. The Kaplan-Meier estimates also easily facilitate the computation of central tendency statistics (i.e. mean LoS). This method has been used by Wilson et al. (1999), who used Kaplan-Meier curves to estimate the LoS for patients who received different preoperative optimisation methods of oxygen delivery before major elective surgery. A similar approach was adopted by Forster et al. (2012), who applied the same method to describe the LoS of patient with hospital-acquired infection (i.e. Clostridium difficile).

2.3. Case-mix Analysis Base

The case-mix is defined as the clinically meaningful grouping that broadly describes the types of patients treated by a hospital or a healthcare service. Case-mix is also used as a generic term to describe scientifically developed grouping mechanisms used to categorise patient care episodes in order to facilitate effective planning and management of healthcare (Heavens, 1999). In this context the models described in this section correspond to those which have a component of "patient classification" according to internal or external factors.

2.3.1. Mixture Models

Quantin et al. (1999) suggest that the reason why none of the distributions under study seemed to fit satisfactorily in a wide variety of samples as Marazzi et al. (1998) pointed out, was because disparities in patient care and medical practice for a given DRG might lead to the formation of subgroups that systematically differ with respect to LoS, and whose proportions may differ from one hospital to another. They suggest that the observed distribution of LoS within the same DRG may in fact represent a mixture of several different distributions. This type of models is commonly referred to as finite mixture models, where a continuous variable in a large sample consists of two or more clusters of observations (components) with different means and perhaps different standard deviations within each sample. In other words, the observed continuous variable is a mixture or sum of two or more distributions with different parameters (Maclachlan and Krishnan, 1997) and each cluster provides a local approximation to some part of the true distribution (Deb et al., 2011).

In fact, the idea of finite mixture models goes back as early as the work of Pearson (1894), where he stated that: *"In the case of certain biological, sociological and economic measurements there is, however, a well-marked deviation from a normal shape, and it becomes important to determine the direction and amount of such deviation. The asymmetry may arise from the fact that the units grouped together in the measured material are not really homogeneous. It may happen that we have a mixture of 2, 3, ... s homogeneous groups, each of which deviates about its own mean symmetrically and in a manner represented with sufficient accuracy by the normal curve. Thus an abnormal frequency-curve may be really built up of normal curves having parallel but not necessarily coincident axes and different parameters".*

Pearson was also the first to introduce a method of moments approach to the estimation of a finite mixture of distributions, which implied complex calculations (i.e. it involved the solution of a ninth degree polynomial). Later, Rao (1948) suggested the maximum likelihood algorithm as an estimation method, and this has become one of the most popular methods of estimation in the recent years.

Another very popular method of estimation, which can be thought as an extension of the maximum likelihood algorithm, is the expectation-maximization (EM) algorithm formulated by Dempster et al. (1977). Finally, the use of Bayesian approach to estimate finite mixture models has become very familiar with the development of Markov chain Monte Carlo (MCMC) methods (Diebolt and Robert, 1994).

One of the main advantages of the finite mixture models is that the problem of choosing the right number of clusters and an appropriate clustering method can be reduced to a statistical model choice problem (Fraley and Raftery, 2002) where the task of choosing the optimal number of components or comparing among different models can be performed via the usual methods (see Section 4.1.1).

In the field of healthcare data, Mihaylova et al. (2011) suggested that mixture models often perform better than model alternatives based on single distributions for total resource use. Abbi et al. (2008) found that a six-component Gaussian mixture model was able to model several types of LoS within stroke patients. On the other hand, Atienza et al. (2008) explored a mixture of different families through a mixture of the union of Gamma, Weibull and Lognormal families to model LoS within several DRG's.

However most of the attention on mixture models is concentrated on those models that include analysis of covariates. These models are defined as finite mixtures of generalised linear models, where a set of covariates (independent variables) is related to the mean of each component of the mixture by means of generalised linear models. For example, Xiao et al. (1999) used a Poisson mixture regression model to analyse the potential factors that might influence LoS of certain obstetrical DRG's. Lee et al. (2001) fitted a Gamma mixture regression model to identify the factors associated with maternity LoS within obstetrical DRG's. Singh and Ladusingh (2010) found that a negative binomial regression model provided reasonable fit to nationwide LoS data in India. Yau et al. (2003) extended the application of a Gaussian mixture regression model on neonatal LoS data to account for between hospital variations through random effects in the linear predictors.

2.3.2. Data Mining Techniques

Data mining is a relative new methodology and technology from the last two decades. It aims to identify valid, novel, potentially useful, and understandable correlations and patterns in data (Trybula, 1997) According to Koh and Tan (2011), data mining in healthcare is becoming increasingly popular, if not increasingly essential, as a viable tool to address the important heterogeneous patient population problem. One of the main branches of data mining techniques is in the area of predictive models, which aim to describe one or more of the variables in

relation to all the others. This is done by looking for rules of classification and prediction based on the data.

In the context of predicting LoS, there are broadly two types of predictive models: regression-type models and the classification-type models. In regression-type models, such as regression trees, LoS is analysed as continuous variable. However, unlike linear regression models, they do not hold the implicit assumption that the underlying relationships between the predictor variables and the dependent variable are linear or monotonic. Ridley et al. (1998) applied classification and regression trees (CART) with the aim of classifying ICU patients from three hospitals into groups so that the variation in LoS within each group was minimised. Stineman et al. (1998) used CART to establish subgroups of patients, with a similar impairment condition, expected to have similar LoS in rehabilitation care hospitals. These subgroups were later defined as FIM-FRG's, (Functional Independence Measure- Function Related Groups) a version of DRG's for rehabilitation care. Harper (2002) incorporated a CART algorithm in a simulation tool for the planning and management of hospital resources. CART, specifically is used in the construction of homogenous LoS patient groupings for bed management. More recently, Saltzman et al. (2011) evaluated the effect of a new bedside risk score on different healthcare outcomes (including LoS) for patients with acute upper gastrointestinal bleeding.

Examples of more advanced models are: Garg et al. (2010), who proposed a novel decision tree algorithm where phase-type distributions (see Section 2.4.1) are fitted at every step of the construction of the tree. This model, called phase-type survival trees, was used for fitting LoS stroke data. Later Garg et al. (2011) compared this model to a new extension called Gaussian mixture survival tree model (i.e. Gaussian mixture distributions are fitted at each step of the construction of the tree) using the same stroke dataset. The Gaussian mixture survival tree was by far superior to the phase-type tree, not just in terms of better fit but in terms of providing a much simpler model. In the same paper, a survival tree that includes both type of distributions (i.e. phase-type and Gaussian mixture) within the same tree was developed proving to be superior (in terms of maximum log-likelihood) to the other individuals trees.

On the other hand, the dependent variable in the classification-type models is a discretised version of LoS whereby the continuous variable is split in different intervals according to a certain criteria. Each interval in this sense corresponds to a discrete category, for example short LoS, medium LoS or long LoS. The aim of these models is to classify patients into these categories according to their characteristics. Liu et al. (2004) used data from two datasets: clinical data and stroke data where the LoS variable was discretised into three and six groups respectively. They explore two algorithms: Naive Bayesian Classifier and C4.5 tree. A new

version of C4.5, called R-C4.5, introduced by Yao et al. (2005) is applied to predict LoS, which is categorised into three groups.

Artificial neural networks have been used to predict LoS in a Canadian ICU after cardiac surgery (Tu and Guerriere, 1992). Moreover they have been used to predict LoS for psychiatric patients who are involuntarily admitted to American states hospitals (Lowell and Davis, 1994) and for patients with acute pancreatitis (Pofahl et al., 1998).

Ramon et al. (2007) explored two Bayesian networks algorithms: Naive Bayes classifier and Tree Augmented Naive Bayes Network (TAN). In addition to Decision tree learning (DT) and First order random forest (FORF) to predict hospital stays of more than 3days in ICU.

Data mining techniques also play an important role in the Discrete Conditional Survival Models (DCS), which are descendants of the Discrete Conditional Phase-Type models explained in Section 2.4.1 of this chapter. These models consist of two components: (1) a conditional component, which comprises a structure (in the form of a data mining model) that captures the nature of the data by representing the various inter-relationships between variables, and categorise the observations into a number of discrete classes and (2) a process component, which represents the skewed distribution of each discrete class by some form of survival distribution.

The latter approach has been used by Cairns and Marshall (2009) who fitted a DCS model to ambulance response times using multinomial logistic regression as the conditional component and different distribution forms, including Log-logistic and Lognormal as the process component.

The range of options for a data mining algorithm seems limitless, with more sophisticated and new algorithms emerging in the field constantly. Therefore, it may become a real challenge to choose an adequate method, whose success relies on the particular nature of the data. For example, Lim et al. (2000) run an extensive study in which twenty-two decision trees, nine statistical methods and two graphical methods were compared among thirty-two datasets in terms of classification accuracy (i.e. mean error rate), training time, and (in the case of trees) number of leaves. The results show that the mean error rates of many algorithms are sufficiently similar and their differences are statistically insignificant. Therefore they suggested taking into account other criteria such as interpretability of the data mining method.

2.4. Multi-stage Models (MSM)

Multi-stage models are models for “processes”, for example describing a life history of an individual (Hougaard, 1999). One of the simplest predecessor multi-stage models was

developed by Pendergast and Vogel (1988) where clinically meaningful phases of hospital care were defined and basic probability theory was used to estimate the transitions probabilities from one phase to another. Then the number of patients and associated LoS of each phase were estimated and translated to bed requirements.

Statistically speaking, these models are appropriate for a continuous time stochastic process which allows individuals to move among a finite number of states. A change of state is called a transition, or an event, and states can be transient or absorbing, if no transitions can emerge from the state (Meira-Machado et al., 2009). Recently there have been a number of multi-stage models which directly link LoS with the process of care of the patient, where the stages (or states) could be understood as representing severity of illness, patient recovery process or patient pathways at hospital.

2.4.1. Phase-type Distributions (Ph)

Phase-type distributions (Neuts, 1994) describe the time to absorption of a finite Markov chain in continuous time when there is a single absorbing state and the stochastic process starts in a transient state (Faddy, 1994). They are based on the assumptions that the $1, \dots, n$ states are all transient, so absorption into state $n+1$ from any initial state is certain (Marshall et al., 2005b). The corresponding density function of a continuous nonnegative random variable T is given by Equation 2.7:

$$f(t) = -\boldsymbol{p} \exp(\boldsymbol{Q}t)\, \boldsymbol{Q}\boldsymbol{e} \tag{2.7}$$

Where $\boldsymbol{p}$ is the initial state probability, $\boldsymbol{Q}$is the sub-matrix of transition rates restricted to the transient states and $\boldsymbol{e}$ is a $n \times 1$ vectors of ones. The Ph distribution is said to have a representation ($\boldsymbol{p}$, $\boldsymbol{Q}$) of order n.

Fackrell (2009) fitted a general phase-type model of order 6 (i.e. six transient states) to truncated LoS data from a hospital in Australia, and compared it with other types of Ph distributions (e.g. exponential, hyper exponential, generalised Erlang and Coxian)

Most of the literature on the field focuses on a subclass of phase-type distribution known as Coxian phase-type distributions (see Figure 2.2). In this type of distribution, the transient states are ordered with the process starting in the first and then developing in either sequential transitions through these phases or transitions out into the absorbing state (Marshall and McClean, 2004).

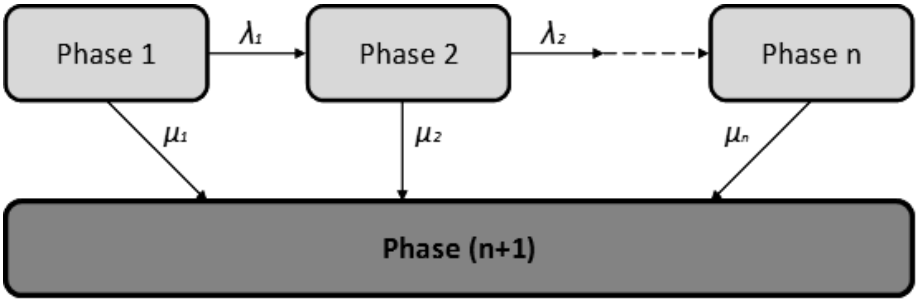

Figure 2.2: Coxian Phase-type distribution. (Adapted from Marshall and McClean, 2004)

For example, a Coxian phase type-distribution of order 3 has a representation summarised by Equations 2.8 and 2.9:

$$p = (1\ 0\ 0) \tag{2.8}$$

$$\boldsymbol{Q} = \begin{pmatrix} -(\lambda_1 + \mu_1) & \lambda_1 & 0 \\ 0 & -(\lambda_2 + \mu_2) & \lambda_2 \\ 0 & 0 & -\mu_3 \end{pmatrix} \tag{2.9}$$

Faddy and McClean (1999) fitted a Coxian phase-type distribution to male geriatric patient data, and incorporated covariates as age at admission and year of admission. McClean et al. (2005) also used the Coxian phase-type model to classify patients into categories according to how long they might spend in hospital (LoS) and what phases of care they are likely to pass through. Faddy et al. (2009) compared the Coxian phase-type with Gamma and Lognormal distributions fitted to a dataset from two hospitals in Australia and found that the phase-type was slightly superior to the other two distributions. They argued that the reason for the better fit of models using phase-type distributions is that these distributions are better able to accommodate extreme values, because of the hidden Markov process which differentially characterises short and long stays in hospital.

On the other hand, Xie et al. (2005) proposed a more complex representation of a phase-type distribution to model the LoS of geriatric patients in residential and nursing home care, where transitions between both adjacent and non-adjacent states are allowed. However when the model was fitted to data on funded admissions to residential care and nursing home care in London, a more parsimonious model in the form of Coxian phase-type was needed.

Tang et al. (2012) extended the work done by Faddy et al. (2009) to incorporate covariates to the phase-type structure and applied a Bayesian method to select the number of phases. The methodology is illustrated on LoS data from patients with acute myocardial infarction (AMI).McClean et al. (2010) extended the basic model to the mixed Coxian phase-type with

multiple absorbing states. They used it to model length of stay on stroke patients with three absorbing states or destinations (e.g. home, death and other destinations).

Moreover, phase-type distributions have been incorporated as an element of more complex models: Gorunescu et al. (2002) developed a queueing model where the servers are hospitals beds and the service time (i.e. time the bed is occupied) can be modelled using a phase-type distribution. In addition, various scenarios were tested using different admission rates, length of stays and bed allocations to measure the influence on bed occupancy, bed emptiness and rejection in departments of geriatric medicine. Later, Marshall et al. (2002) developed a new approach named conditional phase-type (C-Ph) distribution which uses Coxian phase-type distributions conditioned on a Bayesian network. The conditional phase-type model is defined as consisting of Causal Nodes $\boldsymbol{C} = \{C_1, \dots, C_m\}$ belonging to the causal network, and Process Nodes $\boldsymbol{Ph} = \{Ph_1, \dots, Ph_n\}$ representing the Coxian phase-type distribution (Marshall et al., 2005a). Marshall and McClean (2003) applied these models to geriatric LoS data. The later model was expanded to Discrete Conditional Phase-type models (DC-Ph) which consists of two components. Firstly there is the conditional component, which comprises a structure that captures the nature of the data by representing the various inter-relationships between variables, and thus can categorise the observations into a number of discrete classes. Secondly, there is a process component, which represents the skewed survival distribution of each discrete class by a Coxian phase-type distribution.

Marshall et al. (2007) fitted a DC-Ph model to LoS data at an A&E department where the conditional component was a Naive Bayes classifier.

Similarly, Harper et al. (2011) introduced a DC-Ph model where the conditional component is a classification tree to predict complications during childbirth and the duration of childbirth on the labour ward is then modelled by Coxian phase distribution.

One of the drawbacks of the PH distributions is the lack of ease in fitting the parameters of the distribution. As a result they are not used extensively as there is no standard statistical software available to carry out the estimation (Marshall and Zenga, 2010). Moreover, Marshall and Zenga (2009) confirmed that in the case of the Coxian phase-type distribution, the choice of initial values heavily influences the fitted results and that it was possible for different estimation procedures to obtain equivalent solutions with different parameter estimates. Marshall and Zenga (2010) proposed instead the use of the time-consuming Quasi-Newton algorithm for the estimation.

In addition to the problems of estimation, there is the over-parameterisation problem. In the three comparative studies mentioned in this section: Fackrell (2009), Faddy et al. (2009) and Marshall and Zenga (2010), the general PH and Coxian phase-type distributed were superior to

the other distributions in terms of maximum log-likelihood, however the difference on log-likelihoods, in comparison with other more parsimonious models, is very small. For example, in Faddy et al. (2009) the best fit was a Coxian phase-type distribution order six with a log-likelihood value of -4903.53 and the second best fit was a Lognormal distribution with a log-likelihood value of -4922.81[2]. However, the phase-type distribution has 11 parameters to estimate and the Lognormal distributions have just two parameters. In Marshall and Zenga (2010) the difference between the log-likelihood of a Coxian phase-type distribution (order 4) and a Weibull is just 11.38 units, however the former model contains seven parameters and the latter just 2.

Finally, the phase-type distributions as other multi-stage models aim to represent the patient process of care. However, the resulting (optimal) number of states often does not have a clinical or operational meaning (although they are statistically significant). Perhaps one of the most representative examples of this problem is found in Fackrell (2009) work. They found that a Coxian phase-type distribution of order 25 (25 states and 49 parameters to estimate)was the best fit for Australian LoS data. Later this model was dismissed for a simpler model, a general PH with six transient states[3] which also provided a superior fit. However, the problem is that the fitted data was truncated at 30 days, eliminating all the outliers and reducing considerably its variability (the coefficient of variation is 1.17, which is very close to 1). LoS data of such nature should not require that level of model complexity, where 6 (or 25) states become obviously meaningless from any clinical or operational point of view.

Summarising, phase-type distributions are models that are often over parameterised and their estimation and interpretability is complex. In this case, the final user is the one who can assess the best compromise model between fit and complexity.

2.4.2. Mixed Exponential and Compartmental Models

Another type of phase-type distribution (although it is rarely recognised as such) is the hyper exponential distribution (Fackrell, 2009), which is commonly known as mixed exponential: it has probability density of the form

$$f(t) = \sum_{s=1}^{S} p_s \lambda_s e^{-\lambda_s t} \tag{2.10}$$

where $p_s > 0$ and $\sum_{s=1}^{S} p_s = 1$

[2] In addition, Faddy et al. (2009) reported the BIC (Bayesian Information Criterion) which assesses the goodness of fit and penalise the number of parameters. Nevertheless phase-type distribution was a better fit than log-normal, although the difference was very small

[3] A standard general PH distribution of order p requires $p^2 + p - 1$ parameters

Millard (1988) demonstrated that Equation (2.6) with $s = 2$, a two-term mixed exponential, reflected the time elapsed since admission for two different types of patients requiring acute/rehabilitative and long stay care on bed occupancy midnight census data, where each group is represented by a single exponential with mean LoS equals to p_s/λ_s. McClean and Millard (1993) fitted a two-term mixed exponential to geriatric data in order to examine the pattern of LoS in the ward of admission until death/discharge or transfer.

Similar to the compartmental flow models used in pharmacokinetics, Millard and Tooting (1992) suggested that the best fit mixed exponential (two or three terms) model has a corresponding flow model with two or three compartments. In other words, a system with *n* compartments will have *n* linear differential equations, which solution can be written as *n*-term mixed exponential distribution. The change from continuous time (mixed exponential) to discrete time (compartmental models) makes the equations easier to interpret, especially for clinicians (Marshall et al., 2005b).

In this type of model, developed by Harrison and Millard (1991), the patients flow into the department to receive different type of care (i.e. acute care , rehabilitation or long-term care), giving a better understanding of the movement of patients in and out of hospital.

Harrison (1994) and McClean and Millard (1998) extended the models on geriatric data to three terms mixed exponential. Harrison (2001) discussed different compartmental models defined as cascading flow models and separate flow model in order to mimic the admission/discharge process in geriatric departments. Vasilakis and Marshall (2005) compared Coxian phase distribution with compartmental models for length of stay in stroke patients. Their results indicated that the expected length of stay in each phase/compartment was very similar.

More recently Harrison and Escobar (2010) used compartmental models to model the paths through community hospitals of cohorts of patients with specific conditions such as diagnosis, severity of illness and mortality risk.

Unlike other models presented in this review, mixed exponential and compartmental models have been widely used for measuring bed occupancy and modelling bed occupancy patterns, what has made it quite acceptable among the research community in the field.

2.4.3. Markov Models

Irvine et al. (1994), Taylor et al. (1997) and McClean et al. (1998) extended the model of Harrison and Millard (1991)to its continuous stochastic analogue in the form of a Markov model, where the distribution of time en each state (phase) follows an exponential distribution.

In particular, Irvine et al. (1994) designed a two-stage continuous time Markov model that describes the movement of patients through geriatric hospitals. Unlike the compartmental models where the admission rate is constant (Harrison, 2001), this Markov model assumes that the admission of patients to geriatric hospitals may be described by a Poisson process.

In contrast, Taylor et al. (1997) developed a four-state model consisting of two states representing the hospital (acute and long stay care), one state representing the community and another absorbing state representing death. They described it as a continuous-time Markov model where admission is considered as a Poisson process. The model was later extended by Taylor et al. (2000)to a six-compartment model with three hospital compartments, two community compartments and an absorbing state.

Alternatively, McClean et al. (1998) designed a Markov reward model which attached different costs to short and long stay patients, allowing to derive expressions for the probability distribution of the total future decumulate cost of a group of geriatric patients and estimate their average daily costs throughout future time. This model was extended later by McClean and Millard (2006) who added a phase-type distribution in order to model the time patients spend in hospital and community and determine the full system costs of new admissions and current patients.

2.5. Critical Review

Mihaylova et al. (2011) compared different methods for the estimation of healthcare outcomes using several guidelines such as the ability of the model or method to: account for skewness and heavy tails, including covariates, handle small samples and the ease of implementation. These same guidelines were followed for the evaluation of the literature discussed here, as it is believed that such guidelines should be among the basic requirements for the model that will be developed in this research. In addition, the comparative study performed in this section, was complemented by including whether the methods or models hold a clinical or operational meaning, by the ability to model probabilistic relationships and whether the analytical approach have a "grouping patients" component. Figure 2.3displays the summary of the criteria applied to the analytical approaches discussed in this chapter.

Ability to handle the features of the data: This criterion measures the capacity of the model or method to handle the embedded characteristics of LoS distribution such as skewness and heavy tails. Since this was the primary concern for many of the researchers in the area, most of the methods and models respond relatively well on handling the nature of LoS. The arithmetic methods and linear regression have a limited ability, unless a transformation of the data is

applied. On the other hand, truncation methods and robust statistics seem to cope well with skewness but it may have limited applicability for heavy-tailed data. Moreover, the applicability of the probability distributions depends on the chosen distribution: Lognormal, Gamma and Weibull works well. Finally, GLMS, proportional hazards models, the methods based on a case-mix approach and multi-stage models seems to be the most appropriate methods on acknowledge the skewness and heavy tails.

Analysis of covariates: This criterion judges the ability to account for covariates. Although the main objective of most of the methods and models described in the previous section was to predict LoS, some of them focused as well on understanding the factors that influence the LoS. Probability distributions may account for the analysis of covariates if these are incorporated in a way comparable to GLMs. Multi-stage models like phase-type distributions and Markov model can be extended to account for explanatory variables however this highly complicates the estimation and interpretability of the models. Linear regression models, GLMS and models based on the case-mix approach are especially designed to allow adjustments for covariates.

Clinical or operational meaning: The case-mix approaches and multi-stage models have high application to understand the hospital dynamics in terms of patient flow, whether this is clinical or operational, through the understanding of the patient population and its interaction with the system, i.e. hospital or community. Moreover they allow the grouping of patients into homogenous clusters, which helps to simplify the system as well as improve the understanding of the diverse patient population (Harper, 2005). Furthermore, they provide a solution to tackle the heterogeneity population problem (mentioned in Section 1.2) by grouping the population into a set of comprehensible and homogenous groups (Gorunescu et al., 2010).

On the other hand, some of these models have already been implemented in more advanced simulation models like Gorunescu et al. (2002) or the so-called bed management systems like BOMPS [4] (McClean and Millard, 1995) and PROMPT [5] (Harper, 2002), which have been designed to improve the quality of decision-making in bed management on both a short and long-term basis.

Finally, survival analysis brings valuable information about patient behaviour (in terms of LoS) at admission but more importantly allows predictions at any moment during the hospital stay.

[4] The acronym BOMPS stands for Bed-Occupancy, Management, and Planning Software
[5] The acronym PROMPT stands for Patient Resource Operational Management Planning Tool

Analytical approach	Handling features of data	Analysis of covariates	Clinical/ Operational meaning	Probabilistic relationships	Works with small samples	Ease of estimation
Arithmetic methods	◔	○	○	○	●	●
Truncation rules	◑	○	○	○	◔	●
Robust statistics	◑	○	○	○	●	●
Probability distributions	◕	◕	◑	●	◔	●
Survival analysis	●	◕	●	●	◑	◕
Linear regression	◔	●	◔	○	◔	●
Generalised linear models	●	●	◑	◕	◔	◕
Non-parametric methods	◕	○	◑	◑	◑	◕
Mixture models	●	●	●	●	◑	◑
Data mining techniques	◕	●	●	◑	◑	◑
Phase-type distributions	●	◑	●	●	○	○
Mixed exponential/com	●	◔	●	◑	●	◑
Markov models	●	◑	●	●	○	○

○ Low applicability ◔ Low-medium applicability ◑ Medium applicability ◕ Medium-high applicability

● High applicability

Figure 2.3: Summary of some characteristics of the reviewed methods for predicting and calculating hospital LoS (Adapted from Mihaylova et al., 2011)

Modelling probabilistic relationships: It is more useful for understanding the operational function of a healthcare system, if the model to develop in this thesis ascertains the probability or likelihood with which specific values of the variable LoS will occur, rather than a crude estimation such as arithmetic methods or linear regression computed LoS. These models are usually defined as non-deterministic of static nature such as probability distributions, survival analysis, mixture of models, phase-type distributions and mixed exponential; and non-deterministic of dynamic nature such as Markov models.

Sample size implications: The information about how well the model and methods described in this review performs on small datasets is very limited and unclear. In Lim et al. (2000) some data mining techniques performed very well in dataset containing few hundreds of observations whereas other performed very bad. On the other hand, to fit mixed exponential models only midnight hospital census data from one week is enough to obtain valid estimates according to Harrison (2001).More recently, Marshall and Zenga (2010) proved that for phase-type distributions the larger the samples the more accurate the parameters were estimated.

Ease of estimation: The arithmetic and statistical methods can be easily implemented either using spread sheets or standard statistical software. Case-mix approach methods and mixed exponential can be as well implemented on statistical software, but they require some expertise in statistical modelling and computation, which is not always available. On the other hand, some of the drawbacks related to the estimation of phase-type distributions have been already discussed in Section 2.4.1 of this chapter.

2.6. Conclusions

Drawing on the consideration of the above examined criteria, the models with a case-mix analysis base,(i.e. finite mixture models and data mining techniques), seem to be those that approximate the closest, among the literature review studied here, to a statistical model to predict LoS in public hospital at Mexico which satisfies the research objectives proposed in this thesis.

Having in mind the objectives of developing a statistical model to predict patient length of stay (LoS) in public hospitals in Mexico which A) approximates to the distribution of the LoS, B) recognises and addresses the heterogeneity population problem, C) supplies LoS predictions for individual patients and cohorts of patients (within and between hospitals) and D) demonstrates a solid application into the decision-making process; it is found that the models with a case-mix

analysis base contribute to objective A by handling properly the features of the LoS data. They contribute to objective B by identifying the factors (covariates) that affect LoS and by clustering patients into homogenous groups. In addition, they can be used for making inferences on individual patients or groups of patients, they model probabilistic relationships, they are relatively easy to estimate and work well with small samples.

However to fulfil thoroughly the research objectives, mentioned above, and answer the research questions, the models that will be proposed in this thesis should encapsulate advances from previous models developed by others. These advances will be discussed in detail in the next chapter.

Having this theoretical background clearly defined, the next chapter will highlight the methodological approach that is adopted by the researcher to model LoS at public hospitals in Mexico, where the previous points will be developed and explained in more detail.

3 METHODOLOGY AND PRELIMINARY ANALYSIS

In the previous chapter, the relevant research in the field of estimating and predicting LoS was explored and discussed. As a result of this review, the models that are more suitable to model LoS, according to the criteria previously defined, were identified (i.e. finite mixture models, data mining, survival analysis, etc.). However these models have certain limitations to overcome as well as areas of opportunity to take advantage of.

This chapter firstly identifies the primary characteristics of the model that will help to overcome the limitations and tackle the areas of opportunity mentioned in the literature review (Section 2.6). Secondly, the chapter outlines the formal methodology to build the model that satisfies the research objectives and answers the research questions. Thirdly, the last part of the chapter is devoted to a brief preliminary analysis of the data.

3.1. Areas of opportunity

Let us recall the research objectives of developing a statistical model to predict patient length of stay (LoS) in public hospitals in Mexico which A) approximates to the distribution of the LoS, B) recognises and addresses the heterogeneity population problem, C) supplies LoS predictions for individual patients and cohorts of patients (within and between hospitals) and D) demonstrates a solid application into the decision-making process.

In order to fulfil thoroughly these objectives and answer the research questions, the models that will be proposed in this thesis should encapsulate the following advances from previous models:

1. Integrating the full case-mix: Most of the models and methods described in the literature review are focused on a particular patient cohort or patient population, for example geriatric patients, stroke patients, patients with mental diseases, patients

undergoing major surgery, etc. However if the aim of this research is to provide a model to predict LoS to improve the hospital decision-making process, it is desirable that it includes the whole hospital case-mix. Therefore the model proposed in this thesis will consider almost all types of patients admitted to hospital during the period of study. Moreover, in the literature review the analysis of factors influencing LoS focused only on those associated with the medical condition or cohort of patients under study. In this research, all the factors that are universal for the full case-mix are considered.

2. Combining models: Although both finite mixture models and data mining techniques have certain model attributes that meet the research objectives, the researcher believes that a combination of these two models and others, discussed in the previous sections, should boost their performance compared to alternatives based on single models. See points 4 and 5 for more details.
3. Exploring other probability distributions: Although the finite mixture models have been already used to model LoS, most of the research has focused on the Gaussian mixture model. However one of its limitations is that it is also defined on negative real numbers, which is unrealistic for LoS. The model proposed in this thesis will explore other probabilistic distributions to create the mixture of distribution that have been already successful on modelling LoS as single models like Lognormal, Gamma, Gaussian and Poisson, in order to try to find the best fit for the Mexican LoS data.
4. Understanding the associated factors on three different levels: In order to meet objective B, while addressing objectives C and A, three different and independent research approaches will be fully explored:
    a. Individual patients: By understanding how patient attributes shape the LoS distribution of each individual (which is in the form of a mixture of distributions). Here, the finite mixture model, which proved to be the best fit, will be extended to account for covariates using a generalised linear model principle.
    b. Cohorts of patients within a hospital: By making each group or cohort correspond to a component of the finite mixture model. In this research line different data mining techniques will help to delineate the relationships between each group and patient attributes.
    c. Cohorts of patients between hospitals: This research line requires an understanding of how the environment or context of each hospital affects the LoS by using a multilevel or hierarchical structure.
5. Enhancing its application: To address the last objective of this thesis, the model proposed in this thesis will review the role of survival analysis in the field of finite

mixture models for a better understanding of patient flow and its translation into bed requirement calculations. The nature of survival analysis provides valuable information that has not been yet fully recognised in the field of operational research for healthcare.

3.2. Methodology

The methodological approach used through this thesis is based on the classical data analysis approach, which has the structure described below in Figure 3.1:

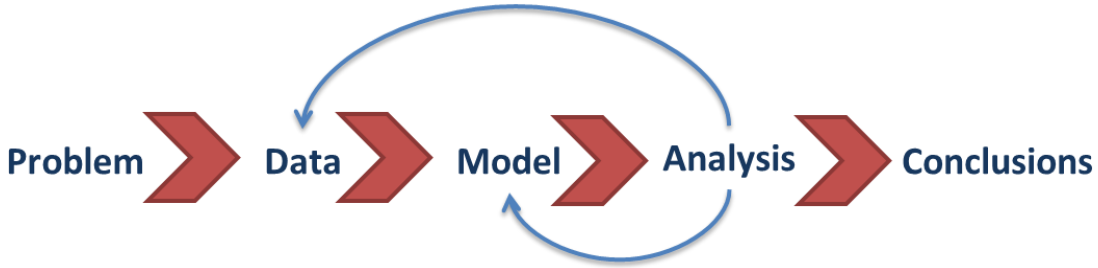

Figure 3.1: Classical data analysis approach

Unlike other approaches (e.g. in Exploratory Data Analysis the model is selected after data is collected and analysed), the main feature of the classical approach is that a model (or models) has (have) been already selected, according to the research objectives and the literature review of previous chapters. Therefore, the stage of analysis focuses on evaluating the appropriateness of the proposed model to the data. However, in this research study, the analysis of the model also contributes to redefine the data that enters the model and the model itself.

Accordingly, the classical methodological approach is broken down into five main stages: 1) definition of the problem, 2) data collection and manipulation, 3) model building, 4) analysis of the model outcomes and 5) conclusion and recommendation.

In this context, the problem definition has already fully described and discussed in Section 1.2 which can be summarised in few words as the need for more complex and sophisticated tools to understand length of stay at public hospitals in Mexico. The second and third components of the methodology (data and model) are described on the oncoming sections.

3.3. Data

Cluster sampling was used as a sampling technique, where the hospital population (i.e. the Mexican hospitals described in Section 1.3) was divided in natural groups (medical wards) and a sample of these wards was selected. The selected wards were Adult medicine, General surgery

and Trauma (ISSEMyM only)[6], which account for the majority of admissions at both hospitals. Notice that Obstetrics was omitted since the length of stay is controlled by public and organizational policies.

Then, the data used to build the models for this thesis came from a secondary source in the form of routine patient records. It corresponds to almost 14,600 patient records from 2005 to 2009 (the ISSEMyM dataset contains 1300 observations and MRC dataset contains 13218). Every record represents a hospital episode. Although each hospital collects different information about the patient, each record contains the value of LoS in days and the information of the clinical history form (which is filled upon patient admission). Table 3.1 contains the names and descriptions of the variables found in both hospital datasets. The reason why merely routinely collected data was selected to build the models was because it is of interest here to answer the research question whether an efficient model to predict LoS can be derived from this type of data only.

Variables	**Description**
Age	Age of the patient
Diagnosis	Main health problem or disease, cause of the hospitalisation, coded according to the International Disease Classification version 10 (ICD10)
Gender	Gender of the patient
Length of stay (LoS)	Number of nights spent at hospital
Origin	First hospital area from which the patient was referred for hospitalisation: Accident Emergency, Outpatient clinic or other (ISSEMyM only)
Previous visits	Number of previous hospitalisations
Surgical procedure	Main surgical procedure coded according to the International Procedure Classification version 9. This is not a mandatory field since not all the patients require surgery.
Number of surgical procedures	Total number of surgical procedures that patients underwent during their stay. This is not a mandatory field since not all the patients require surgery.
Ward	Ward where the patient was treated: Adult medicine, General surgery or Trauma (ISSEMyM only)

Table 3.1: Variables in the dataset for both public hospitals

The reason why the ISSEMyM dataset contains a smaller number of entries than MRC is due to the fact that most of the data was still in paper format; and it took a considerable amount of time to collect and transcribe the data. However, the result was a richer dataset containing other very important socioeconomic, demographic and clinical data. The next table (Table 3.2) summarises the variables found just in the ISSEMyM dataset.

[6] In the MRC hospital, the Trauma ward does not exist

Variables	Description
First Diagnosis	Health problem or disease diagnosed during the first medical evaluation at hospital, coded according to ICD10
Number of diagnosis	Total number of diagnosed medical conditions
Number of comorbidities	Total number of previously diagnosed medical conditions in addition to the primary disease or disorder.
Occupation	Patient principal economic activity (i.e. employed or unemployed, etc.)
Education level	Patient maximum level of studies (i.e. none, primary school, technical education or university)
Inherited family history • Diabetes • Hypertension • Neoplastic • Other	Binary variable(s) which describe whether any of these medical conditions are present in the patient family
Personal non-pathologic history • Environmental factors • Smoker/Drinker • Exposition to pollution	Categorical variable(s) which score(s) the level of quality of patient housing and hygiene; and level of exposition to drugs, cigarettes, alcohol or polluters
Personal pathologic history • Previous surgical procedures • Allergies • Transfusions • Others	Binary variable(s) which describe whether the patient had in the past any of these medical conditions or interventions

Table 3.2: Variables in the dataset for ISSEMyM hospital

One of the first challenges to address is that the variables "first diagnosis", "diagnosis" and "surgical procedure" described in Table 3.1 and Table 3.2 contain around 330, 850 and 200 different ICD codes respectively; complicating the inclusion of these variables for further statistical analysis. To handle the problem by reducing the number of categories per variable, hierarchical cluster methods based on the chi-square dissimilarity measure will be used (more on this in Section 4.2)

Furthermore, with such a rich dataset including so many variables, there is a need for conducting a variable selection process in order to enhance the predictive performance of the successor models for this thesis. In Section 4.4 a formal methodology to select the significant variables for the LoS will be applied combining multiple stepwise regression and bootstrapping.

On the other hand, some other data and valuable information were obtained, during an initial stage of this research, from casual and informal interviews conducted by the researcher with medical staff and decision makers from both hospitals under study. These interviews were very useful to identify the research problem and to define the objectives of this thesis. Moreover, they were used to understand some aspects about their day to day processes such as bed management, data collection, admission and discharge policies, etc.[7] In addition, other interviewed subjects included the staff in charge of the data collection and clinical files. This was done in order to know more about the variables included in the database, how these variables were recorded (and coded) and most important to assess the reliability of coding of conditions and interventions and the completeness of data (Black and Payne, 2003).

The last two points represented a challenge because the ISSEMyM dataset did include ICD coding or any coding for diagnosed conditions and interventions. In this case, an ICD coding specialist from the MRC general hospital was hired to code the entire database. Moreover, the completeness of the paper format data (ISSEMyM data) was highly questionable and unfortunately some of the records had to be excluded from the study given their percentage of missing values was more than 20% (Hair, 2009).

3.4. Model Building Process

Based on what has been discussed in the last two chapters, this stage of the methodology can be divided into three main phases: the probabilistic model, the analysis of the associated factors and the application of the model to the decision-making process. In the following subsections, these phases will be described and discussed briefly (for more details the reader is referred to the chapters where each phase is presented)

3.4.1. The probabilistic model

One of the main objectives of this research is to develop a probabilistic model that approximates to the distribution of LoS. The literature review already stated the appropriateness of the finite mixture models for such task, since it is believed that the observed distribution of LoS in fact, may be represented by a mixture of several different distributions (Quantin et al., 1999), which may help on handling the data characteristics of skewness and heavy tails. In addition, the finite mixture model clusters patients into homogenous groups, which partially address the heterogeneity problem. In Chapter 4, four different mixture of distribution will be explored and compared: Gaussian, Lognormal, Gamma and Poisson.

[7] Whenever, the data or information comes from an interview process, the reader will be referred to the exact source.

In a first step, the different mixtures of distribution will be fitted to a joint dataset containing records from both hospitals; later the same procedure will be repeated but using a separate dataset per hospital. This will be done with the aim to test the research question whether the selected model can be used indiscriminately for other hospitals or each hospital needs its customised version.

3.4.2. Understanding the associated factors

Once the best probabilistic model has been selected, it will be extended to account for covariates. In this context, internal and external factors associated with LoS will be explored in three different research approaches.

In the first approach, the patient attributes not only predict the LoS homogenous group (LoS category) to which the patient belongs, but they shape its LoS probabilistic density curve. The finite mixture model, defined in the first stage of the model building process, will be extended to accommodate patient characteristics using a generalised linear model principle.

In the second approach, the patient attributes or variables will be used to predict the LoS category to which the patient belongs; where each group or cohort correspond to a component of the finite mixture model. Subsequently, some of the most common data mining prediction techniques will be evaluated, namely: Logit regression, decision trees (CART, QUEST, C4.5 and CHAID), Naive Bayes and hybrid methods (Naive Bayes trees and Logistic Trees) in order to find the best method to delineate the relationships between each group and patients attributes.

In the third approach, a multilevel or hierarchical structure will expand the previous approach, in order to understand the environment in which the patient is treated and how it affects LoS, providing a model that adapts itself from a local level (hospital) to a regional or institutional level.

There is a fundamental difference between the first and second approach: in the first approach, named individual-based, all patients are different; their individual characteristics predict firstly the membership to one of the LoS categories and secondly those same characteristics define the shape of their LoS probabilistic curve and its associated expected LoS (see Figure 3.2). In the second approach named “group-based”, all patients within LoS categories are the same. Although their individual characteristics help to predict the membership of LoS, their length of stay probabilistic curve and associated expected LoS is defined by the parameters of the category itself (see Figure 3.3).

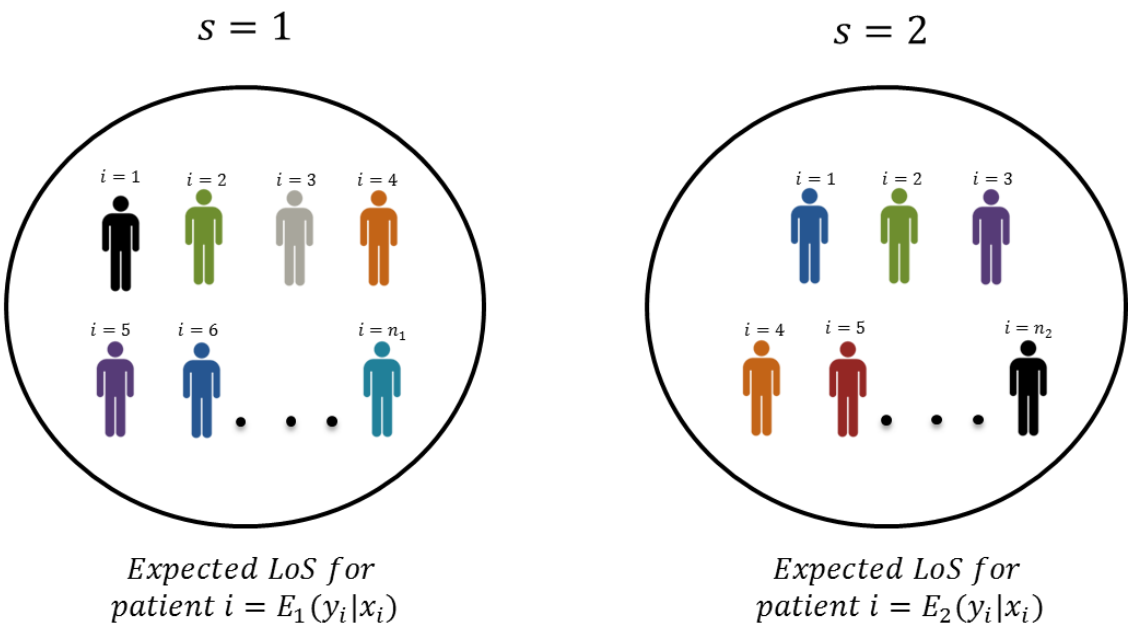

Figure 3.2: The individual-based approach, where it is assumed that within each component or LoS category, every patient is different with an associated expected value of LoS and a probabilistic density curve. Therefore the expected LoS of patient is equal to the conditional mean value of $y_i$given the values of $\boldsymbol{x_i}$, where y_i is the LoS of patient *i* and $\boldsymbol{x_i}$, is the vector containing the attribute of patient *i*

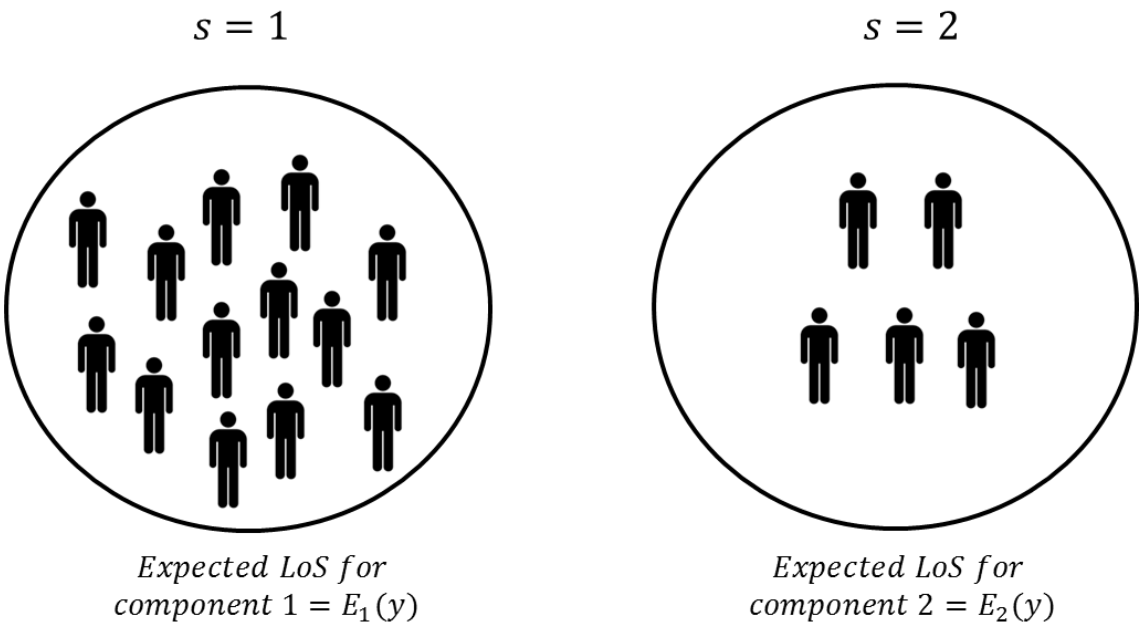

Figure 3.3: The group-based approach, where it is assumed that all patients within each component or category *s* have the same LoS probability density and associated LoS expectancy. Therefore the expected LoS of a patient is equal to mean value of *y* of the component where he or she belongs.

It is important to highlight that these three research lines will allow a better understanding of the heterogeneity patient population problem and will enable LoS predictions for individual patients and cohort of patients within and between hospitals.

3.4.3. Application

In the final part of the model building process, survival analysis will complement the three different approaches, in order to provide insights into patient flows and the translation into bed requirements calculations. This requires the following steps: first the individual and group-based approaches are extended to contribute to a better understanding of patient flow in a bed management context. Second, the current methodology to calculate bed requirements will be discussed and two other methods based on the finite mixture models would be suggested as an alternative approach. Finally the applications of the multilevel group-based model will be explored with a discussion of how this can be extended to provide predictions for institution, hospital and patient levels.

3.5. Preliminary Analysis

Any modelling process should start with an exploratory statistical analysis of the data which allows one to gain insights into the dataset, uncover underlying structures, detect outliers (or other anomalies) and test assumptions (Nist/Sematech, 2003). This last section of the chapter is devoted to such preliminary analysis of the LoS data for both hospitals.

Length of stay (LoS) is the object of study in this thesis. It represents the number of nights a patient spends in hospital. As the variable is measured by counting, it is defined on a ratio scale (i.e. equal distance between each value and an absolute zero point) and therefore it will be treated as a continuous variable. The distribution of the length of stay of the whole sample is shown in the histogram in Figure 3.4 and Figure 3.5 shows the distributions for each hospital. The basic descriptive statistics of the distribution are shown in Table 3.3.

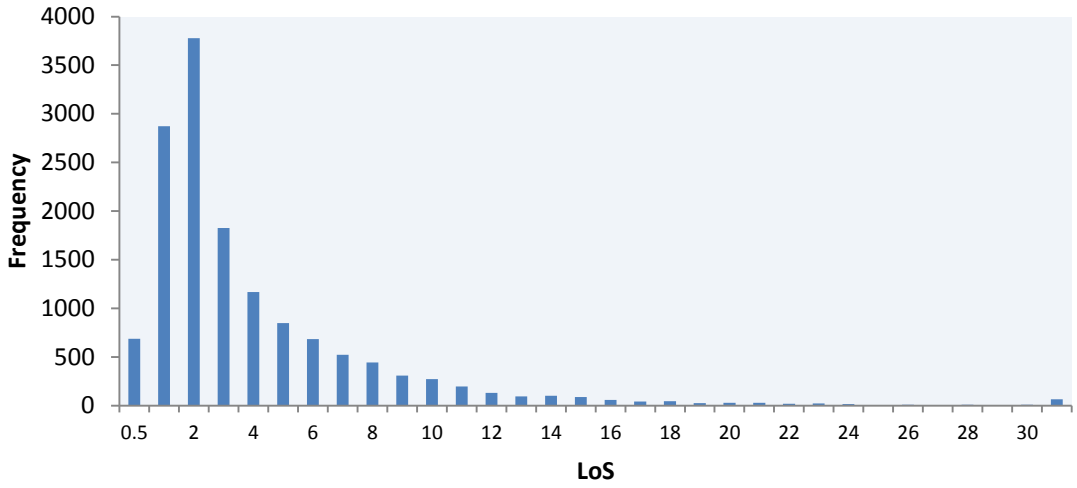

Figure 3.4: Histogram for length of stay

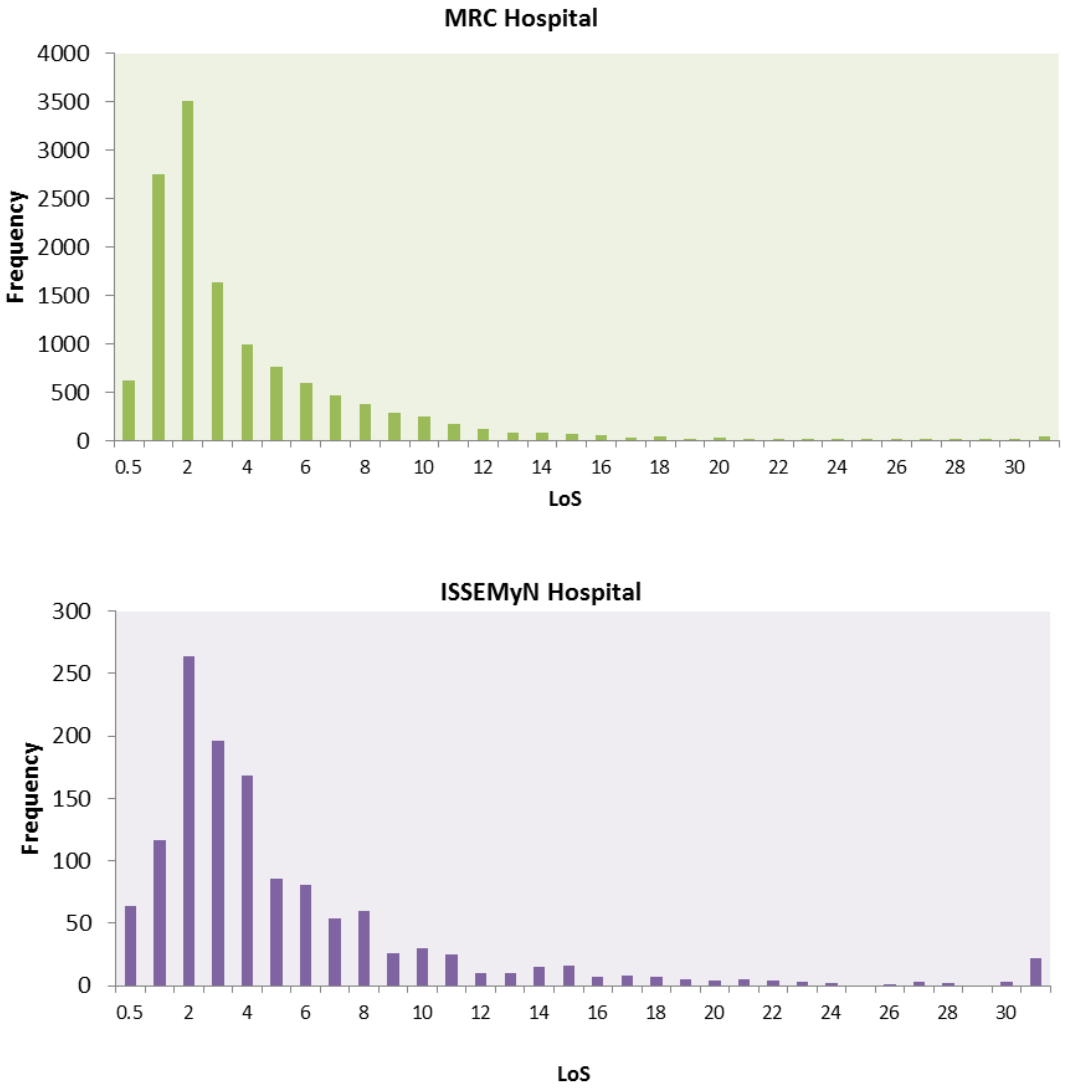

Figure 3.5: Histograms for MRC and ISSEMyM Hospital

	Mean	Std. Deviation	Variance	Minimum	Maximum	Skewness	Kurtosis
Overall LoS	4.13	4.95	24.58	0.50	285	4.89	45.59
ISSEMyM	5.73	7.77	60.50	0.50	285	4.92	35.25
MRC	3.97	4.55	20.77	0.50	196	4.31	34.73

Table 3.3: Descriptive statistics for LoS

In general terms, one can see that the distribution of the sample for each hospital does not differ from the entire sample: while the high and positive values of skewness indicate a peak of records on the left of the distributions, the positive and high values of kurtosis indicate pointy and heavy-tailed distributions and distinct peaks near to the mean. The coefficient of variation for the whole sample is 119.8%, indicating a very high variability of the data.

Most of the classical statistical tests rely on the assumption of normality of the outcome variable (LoS). By visual inspection, the histograms look far from normal: in addition the high values of skewness, kurtosis and the presence of outliers suggest that the distributions do not follow a normal distribution. The normal Q-Q plot for the whole sample in Figure 7 confirms the lack of normality of the whole sample (i.e. in a Q-Q plot generated from normally distributed data it is expected that observed values would follow the straight line).

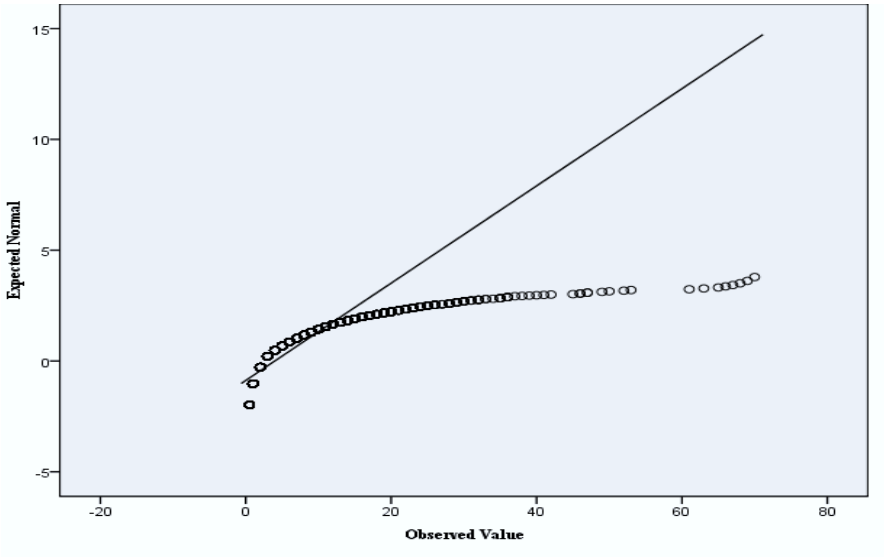

Figure 3.6: Normal Q-Q plot of LoS

As the assumption of normality is important for future statistical tests, LoS need to be transformed to correct the non-normality. Two different transformations were tested: the log-

transformation and reciprocal transformation. These transformations are often used to correct unequal variances (another important assumption) and positive skewness like the one present in the data (Field, 2009).

Best results were obtained from the log-transformation which reduced considerably the skewness (from 4.9 to .29) and kurtosis values (from 45.6 to -.29). The Q-Q plot depicted in Figure 3.7 shows a slight yet acceptable deviation from normality, although the presence of outliers is still notable.

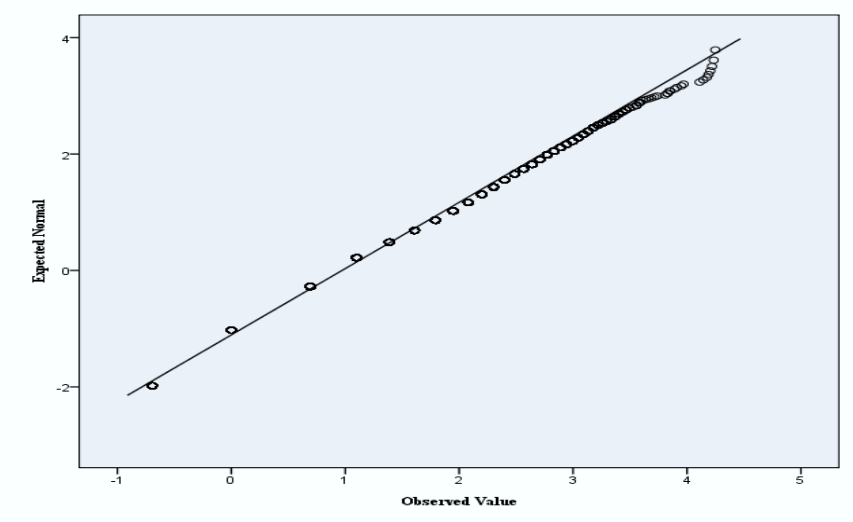

Figure 3.7: Normal Q-Q plot of transformed LoS

In respect to the outliers, there is no evidence that they are the result of errors in data entry so they will not be erased from the dataset. Therefore, all the outliers were changed to be one unit above the next highest observation in the dataset (i.e. 70 days). This approach was suggested by Field (2009) to deal with outliers when transformations fail.

3.6. Summary

This chapter starts with a brief explanation of the methodology to follow in the upcoming chapters. Broadly speaking, the methodological approach is based on the classical data analysis approach where a model has been already selected before any preliminary analysis of the data based on other criteria or knowledge (i.e. literature review).

This chapter also described the data that will be used on this research. In particular, it emphasised the appropriate sampling technique to use, the characteristics of the datasets and the most suitable data processing techniques to conduct, based on such characteristics (i.e. the

reduction of the size of some categorical variables and a preliminary variable selection process will be discussed in the following chapter).

Furthermore, this chapter stated how the model building process is divided into three main stages: the probabilistic model, which includes the selection of the best finite mixture model for the LoS data; the understanding of the associated factors, which is explored through three different research lines; and the application of the models to the decision-making process.

Finally, this chapter concluded with an exploratory analysis of the length of stay to reveal the shape and distribution of the data. This analysis revealed that a peak of records was evident in the left of the distribution with a heavy tail on the right, and thus confirming the lack of normality and the presence of outliers.

Because the normality assumption is required for some of the statistical tests to be performed later, some transformations were tested. As a result, the log-transformation was identified as being the most appropriate for the data. However the transformation did not improve the presence of outliers which then had to be modified to be one unit above the next highest score in the dataset.

4 THE PROBABILISTIC MODEL

This chapter explores the role of finite mixture models in modelling LoS, aiming to approximate to the distribution of LoS and to capture the variability in the data.

However, to address the issue that the dataset contain a variety of different codes for "first diagnosis", "diagnosis" and "surgical procedure", the data will be clustered into groups with similar LoS using hierarchical cluster methods.

4.1. A Model-based Cluster Approach

The logic of the finite mixture models is based on the idea that a continuous variable in a large sample could consist of two or more clusters of observations with different means and perhaps different standard deviations within each sample. Therefore the observed continuous variable is a mixture or sum of those two or more distributions with different parameters (Maclachlan and Krishnan, 1997).

It is because of the clustering element that finite mixture models are sometimes known as model-based clustering. These clustering algorithms based on probability distributions are an alternative to heuristic-based models like k-means or hierarchical clustering (Yeung et al., 2001). One of the main advantages of this approach is that the problem of choosing the right number of clusters and an appropriate clustering method can be reduced to a statistical model choice problem (Fraley and Raftery, 2002) where the task of choosing the optimal number of components or comparing among different models can be performed via the usual methods such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and log-likelihood

Some of the other important advantages are that predictors can be integrated into the model as covariates; multilevel data structures (e.g. patients nested within hospital) can also be taken in account. In addition, the model can be adapted to consider dependent observations or repeated measurements (Dias, 2004); also outliers can be handled by adding extra components to the model (Fraley and Raftery, 2002).

Moreover, the clusters can be understood as homogenous groups of individuals or objects. This concept of clustering individuals is not new in the modelling length of stay literature. Many studies classify patients according to their length of stay in clusters or categories (e.g. Harrison and Millard (1991), Faddy (1994), Taylor et al. (2000), Harper (2002), Marshall et al. (2005b) and Abbi et al. (2008)).The specific characteristics of each category (i.e. the intervals of time that each category considers) change arbitrarily from researcher to researcher, according to the individual objectives of such classification[8]. In this context, model-based clustering using finite mixture models can also be used to define carefully such characteristics of each category to make them not just statistically but also clinically meaningful.

4.1.1. Finite Mixture Models

Let us consider a sample of *n* individuals (patients). Each individual is denoted by $i=1,\ldots,n$ and it is characterised by the random vector $\mathbf{Y_i}$, containing the random variables corresponding to the measurements of some features of the individuals under study (e.g. length of stay), and where y_i are the observed values.

The finite mixture model with *S* components or clusters is defined by Equation 4.1

$$f(y_i;\varphi) = \sum_{s=1}^{S} \pi_s f_s(y_i;\theta_s) \tag{4.1}$$

where π_s are the mixing proportions that can be interpreted as the component relative size and satisfy $\sum_{s=1}^{S} \pi_s = 1$ and $\pi_s > 0$. Within each component, observation y_i is characterised by density $f_s(y_i;\theta_s)$, and φ represent all parameters in the model. Each component belongs to the same parametric family of distribution but with different parameters.

The estimation of the mixture is carried out using the maximum likelihood (Equation 4.2)

$$\max_{\pi,\theta} \ln L = \sum_{i=1}^{N} \left(log \left(\sum_{s=1}^{S} \pi_s f_s(y_i;\theta_s) \right) \right) \tag{4.2}$$

[8] The most common approach is to find the classification which matches the resource planning methodology at the hospital, which is based on personal or clinical judgement (Côté, 2000)

The reader is referred to (McLachlan and Peel, 2000) and (Dias, 2004) for a more detailed definition of finite mixture models.

The choice of the distribution is usually from among the well-known distributions from the exponential family. In this research, the probabilistic distributions described in Table 4.1, were fitted to the LoS data to find the most appropriate finite mixture model to describe it. The Gaussian distribution is usually the first choice for most of the applications and is well accepted by non-technicians. Abbi et al. (2008) fitted a Gaussian mixture model to stroke LoS observations; however one of its limitations is that it is also defined on negative real numbers, which is unrealistic for LoS.

Lognormal, Weibull and Gamma distributions have proved to work well with LoS data (Marazzi et al. 1998; Xiao et al. 1999; Graves et al. 2009). Finally, a Poisson distribution was also fitted to the LoS data based on previous work by Singh and Ladusingh (2010) where LoS is defined as a count variable.

Distribution	Notation	Domain
Normal/Gaussian	$N(\mu, \sigma^2)$	$(-\infty, \infty)$
Lognormal	$\ln N(\mu, \sigma^2)$	$(0, \infty)$
Gamma	$G(\alpha, \beta)$	$(0, \infty)$
Poisson	$P(\lambda)$	0,1,2…

Table 4.1: Distributions used to fit LoS

The five models with up to 4 components each were fitted on STATA using `fmm` command (Deb and Trivedi, 1997; Deb and Holmes, 2000; Deb et al., 2011) where $S = 2, 3\ and\ 4$. The researcher has decided to constraint the finite mixture model to those values to preserve the simplicity of the model and to provide a clinical and natural interpretation of the number of components.

To choose the optimal number of components (i.e. the best fit) and compare among different models, log-likelihood, AIC and BIC statistics were calculated[9] (see Equations 4.2-4.4).

$$AIC = 2k - 2ln(L) \tag{4.3}$$

[9] The standard log-likelihood ratio statistic used to compare nested models is inappropriate for mixture models because they do not follow an asymptotic chi-squared null distribution. (Böhning et al., 1998)

where k is the number of parameters and L the maximized value of the likelihood function for the estimated model.

$$BIC = -2ln(L) + k * \ln(n) \tag{4.4}$$

where n is the sample size.

AIC and BIC can be used indiscriminately, however log-likelihood can be used just when comparing models with the same number of parameters. In addition, especial attention should be paid when comparing the Log-likelihood of the Poisson mixture against the continuous models (Weiss, 2010)[10].

Once the preferred model was elected, the intervals of time for each mixture component were defined using the highest posterior probability that observation y_i belongs to component s (see Equation 4.5)

$$\Pr[y_i \in components \mid y_i;\ \theta] = \frac{\pi_s f_s(y_i|\theta_s)}{\sum_{j=1}^{S} \pi_j f_j(y_i|\theta_j)} \tag{4.5}$$

where posterior probabilities is derived using the Bayes theorem and it can be broken down in three elements: the prior probability π_s, the conditional probability $f_s(y_i|\theta_s)$ and the unconditional probability $\sum_{j=1}^{S} \pi_j f_j(y_i|\theta_j)$.

4.1.2. Results

Table 4.2 shows the results of fitting finite mixture models to both hospitals dataset

[10] In theory the log-likelihood of a discrete probability model cannot be compared against the log-likelihood of a continuous probability model. The former is exactly equal to the joint probability of the observer data whereas the latter is equal to the joint density of the observed data. Therefore,$P(X = x_i)$ which is undefined in the continuous realm is reinterpreted as $P(x_{i-1} < X < x_{i+1})$ using a midpoint approximation (i.e. the standard way to estimate discrete probabilities when using continuous models). The midpoint approximation states that:
$P(X = x_i) \approx f(x_i)\,\Delta x/2$ where $\Delta x = x_{i+1} - x_{i-1}$

Distribution	Number of components	Log-likelihood	AIC	BIC	Number of parameters
Normal	2	-36331.19	72672.39	72708.27	5
	3	-34284.60	68585.21	68642.62	8
	4	-33861.12	67744.24	67823.18	11
Lognormal	2	-33213.29	66436.58	66472.46	5
	3*	**-33195.81**	**66407.62**	**66465.03**	**8**
	4		No convergence		11
Gamma	2	-33496.90	67003.80	67039.69	5
	3	-33266.12	66548.23	66605.65	8
	4		No convergence		11
Poisson	2	-36698.75	73403.50	73425.03	3
	3	-34840.39	69690.77	69726.66	5
	4	-34840.39	69694.77	69745.01	7

Table 4.2: Results when fitting mixture distribution models. * is the preferred model according to measures of goodness of fit.

The Lognormal and Gamma mixture models with four components failed to converge after nearly 300 interactions. This might suggest that the 4th component was an attempt to fit a small number of outliers (Deb et al., 2011).

According to the log-likelihood, AIC and BIC values (see Table 4.2) a Lognormal mixture model with three components is the best fit (See Figure 4.1). These results are not surprising: McLachlan and Peel (2000) showed how a Gaussian mixture model can be satisfactorily replaced by a Lognormal mixture model with fewer components in certain cases when modelling skewed data, such as the patient length of stay. Faddy (1994) found a single Lognormal model superior to a Gamma distribution when describing geriatric length of stay. Further, Marazzi et al. (1998) carried out a study on 3279 samples using single component Lognormal, Gamma and Weibull models to describe the distribution of LoS: the Lognormal model was found to be the best fit for most of the samples. On the other hand, the Poisson distribution was the least appropriate model, suggesting that treating LoS as a discrete count variable does not fully capture the nature of the data.

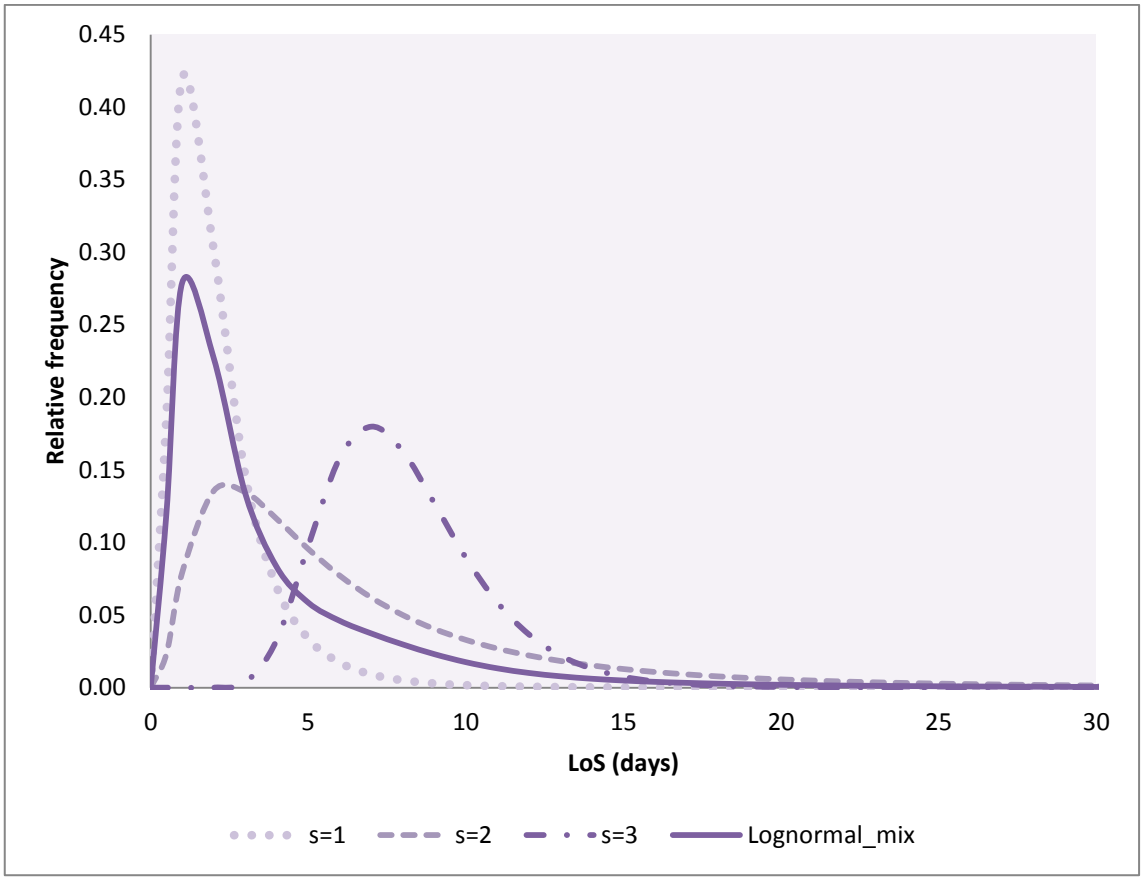

Figure 4.1: Three-component Lognormal mixture

However, the third component of the Lognormal mixture was found to be unnecessary, on calculation of the posterior probabilities. Figure 4.2 shows that the posterior probabilities for the second component are consistently higher than those from the third component, reducing the chance of any observation of belonging to third (and last) component. This phenomenon can be understood from the way that posterior probabilities are calculated. The prior probability of the third component (mixing proportion π_3) is very small (.059), indicating that this component is fitting a small group of outlier observations (i.e. just very few patient having very long LoS). In addition, Figure 4.3 shows that for LoS > 14 days, conditional probabilities for the second component (*s=2*) tend to be slightly higher than the conditional probabilities for the third component (*s=3*). By way of explanation, the probabilities of a longer length of stay occurring are higher in the second component than in the third component, even when the third component is added to the model to precisely accommodate more distant observations.

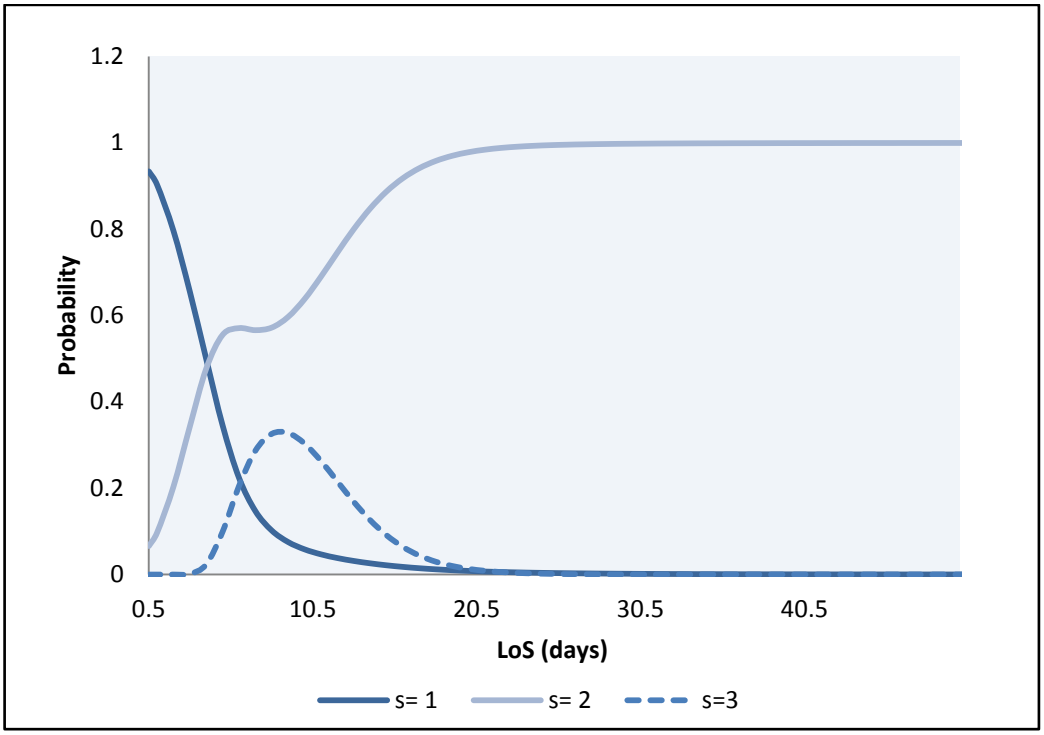

Figure 4.2: Posterior probabilities for the three-component Lognormal mixture model. Notice that that the posterior probabilities for the second component are consistently higher than those from the third component.

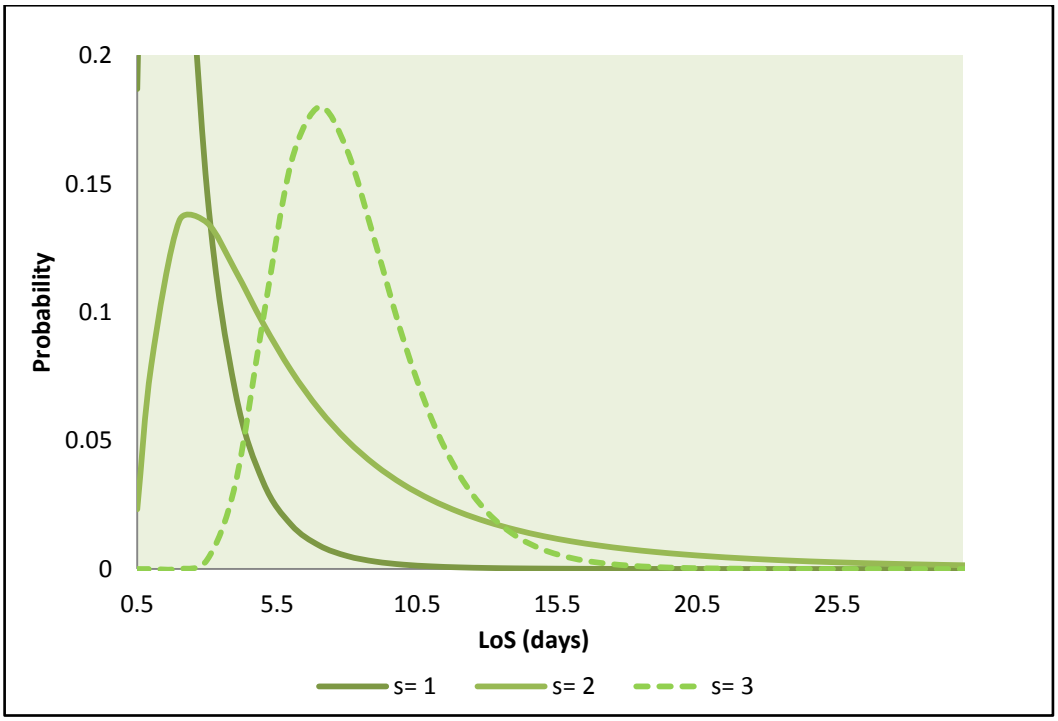

Figure 4.3: Conditional probabilities for the three-component Lognormal mixture model. Notice that for LoS > 14 days, conditional probabilities for the second component tend to be slightly higher than the conditional probabilities for the third component.

Therefore a more parsimonious model seems to be appropriate: a two-component Lognormal mixture model. Further analysis and results will be based on this model. Figure 4.4 displays the two components of the mixture (dotted and dashed lines). Notice that each component provides a local approximation to some part of the true LoS distribution (Deb et al., 2011). The mixture (light purple line) provides a very good fit of the LoS data. Figure 4.5 displays cumulative density functions (CDF) for the first and second component, and the mixture of distributions. Notice that the CDF of the mixture is very close to the LoS empirical distribution function (EDF)[11].Table 4.3 lists the parameters estimates for the selected model.

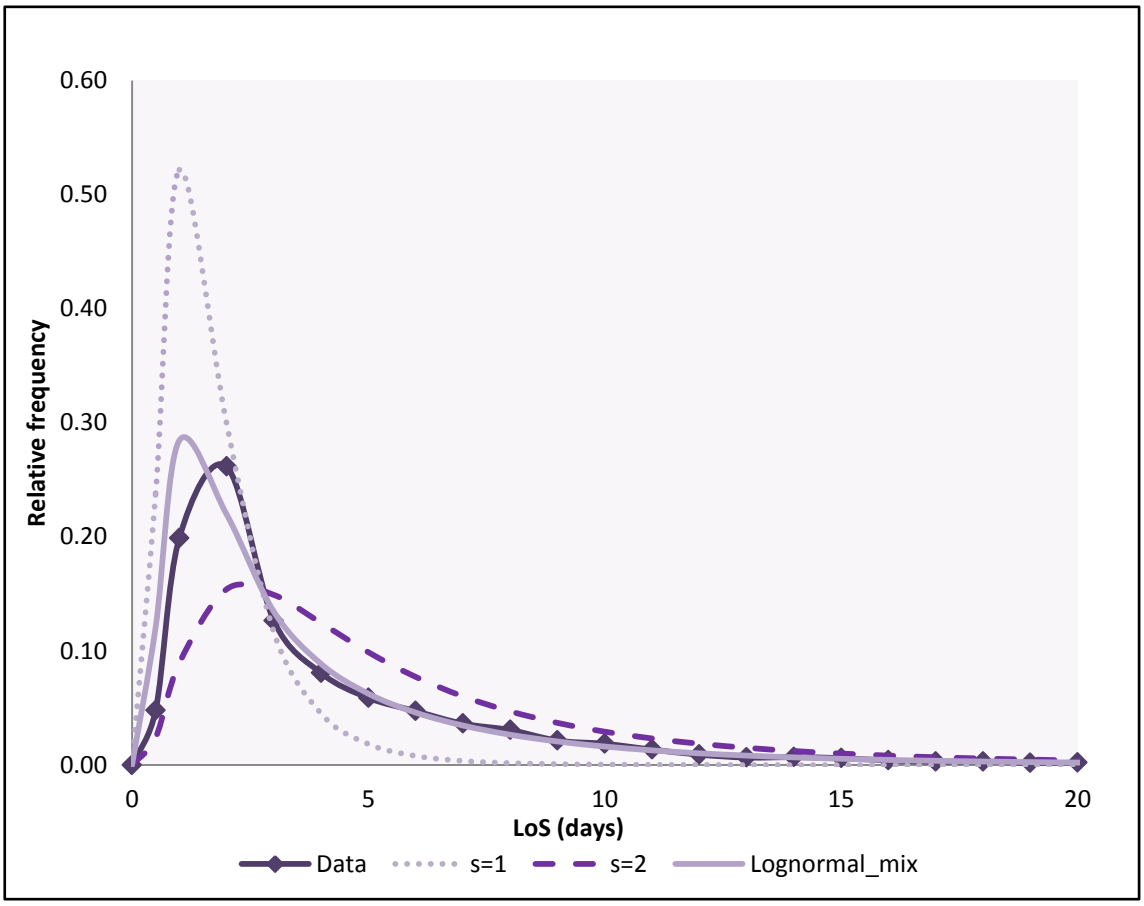

Figure 4.4: Empirical distribution of LoS approximated by two-component Lognormal mixture

[11] According to the Glivenko-Cantelli lemma, when the sample size (from where the EDF is computed) is large, the EDF is quite likely to be close to the CDF over the entire real line. In this sense, when the CDF is unknown, the EDF can be considered to be an estimator of CDF (DeGroot, 1986)

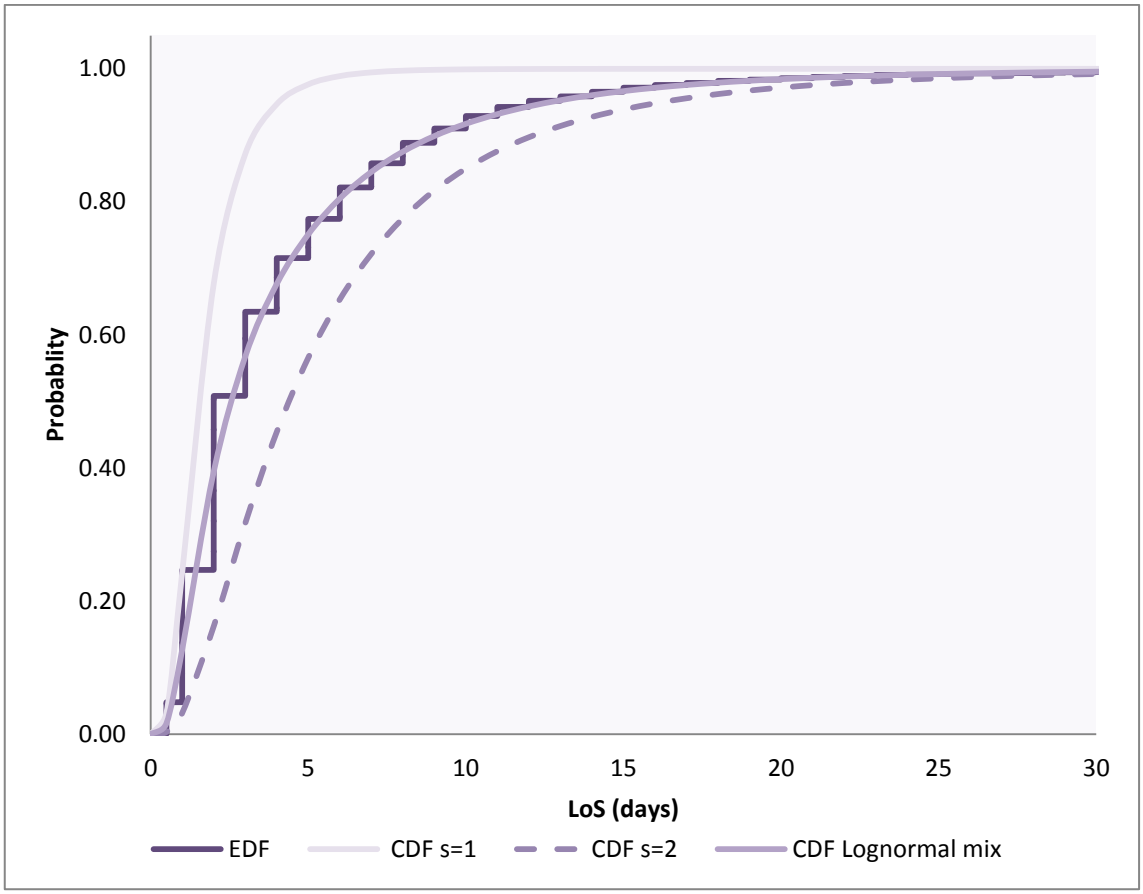

Figure 4.5 Empirical distribution and cumulative distribution functions

Parameter	first component	2nd component
μ	0.42 (.03)	1.48 (.06)
σ	0.60 (.01)	0.80 (.02)
π	0.45 (.05)	0.55 (.05)

Table 4.3: Two-component Lognormal mixture parameter estimates (standard errors)

It was already mentioned that one way to interpret finite mixture models is that each component can be referred as a subpopulation; therefore patients can be grouped in any of these subpopulations, clusters or categories according to the highest posterior probability of their LoS observation belonging to one of the mixture components. After fitting the fmm, posterior probabilities were calculated for each patient of the dataset. Then patients were clustered into components or categories according to those calculations. Table 4.4 highlights some descriptive statistics for each component.

Statistic	Sample	s=1	s=2
N	14518	7380	7138
Mean	4.13	1.46	6.89
Std. deviation	4.95	0.56	5.89
Skewness	4.89	-0.26	4.53
Kurtosis	45.52	1.39	36.63
Min	.05	0.5	3
Max	70	2	70

Table 4.4: Summary statistics for the two component Lognormal mixture and the LoS sample

From Table 4.4, it can be read that LoS data consists of two components or clusters: the first cluster is patients with LoS up to 2 days, referred to now on as patients with "short LoS" and the second cluster is patients with LoS from 3 days, referred to from now on as patients with "medium/long LoS". Accordingly, a new variable named "LoS category" was added to the dataset, which classifies patients into two statistically and clinically meaningful categories: "short LoS" and "medium/long LoS".

4.1.3. A Model for Each Hospital

The previous results were based on the data from both hospitals, so the question arises whether the same model can be applied to the two hospitals separately. To answer this questioning, the same methodology described in previous sections of the chapter, was applied for each hospital independently.

One can see from Table 4.5, that a two-component Lognormal mixture model was the best fit for the MRC hospital (AIC=59404.2 and BIC=59440.08). The Lognormal mixture with three and four components and the Gamma mixture model with four components failed to converge after 200 interactions. Notice that the parameter estimates of the two-component Lognormal mixture model (μ_1=0.44, μ_2=1.6, σ_1= 0.54, σ_2=0.73, π_1=0.55 and π_2=0.45) were very close to those obtained in Section 4.1.2, which is not surprising as most of the sample used in the previous section is from the MRC hospital. Figure 4.6 and Figure 4.7 show a reasonable fit of the two-component Lognormal mixture and Table 4.6 shows some statistics of each component after grouping patients. Almost equal to the results for both hospitals, the first cluster is patients with LoS up to 3 days, referred to from now on as patients with "Short LoS" and the second cluster is patients with LoS from 4 days, referred as patients with "Medium/Long LoS".

Distribution	Number of components	Log-likelihood	AIC	BIC	Number of parameters
Normal	2	-32219.47	64448.95	64484.35	5
	3	-30603.82	61223.63	61280.27	8
	4	-30264.22	60550.44	60628.32	11
Lognormal	**2***	**-29697.10**	**59404.20**	**59440.08**	5
	3		No convergence		8
	4		No convergence		11
Gamma	2	-29900.02	59810.03	59845.43	5
	3	-29742.85	59501.69	59558.33	8
	4		No convergence		11
Poisson	2	-32416.33	64838.66	64859.90	3
	3	-31111.70	62233.41	62268.81	5
	4	-31111.70	62237.41	62286.97	7

Table 4.5: Results when fitting mixture distribution models to MRC hospital. * is the preferred model according to measures of goodness of fit.

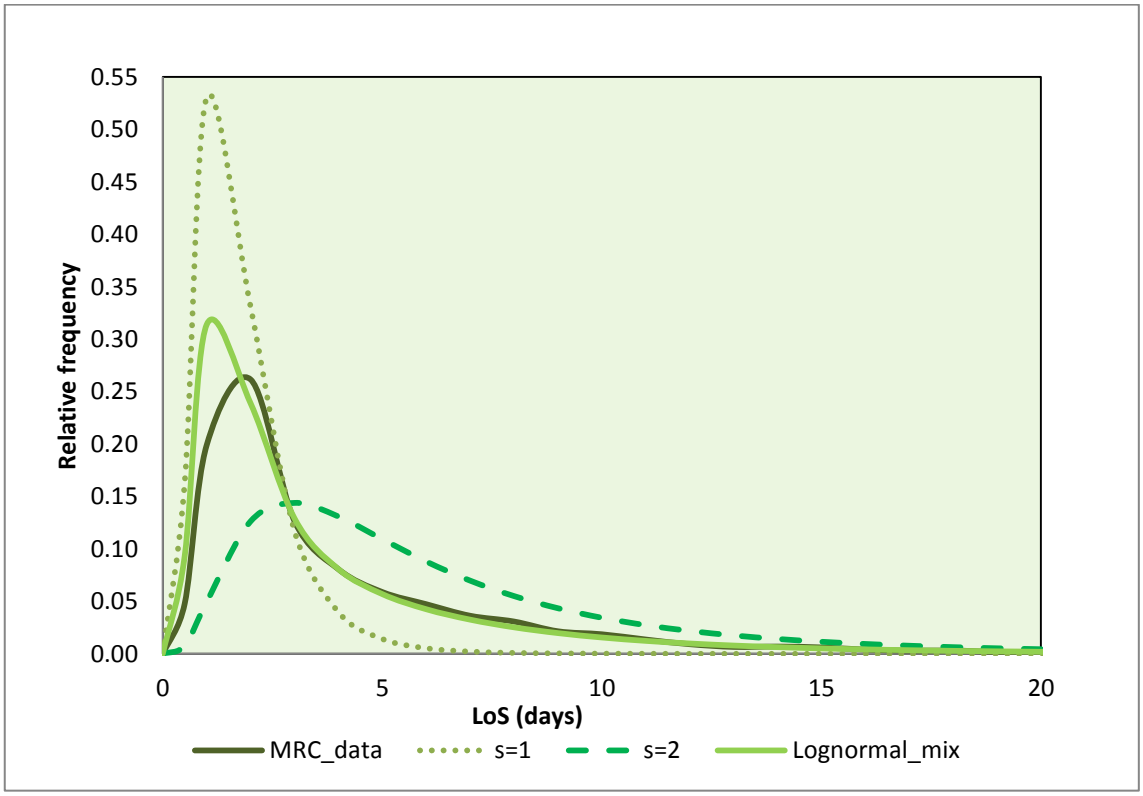

Figure 4.6: Empirical distribution of LoS at MRC hospital approximated by two- component Lognormal mixture

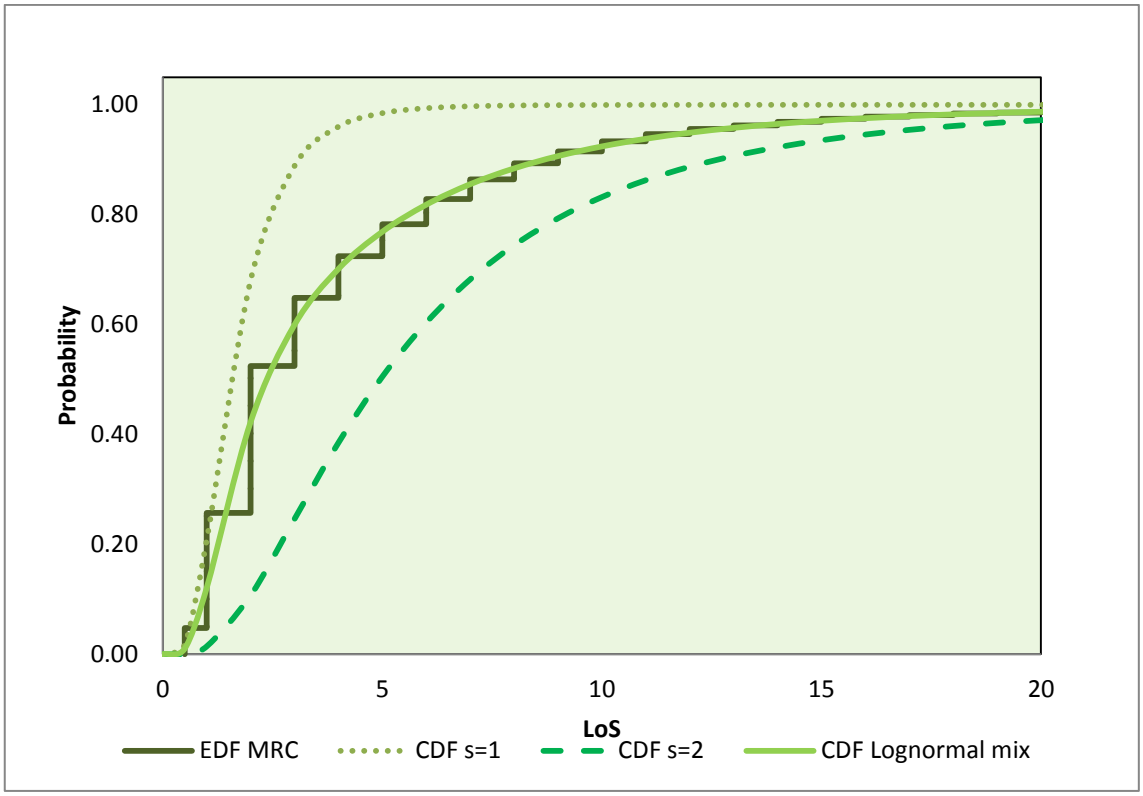

Figure 4.7:Empirical distribution and cumulative distribution functions for MRC hospital[11]

Statistic	Short LoS s=1	Medium/Long LoS s=2
N	8573	4645
Mean	1.75	8.07
Std. deviation	0.78	5.66
Min	0.5	4
Max	3	70

Table 4.6: Summary statistics for the two component Lognormal mixture for the MRC hospital

According to the results depicted in Table 4.7 for the ISSEMyM data, the Lognormal mixture with three and four components and the Gamma mixture model with four components failed to converge after 500 interactions. The best fit was a three-component Gamma mixture model (AIC=6827.02 and BIC=6865.39). However the mixing proportion for the third component is very small (π_3=0.007) and this could cause future problems when trying to classify patients into this group. Therefore, based on the evidence, it was decided to select the second best fit as the preferred model for ISSEMyM: a two-component Gamma mixture model (α_1=2.08, α_2=1.30, β_1=1.81, β_2=11.24, π_1=0.82 and π_2=0.18; AIC=6857.40 and BIC=6881.39). Figure 4.8 depicts

the first and second component, and the Gamma mixture which provides a good fit. Figure 4.9 displays the CDF and EDF curves. Notice that the CDF of the mixture is very close to the ISSEMyM EDF, indicating a good fit. Table 4.8 summarises some statistics for each component of the model.

Distribution	Number of components	Log-likelihood	AIC	BIC	Number of parameters
Normal	2	-3663.28	7336.56	7360.54	5
	3	-3497.79	7011.57	7049.95	8
	4	-3448.03	6918.05	6970.82	11
Lognormal	2	-3426.63	6863.26	6899.15	5
	3		No convergence		8
	4		No convergence		11
Gamma	2	-3423.70	6857.40	6881.39	5
	3*	**-3405.51**	**6827.02**	**6865.40**	**8**
	4		No convergence		11
Poisson	2	-3927.48	7860.97	7875.36	3
	3	-3541.87	7093.73	7117.72	5
	4	-3451.76	6917.52	6951.10	7

Table 4.7: Results when fitting mixture distribution models to ISSEMyM hospital. * is the preferred model according to measures of goodness of fit.

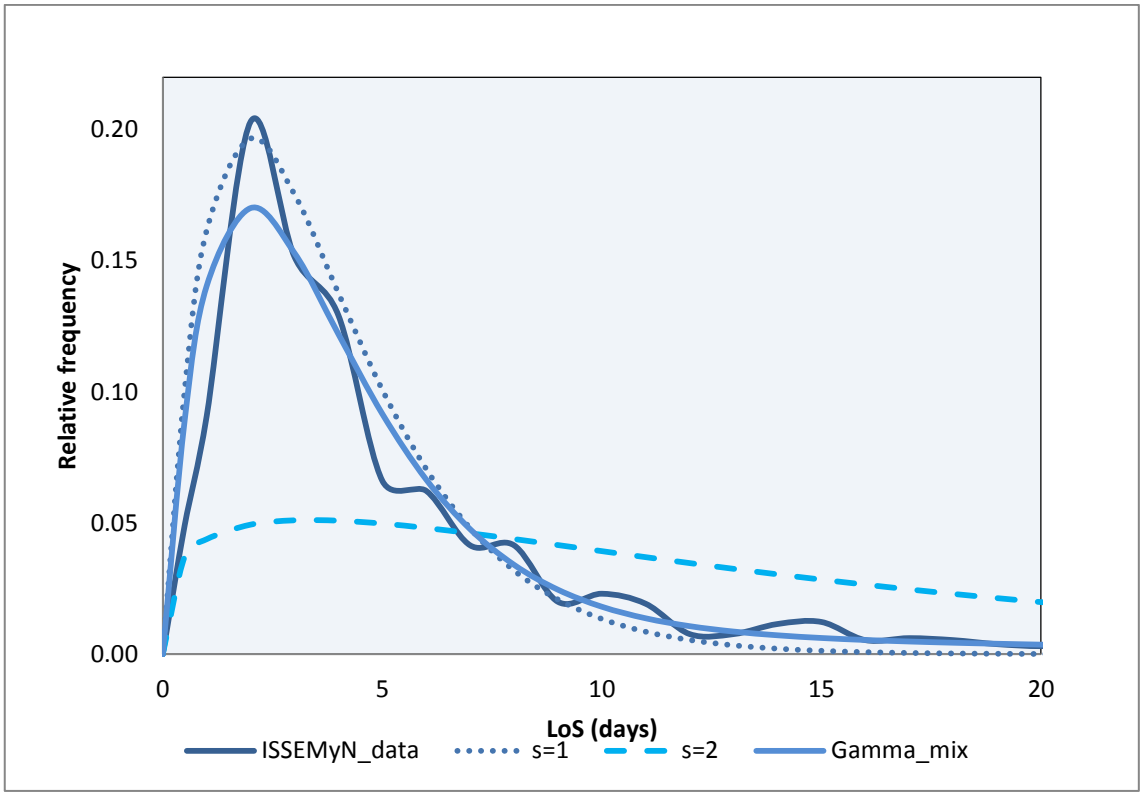

Figure 4.8: Empirical distribution of LoS at ISSEMyM approximated by two-component Gamma mixture

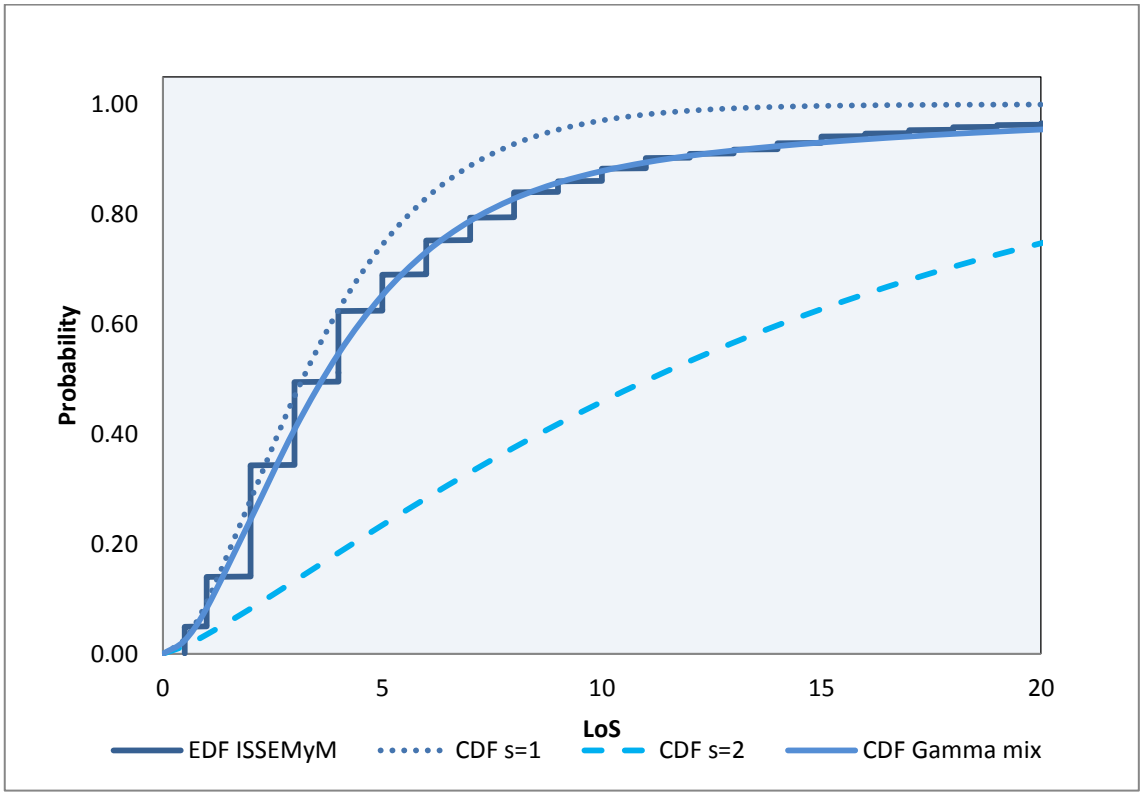

Figure 4.9: Empirical distribution and cumulative distribution functions for ISSEMyM hospital[11]

Statistic	Short LoS s=1	Medium/Long LoS s=2
N	1173	127
Mean	3.85	23.07
Std. deviation	2.59	15.01
Min	0.5	12
Max	11	70

Table 4.8: Summary statistics for the two component Gamma mixture for the ISSEMyM hospital

Based on the results, it seems that the model defined in Section 4.1.2 can be applied to the MRC dataset. However the results for ISSEMyM suggest that the data (or population) is of a distinct nature. In particular, the first cluster is patients with LoS up to 11 days, referred as patients with "Short/Medium LoS" and the second cluster is patients with LoS from 12 days, referred as patients with "Long LoS".

Therefore, for further analysis, when studying the hospitals as separate datasets, the variable LoS category for the ISSEMyM hospital will be defined as a nominal variable with two categories: Short/Medium (patients with LoS up to 11 days) and Long (patients with LoS more than 12 days). In contrast, the definition of the categories for the MRC hospital will be: Short (patients with LoS up to 3 days) and Medium/Long (patients with LoS more than 4 days).

4.2. Clustering Diagnoses and Surgical Procedures

The variables first diagnosis, diagnosis and surgical procedure, described in Section 3.3, contains around 330, 850 and 200 different codes respectively, complicating the inclusion of these variables for further statistical analysis, e.g. logistic regression. To reduce the number of categories per variable different approaches were applied: First, the most natural approach was to group clinically similar codes or diagnoses (e.g. hernia hiatus, hernia umbilical and other hernias were grouped in a single category "Hernia"). This attempt significantly reduced the number of categories; however there were still more than 100 categories for each variable.

Another option is to take the five most common (or frequent) codes and make each of them a category while grouping the rest into one single category named "others"; leading to one nominal variable with 6 categories. The main drawback of this approach was that the resulting categories had very unequal sizes: the category "others" contained more than 50% of the codes.

In later steps of the analysis, this could cause problems during the variable selection and other data mining techniques.

Yet, another approach is to cluster both diagnosis and surgical procedure into categories with similar length of stay. The hierarchical cluster method is a simple and accessible technique to optimise the assignment of objects into a certain number of clusters and it is especially recommended when there are less than a few hundred objects to cluster and not all variables are quantitative. A hierarchical cluster algorithm produces a dendrogram representing graphically step by step the clustering of objects into groups, and groups into larger groups; then it can be broken at different levels to yield different clusterings of the data (Jain et al., 1999)

The clustering starts with calculation of the proximity matrix which is the basis of the hierarchical cluster analysis methods. This matrix has zeros on the diagonal and the values off the diagonal express dissimilarities between the corresponding pairs of objects, variables or categories. Dissimilarities measure the discrepancy between two objects based on several features including the type of variables, data and aims. Rezanková (2005) suggested the use of the "chi-squared dissimilarity measure" when the aim is to cluster categories (i.e. ICD codes) within a variable (i.e. "first diagnosis", "diagnosis" and "surgical procedure"). For the determination of dissimilarity between two ICD codes, a 2 x *K* contingency table is considered, where *K* is equal to 2, corresponding to the number of categories of the newly defined variable LoS category (the column variable): Short and Medium/ Long length of stay. The contingency table for two very common codes E119 (Non-insulin-dependent diabetes mellitus without complications) and K811 (Chronic cholecystitis) is given in Table 4.9:

	Short LoS	**Medium/Long LoS**	**Total**
E119	196	358	554
K811	693	338	1031
Total	889	696	1585

Table 4.9: Contingency table for variables LoS category and ICD 10 codes

The chi-square dissimilarity measure, (D_{CS}), is understood as Equation 4.4 states:

$$D_{CS}(v_{ki}, v_{kj}) = \sqrt{\chi^2} \tag{4.6}$$

where v_{ki} and v_{kj} are the categories i and j of the k^{th} variable and χ^2 is the chi-squared test.Table 4.10 contains the partial proximity matrix displaying the chi-square dissimilarity measure for the two ICD codes:

	E119	**K811**
E119	0	12.17
K811	12.17	0

Table 4.10: Partial proximity matrix for categories of variable Diagnosis

One of the main assumptions of the chi-square test is that, even for large contingency tables, expected frequencies should be greater than five (Field, 2009). However there are ICD codes which appear just one time in their contingency table with LoS category (e.g. in both hospitals during 2005-2009 there was just a single case with malignant neoplasm in the cerebrum). In order to meet the assumption, all those codes with frequencies lower than 5 were grouped in a subcategory: "Non-common diseases (or surgical procedures)".

Let us note that in the ISSEMyM hospital records, the variable "first diagnosis" as part of their regular practice. "First diagnosis" is understood as the health problem or disease diagnosed during the first medical evaluation at hospital. Since this evaluation could take place at the outpatient clinic or A&E, it is assumed that the "first diagnosis" variable is of a different nature from the main diagnosis (i.e. some "first diagnosis" codes are just present in this variable) and therefore a separate proximity matrix for this variable was calculated.

The full proximity matrices for "first diagnosis", "diagnosis" and "surgical procedure" were analysed in SPSS using six different clustering algorithms: single linkage, complete linkage, Ward linkage, average linkage (between groups), centroid linkage and median linkage.

Figure 4.10 depicts the dendrograms generated by each one of the six algorithms for surgical procedures. All of them show significant differences between fusion levels with 2 and 3 clusters. The arrangement of clusters generated by the Ward linkage algorithm (Everitt et al., 2001) was selected as the preferred choice, because it generated three well-defined clusters of relatively equal size (compared with those generated by other algorithms) from early stages of the fusion process. For the full size dendrograms, the reader is referred to the Appendix B.

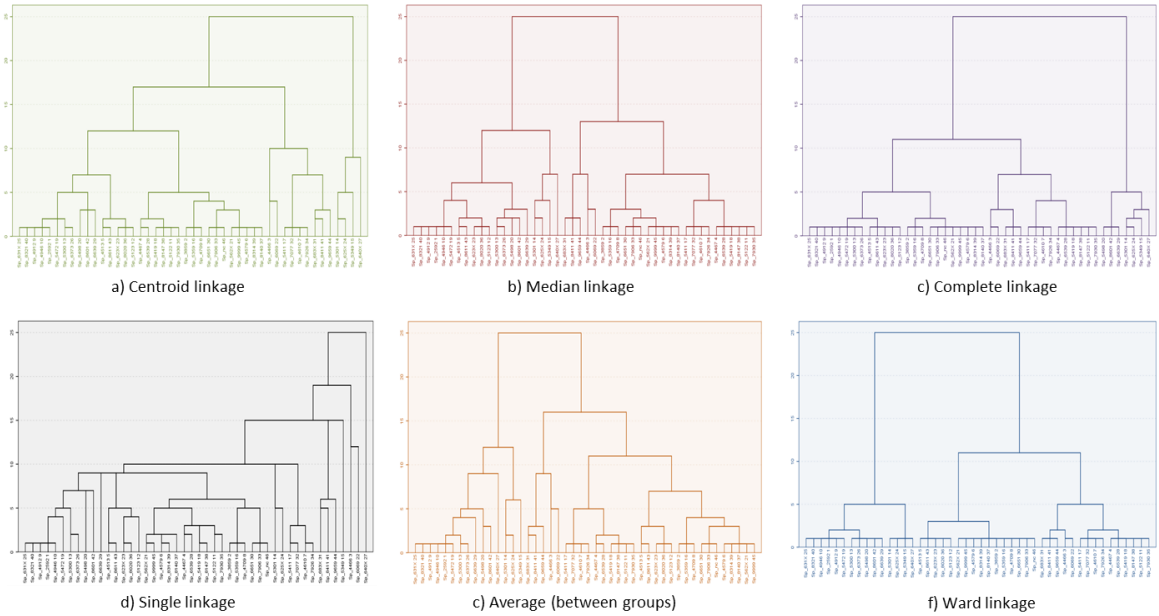

Figure 4.10: Dendrograms using different clustering algorithms for surgical procedures

For the case of the diagnoses codes, .

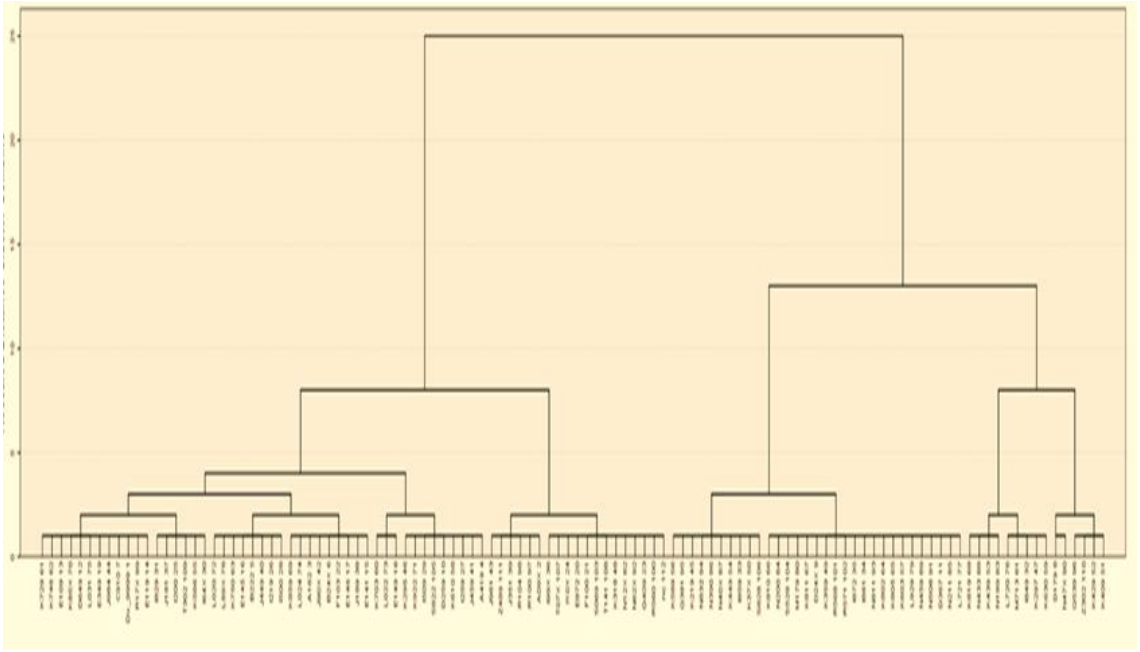

Figure 4.11 shows the preferred dendrogram with three well-defined clusters generated by the complete linkage algorithm. Finally, the dendrogram depicted in Figure 4.12shows significant differences between fusion levels with three clusters, using the Ward algorithm for the "first diagnosis" code. For the full size dendrograms, the reader is referred to Appendix B.

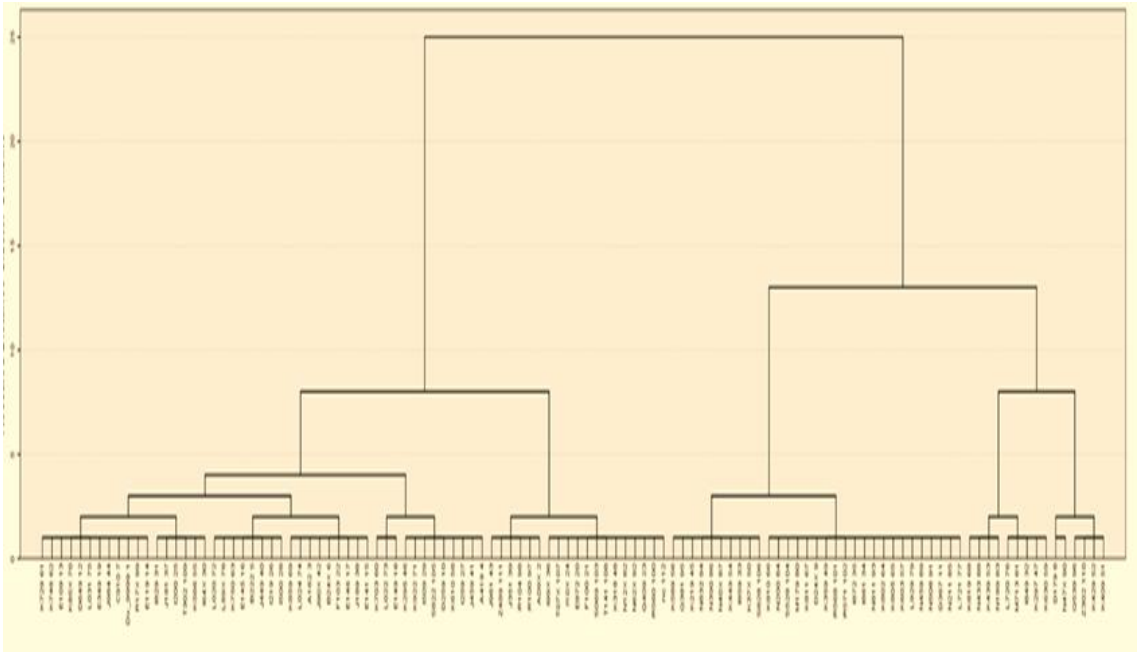

Figure 4.11: Dendrogram generated by complete linkage algorithm for diagnosis

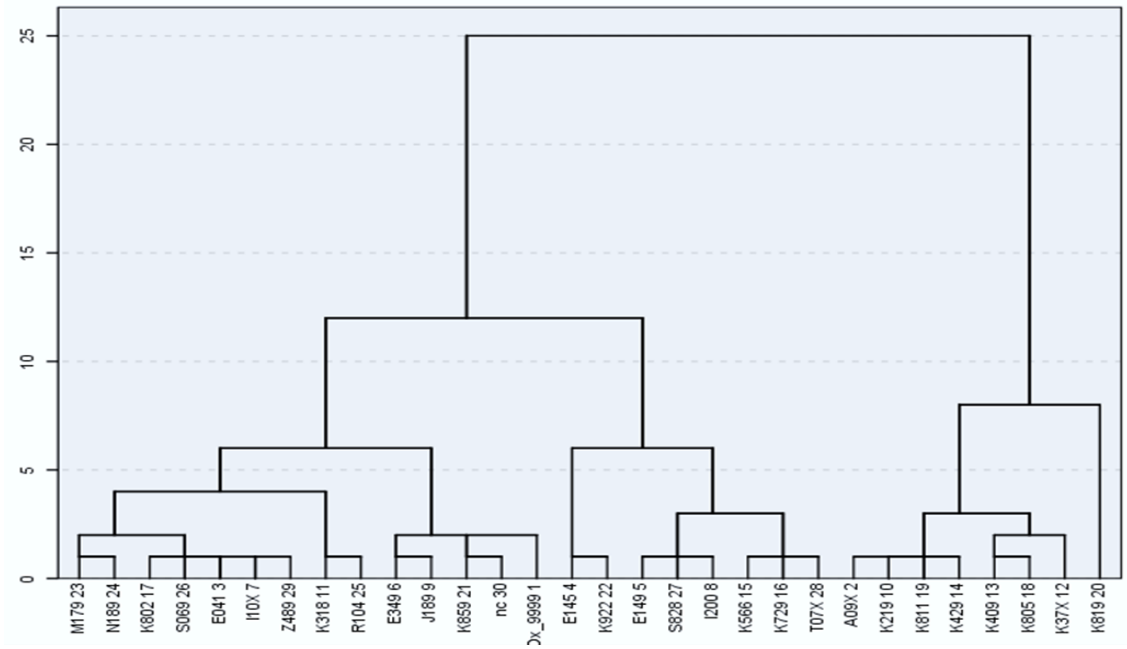

Figure 4.12: Dendrogram generated by Ward algorithm for first Diagnosis

As a result, three new variables were created: "First diagnosis" (For the ISSEMyM dataset only), "Diagnosis" and "Surgical procedure", with three categories each. Figure 4.13-Figure 4.15 show the most common diagnoses and surgical procedures per category as a result of the cluster analysis. The reader is referred to Appendix C for a full list of the medical conditions and surgical procedures included in each cluster.

In addition, one extra category was added to each one of the newly created variables to account for missing values (i.e. except for diagnosis at MRC, which does not contain missing values). The obvious reason why the variable surgical procedure contains missing values is because not all the patients undergo surgical procedures during their stay. However in the case of "first diagnosis" and "diagnosis" at the ISSEMyM data set, the presence of missing values is due to the doctor's personal choice or habit of completing just one field (i.e. either first or second diagnosis).

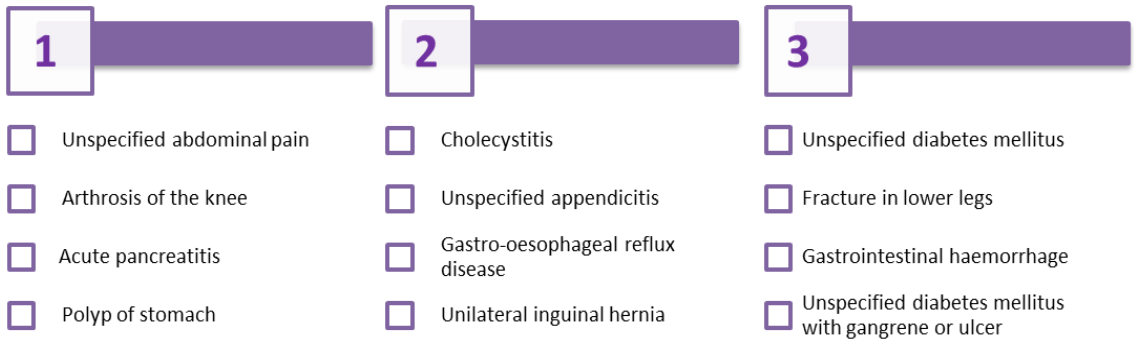

Figure 4.13: Most common first diagnoses per category or cluster

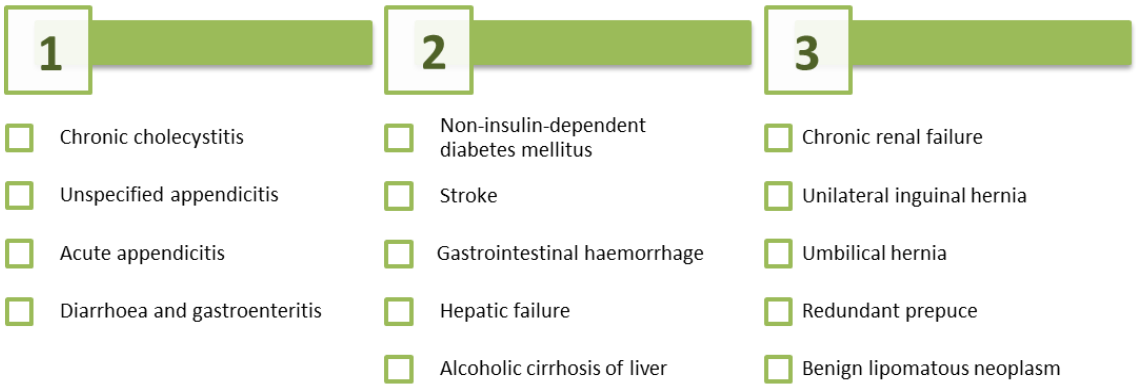

Figure 4.14: Most common diagnoses per category or cluster

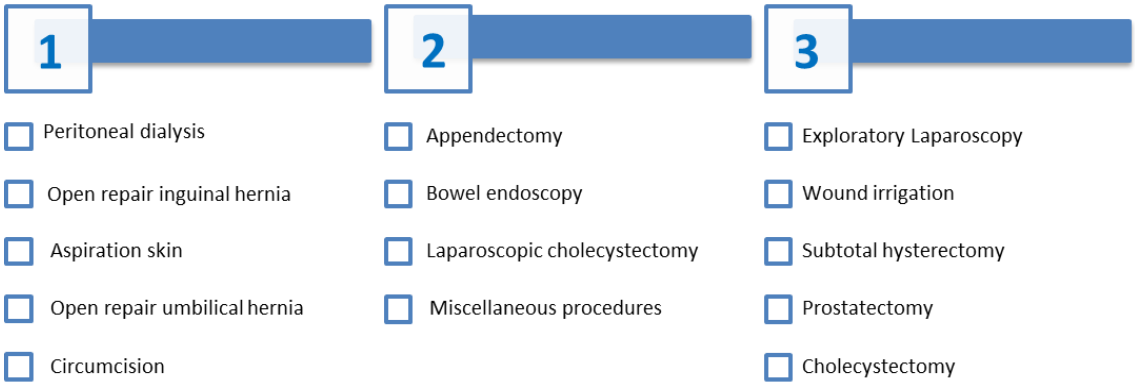

Figure 4.15: Most common surgical procedures per category or cluster

4.3. Summary

A formal approach to find a probabilistic model for LoS was performed using model-based cluster analysis with finite mixture models. A two-component Lognormal mixture model appeared to be most appropriate for describing the length of stay, yielding the creation of a new variable named LoS category with two categories: Short (patients with LoS up to 2 days) and Medium/Long (patients with LoS more than 3 days).

The same approach was used on the data by hospital: a two-component Lognormal mixture model, of very similar nature to the previous model, was the most appropriate choice to describe LoS at the MRC hospital. However a two-component Gamma mixture model was the preferred option for the ISSEMyM hospital. These results yielded a redefinition of the categories in the variable LoS category: Short/Medium (patients with LoS up to 11 days) and Long (patients with LoS more than 12 days) for ISSEMyM hospital, and Short (patients with LoS up to 3 days) and Medium/Long (patients with LoS more than 4 days) for MRC hospital.

Furthermore a methodology was employed to reduce the number of ICD codes of the variables "first diagnosis", "diagnosis" and "surgical procedure" using hierarchical cluster methods based on the chi-square dissimilarity measure. More than 800 diagnosis codes were grouped into three clusters using complete linkage algorithm. In addition using Ward's algorithm, more than 300 first diagnosis codes and 200 surgical procedures codes were grouped into three clusters.

5 INDIVIDUAL-BASED APPROACH

The main objective of this chapter is to explore the internal factors associated with LoS. Therefore the chapter is divided into main parts: the first part (Section5.1) aims to reduce the number of variables that both datasets contain in order to improve performance and simplicity of further models. The second part (Section 5.2) explores the individual-based approach described in the methodology chapter, where the variables previously selected are incorporated to the finite mixture model developed in the previous chapter.

5.1. Variable Selection

After complementing the datasets with the inclusion of new variables generated by hierarchical cluster methods, it is desirable to conduct a variable selection process. The aim is to look for a small number of variables that adequately represent the original set. The variable selection process is designed to improve prediction performance, providing a more parsimonious model, and giving a better understanding of the underlying process that generates the data.

The best way to select a variable is manually, based on expertise and knowledge. Unfortunately, there is not enough evidence in previous research about which factors might influence LoS when the full case-mix of a hospital is considered. Most of the previous models to predict LoS consider the attributes that exclusively relate to the medical conditions or cohorts of patients under study. However, some attributes such as patient age, medical condition and comorbidities seems to emerge as LoS predictors in most of the studies.

Therefore, without enough evidence from the literature to select the influential variables, quantitative methods such as statistical methods, machine learning techniques and data mining techniques are considered as options to help in the selection process

5.1.1. Multiple Regression

One of the most common statistical techniques to determine which variables are important to explain the variability of the output variable is multiple regression using stepwise regression.

In context of stepwise regression, the backward method is usually preferable to the forward method due to the suppressor effects, which occur when a predictor has a significant effect, but only when another variable is held constant (Field, 2009). Forward selection is more likely to exclude predictors involved in suppressor effects than backward selection. In addition, the forward selection runs a higher risk of missing a predictor that does in fact predict the outcome.

Multiple regression analysis was performed in STATA using the `regress` command. The software automatically breaks down all the categorical variables into dummy variables, each coded 0 or 1. For example, the variable "first diagnosis" can take four values: "no first diagnosis", "first diagnosis category 1", "first diagnosis category 2" and "first diagnosis category 3". Thus each category can be re-expressed as an independent binary variable. Besides one of these binary variables is used as the base category or reference group, and later the parameter estimates (i.e. beta coefficients) of the remaining categories are interpreted relative to that reference group. STATA automatically uses the first category of the predictor as the base category (See Appendix D for a full list and description of the dummies variables).

Table 2.2 show the results of the analysis for each hospital. Notice that multiple regression works under the assumption of normality of the outcome and therefore the transformed LoS (defined in Section 3.4) was used as the dependent variable.

	F statistic	**p-value**	R^2	**Adjusted R^2**	**Root MSE**
MRC	367.16	<0.00001	0.2175	0.21690	0.88564
ISSEMYN	11.33	<0.00001	0.1434	0.1307	0.77717

Table 5.1: Multiple regression output using stepwise method

The value of R^2 = .2175 in Table 5.1, indicates that significant variables account for 21.75% of the variance in LoS in the MRC hospital, which is a weak-moderate relationship. The adjusted value for R^2 is .2169; this small shrinkage means that if the model was derived from the population rather than from a sample it would account for approximately 0.06% (21.75%-21.69%) less variance in the LoS. Also, the similarity of the adjusted value to the observed

R^2indicates that the cross validity of the model is very good and that the model is not over fitted. The $F(10, 13206) = 367.16$ ($p<.00001$) indicates that using the model with the significant variables improved the ability to predict LoS rather than using the mean as "best guess".

The results for ISSEMyM can be interpreted in the same way. The $F(19, 1286) = 11.33$ ($p<.00001$) indicates that using the model with the significant variables improved the ability to predict LoS rather than using the mean.

Table 5.2 and Table 5.3show the significant variables coefficients (*p<0.01*).In the case of the MRC hospital (Table 5.2) the only variable that did not result as significant was the number of surgical procedures. Conversely, for ISSEMyM hospital, it seems that the socio-demographic variables such as occupation and educational level are not significant to predict LoS. Furthermore the inherited family history variables (diabetes, hypertension, etc.) were excluded from the model by the backward method due to their *p-values* greater than 0.01. On the other hand, those variables that describe the actual situation of the patient at admission (i.e. age, surgical procedure, origin, ward, diagnosis and number of current illnesses) and the variables that describe the personal pathologic and non-pathologic history of the patient (except from allergies and exposure to polluters) are the ones that explain better the variability of LoS.

	Coef.	**Std. Err.**	**t**	**P>t**	**[95% Conf.**	**Interval]**
Age	0.0042	0.0004	11.120	0.000	0.0035	0.0050
Gender(female)	0.0322	0.0138	2.340	0.020	0.0052	0.0592
Previous adm.	-0.0071	0.0006	-11.480	0.000	-0.0083	-0.0058
Outpatient clinic	-0.3266	0.0249	-13.120	0.000	-0.3753	-0.2778
General surgery ward	-0.1439	0.0207	-6.960	0.000	-0.1845	-0.1034
Diagnosis_category2	-0.2657	0.0218	-12.160	0.000	-0.3085	-0.2229
Diagnosis_category3	-0.6254	0.0208	-30.110	0.000	-0.6661	-0.5847
Sp_category 1	0.0491	0.0243	2.020	0.043	0.0014	0.0968
Sp_category 2	-0.1056	0.0204	-5.190	0.000	-0.1454	-0.0657
Sp_category 3	0.3577	0.0361	9.920	0.000	0.2870	0.4284
_cons	1.2558	0.0271	46.420	0.000	1.2027	1.3088

Table 5.2: Unstandardized β coefficients for the MRC regression model

	β	*Std. Err.*	*t*	*P>t*	*[95% Conf. Interval]*	
Age	0.0036	0.0015	2.400	0.017	0.0006	0.0065
Num comobidities	0.1246	0.0494	2.520	0.012	0.0278	0.2215
1stDiagnosis_category1	0.0835	0.0767	1.090	0.277	-0.0670	0.2340
1stDiagnosis_category2	-0.2011	0.1044	-1.930	0.054	-0.4060	0.0037
1stDiagnosis_category3	-0.0231	0.1085	-0.210	0.831	-0.2360	0.1897
Num previous sp	-0.0527	0.0273	-1.930	0.054	-0.1062	0.0009
Previous admissions	-0.0160	0.0082	-1.960	0.050	-0.0321	0.0000
Outpatient clinic	-0.1061	0.0649	-1.630	0.103	-0.2335	0.0213
Other origin	-0.3644	0.0624	-5.840	0.000	-0.4867	-0.2420
General surgery ward	-0.2749	0.0808	-3.400	0.001	-0.4334	-0.1165
Trauma ward	-0.4048	0.1030	-3.930	0.000	-0.6068	-0.2028
Diagnosis_category1	-0.1894	0.1187	-1.600	0.111	-0.4222	0.0435
Diagnosis_category2	0.2040	0.1324	1.540	0.124	-0.0557	0.4637
Diagnosis_category3	-0.3646	0.1402	-2.600	0.009	-0.6396	-0.0897
Num diagnoses	0.1554	0.0516	3.010	0.003	0.0542	0.2567
Transfusions	-0.1890	0.0876	-2.160	0.031	-0.3608	-0.0172
Sp_category 1	0.2796	0.1543	1.810	0.070	-0.0231	0.5824
Sp_category 2	0.3326	0.0819	4.060	0.000	0.1720	0.4932
Sp_category 3	0.4608	0.1002	4.600	0.000	0.2643	0.6573
_cons	1.2675	0.1517	8.350	0.000	0.9699	1.5652

Table 5.3: Unstandardized β coefficients for the ISSEMyM regression model

Although it is not the aim of this research, it is important to mention that because the dependent variable was transformed, the interpretation for the β coefficients has changed from the traditional interpretation. A unit increase in the predictor variable is now associated with an approximate 100β percent increase in the outcome variable. This approximation works well for $|\beta| < 0.1$, otherwise, the exact relationship is that: a unit increase in the predictor is associated with an average increase of $100(e^{\beta}\text{-}1)$ per cent (Vittinghoff, 2004). This interpretation is true only if the effects of the other 10 variables are held constant. The following examples for the ISSEMyM hospital illustrate this approximation better:

Number of comorbidities (β =.1246): The coefficient indicates that as the number of comorbidities increases by one unit, patient LoS increases 13.17% in respect to the mean[12] (3.43 days).

Surgical procedure category 3 (β=.4608): It is expected that patients undergoing a surgical procedure classified under category 3 will stay 2.12 days more than a patient undergoing a surgical procedure under any other category. This means an increment of 58.6% in respect to the mean.

One of the assumptions of multiple regression analysis is the nonexistence of perfect multicollinearity. Multicollinearity exists when there is a strong correlation ($r > 0.80$) between two or more independent variables. If multicollinearity is found, then one of the variables should be removed from the analysis or replaced by another equally important variable which is not strongly correlated (Field, 2009).

From the correlation matrix including the interval, ordinal and binary variables of both datasets, no multicollinearity was found[13]. However more subtle forms of multicollinearity could be present in the data. STATA produces various collinearity diagnostics, one of which is the variance inflation factor (VIF). According to Field (2009) if the value of VIF is greater than 10 or the tolerance statistics $(1/_{VIF})$ are below to 0.20 indicate a potential problem, and if the average VIF is substantially greater than 1 then the regression may be biased.

Table 12 shows the Variance inflation analysis for both datasets:

[12]Strictly speaking it is the geometric mean of the outcome variable. Regression of the log transformed outcome variable is used to estimate the expected geometric mean of the original variable (UCLA: Academic Technology Services, 2007)

[13]Bivariate two-tailed correlation analysis was carried out to measure linear relationship between log-transformed LoS and the dependent variables of the both data sets. The Pearson's correlation coefficient *r* was used for measuring correlation in ordinal variables whereas Spearman's correlation coefficient *rs* was used for interval and binary variables

ISSEMYN

Variable	VIF	1/VIF
Diagnosis_category1	5.26	0.190129
Diagnosis_category2	4.71	0.212189
Diagnosis_category3	3.26	0.307032
General surgery ward	2.69	0.371407
Sp_category1	2.52	0.396076
Sp_category3	2.1	0.475852
Trauma ward	1.85	0.541024
Sp_category2	1.58	0.634164
Other origin	1.54	0.648438
Outpatient clinic	1.39	0.718774
Num diagnoses	1.32	0.757837
Previous admissions	1.14	0.87926
Num comobidities	1.09	0.916655
Age	1.07	0.93692
Num previous sp	1.04	0.958136
Transfusions	1.03	0.971476
Mean VIF	2.1	

MRC

Variable	VIF	1/VIF
Diagnosis_category3	2.28	0.438671
General surgery ward	2.24	0.445676
Diagnosis_category2	2.2	0.454854
Sp_category1	1.53	0.654663
Sp_category2	1.45	0.690976
Previous adm.	1.39	0.719511
Outpatient clinic	1.38	0.72328
Age	1.24	0.805163
Sp_category3	1.14	0.87468
Gender	1.02	0.976698
Mean VIF	1.59	

Table 5.4:Variance inflation analysis

From the results in Table 5.4 it can be read clearly that none of the variable VIF values are higher than 10 and tolerance statistics stay above 0.2; however the average VIFs for both models are not close enough to 1, which indicates that the regression models might be biased by multicollinearity.

In order to test the remaining assumptions of the multiple regression models, an analysis of the residuals was carried out: subfigures a) and b), in Figure 5.1, depict the P-P plots of normally distributed standard residuals (from MRC and ISSEMyM datasets respectively). In both plots the points follow fairly the straight lines, indicating that the assumption of normality of the residuals is not violated (Nist/Sematech, 2003). However, in subfigures c) and d) points seem to have similar patterns and being more spread out at the right side of the graphs. This could indicate violations of both the homogeneity of variance and linearity assumptions.

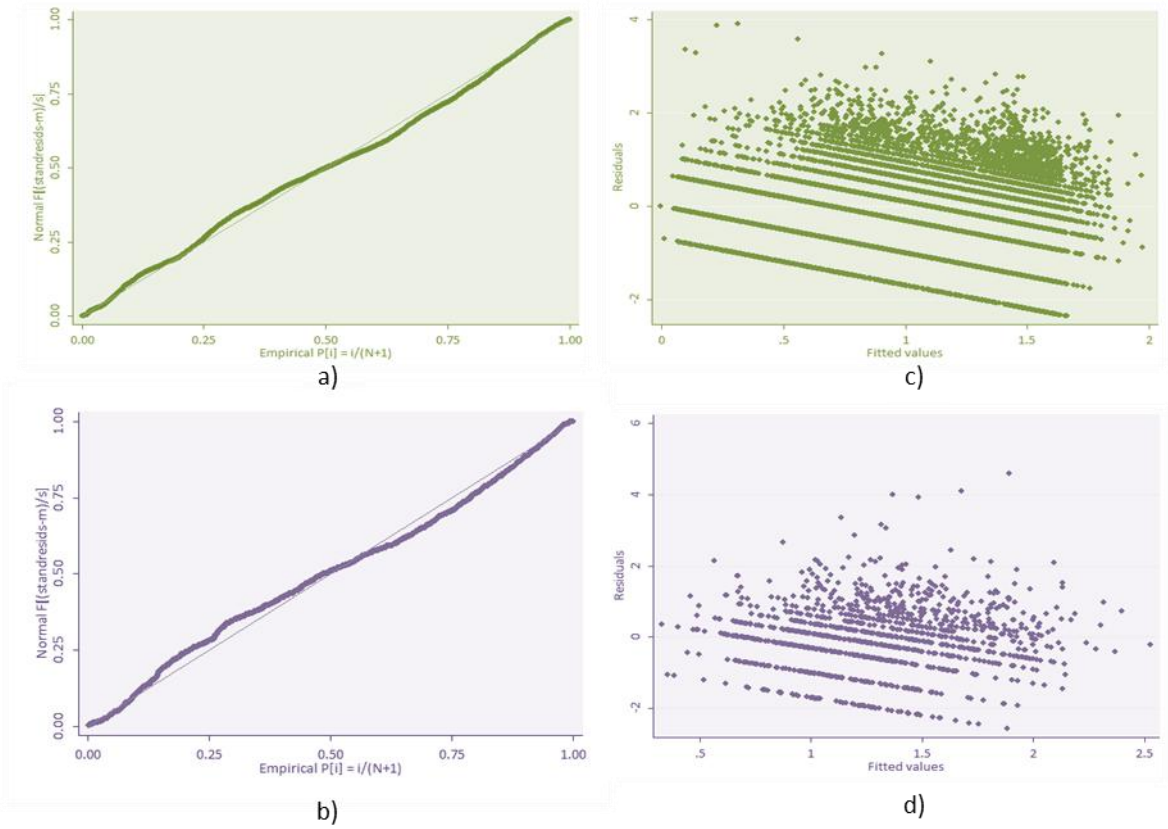

Figure 5.1: P-P plots of normally distributed standard residuals and plots of standardized residuals against standardized predicted values. Subfigures a) and c) correspond to the MRC hospital and subfigures b) and d) correspond to the ISSEMyM hospital

From the previous analysis, it can be concluded that there is some evidence to believe that the assumptions of multicollinearity, homogeneity of the variance and linearity are broken. However, since the final goal of the current multiple regression models is not a prediction of the outcome but a selection of the variables which are significant to LoS, there is no need to be overly concerned about the assumptions of the model.

5.1.2. Bootstrapping

According to Chernick (1999), when there is a large number of variables, the search for the optimal subset might be abstruse. That is why suboptimal selection procedures such as forward, backward and stepwise selection may lead to different results in the same dataset (i.e. different sets of variables may work equally well).

When conducting a variable selection process, it is a frequent practice to mark the selected variables as being useful and the discarded variables as not being useful. However, discarding variables may lead to a loss of valuable information as Gong (1986) demonstrated. Researchers from the Stanford University School of Medicine used logistic regression to predict a patient's chance of survival from acute and chronic hepatitis. The model selected four variables out of 19

as significant predictors. Later, Gong used bootstrapping to validate the results. He generated 500 bootstrap replications of the data and applied the logistic regression on bootstrap replications. In some replications, only one predictor variable emerged. None of the variables were significant in more than the 60% of the replications. This example highlights the importance of not overestimating the results of variable selection procedures and it demonstrates the potential of bootstrapping for assessing the effects of subset selection.

In order to validate the results of the variable selection process from the previous section, a similar methodology was followed: multiple regression was executed 100 times, bootstrapping the β coefficients and standard errors by resampling observations (with replacement) from both datasets. This method is commonly referred to as the non-parametric bootstrap. Then the β coefficients and standard errors (s.e) were used to calculate the *t-statistic* for each one of the independent variables for the 100 replications, where $t = \frac{\beta}{s.e}$. The *t-statistic* tests the null hypothesis that the value of the β coefficients is zero: therefore if it is significant ($p<0.1$) the null hypothesis can be rejected, meaning that the independent variable contributes significantly to predict the outcome variable LoS. The decision to execute just 100 bootstrap replications, contrary to what Gong did previously, was inspired by the work of Efron (1987), who argues that 25 replications gives reasonable results and more than 100 replications do not improve the coefficient of variation for standard errors.

Figure 5.2 shows the percentage of replications where a variable happened to be significant. Just those which were significant in more than the 50% of the replications will be considered for further analysis. For the MRC dataset, the results show little variability. On average 7 out of 8 variables resulted significant in every replication: patient age, number of previous hospital admissions, origin of the patient, ward where the patient is treated, diagnosis, surgical procedure and patient gender were significant in explaining variance in LoS. Conversely, for the ISSEMyM dataset, origin of the patient, surgical procedure, total number of current illnesses, treatment ward, patient age, previous blood transfusions, number of comorbidities drinking/smoking status, diagnosis, first diagnosis and number of previous hospital admissions were significant to explain the variance in LoS; other variables such as demographic variables (i.e. occupation, education, etc.) and inherited family history variables were discarded. However, in some replications, just 5out of 22 variables resulted as significant whereas in others up to 13were significant. In average 10 out of 22 variables resulted as significant in every replication. This indicates a high variation of the results and supports the use of bootstrapping as a useful technique to validate the results of a variable selection process.

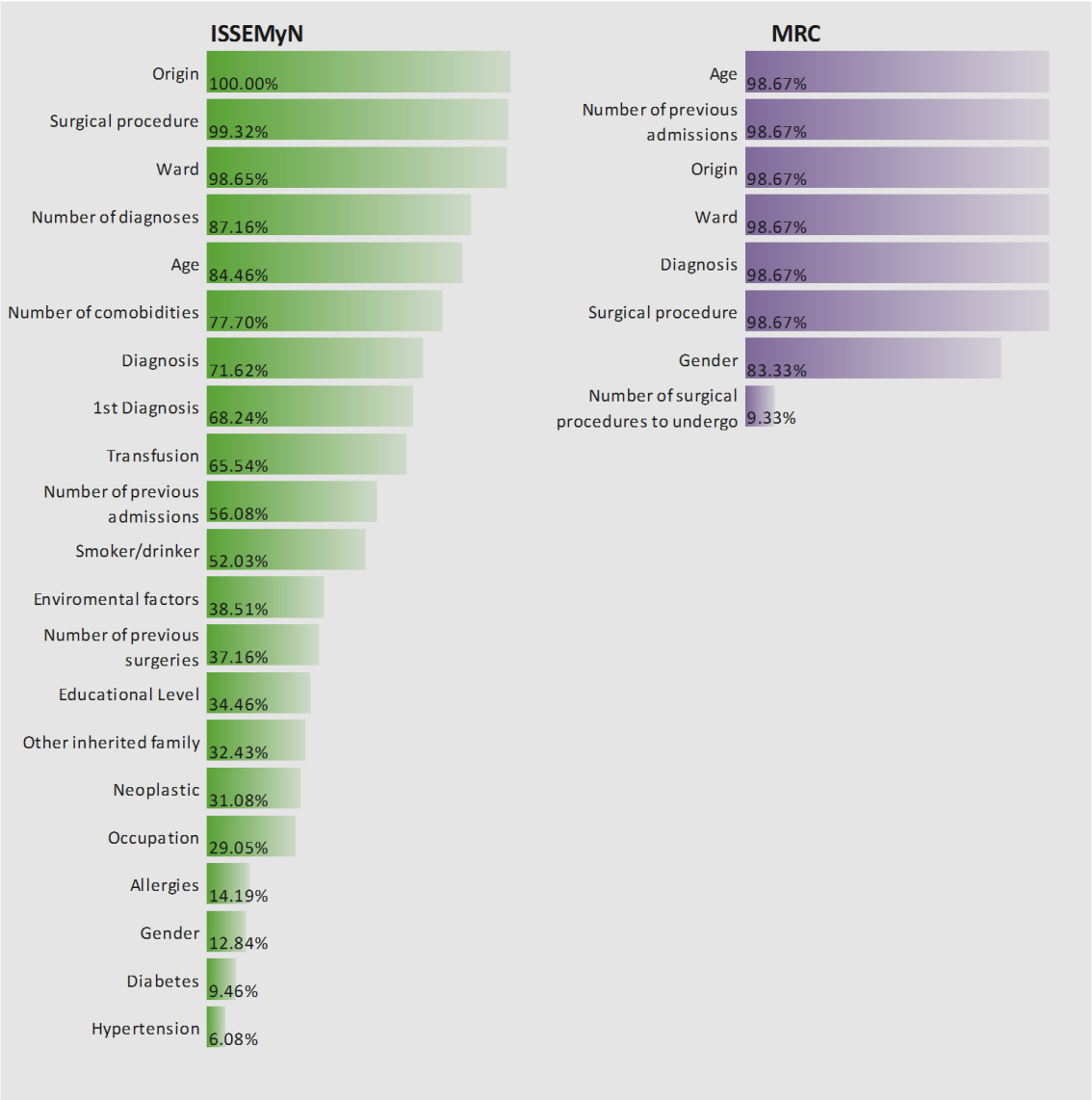

Figure 5.2: Percentage of replications where the variables were significant in explaining the variance in LoS

5.2. Finite Mixture of Generalised Linear Models

According to the results on the previous chapter, LoS data is distributed according to a two components finite model. In this context, a new variable "LoS category" was created and defined as a binary variable with two categories, which corresponded to the two components of the finite mixture model: Short (patients with LoS up to 3 days) and Medium/Long (patients with LoS more than 4 days) for the case of MRC; and Short/Medium (patients with LoS up to 11 days) and Long (patients with LoS more than 12 days) for the case of the ISSEMyM hospital.

In the individual-based approach, the variables selected in the previous section are used to predict the LoS category to which the patient most likely belongs and to shape their LoS

probabilistic curve. In addition, the size and direction of the effects that such variables hold on LoS distribution are explored.

Previously in this chapter, the underlying relationship between LoS and other variables was superficially explored through a linear regression model, which was used as a tool to select out the variables which are significant to explain the variance in LoS. However there was enough evidence to discourage the use of the linear regression model as a formal prediction tool for LoS, due to the violation of the assumptions in which the model is based: non-multicollinearity, homogeneity of the variance and linearity.

Nevertheless, although the linear model is not appropriate for the nature of the LoS data, it is still possible to link it to the two components finite mixture model, using a relative novel approach defined as finite mixture of generalised linear models.

In the finite mixture of generalised linear models, the mean of component *s* from the mixture model is associated with the linear regression model via a canonical link function (see Equations 5.1-5.3):

$$g(\mu) = \vartheta \tag{5.1}$$

$$\mu = g^{-1}(\vartheta) \tag{5.2}$$

$$\vartheta = \boldsymbol{\beta X} \tag{5.3}$$

where μ is the mean, $g(.)$is the canonical link function, $g^{-1}(.)$ is the inverse link function and ϑ is the linear regression model with the explanatory variables affecting the observed outcome. The link function is a transformation of the mean of the dependent variable such that this transformed variable is a linear function of the regression parameters. Dunteman and Ho (2005) stated that one can think of $g(\mu)$ as *"tricking"* the linear regression model into thinking that it is still acting upon normally distributed outcome variables.

In the same context the inverse of the link function $g^{-1}(\vartheta)$ ensures that the regression model ϑ maintains the assumptions for linear models and all the standard theory applies even though the dependent variable takes on a variety of non-normal forms.

Equation 5.1 and 5.2 are the foundation of the well-known Generalised linear models (GLM) (Nelder and Wedderburn, 1972), which are regression models where the dependent variable is specified to be distributed according to one of the members of the exponential family. When it comes to the use of finite mixture models, this type of models is commonly referred as finite

mixture of generalised linear models (Dias, 2004) or mixture regression models (Hagenaars and McCutcheon, 2009).

Each distribution that is a member of the natural exponential family[14] has its own canonical function $g(\mu)$. Table 5.5 summarises the canonical link and the inverse link function for the two distributions used to describe the data of ISSEMyM and MRC.

Distribution	**Canonical Link:** $\boldsymbol{\vartheta = g(\mu)}$	**Inverse Link:** $\boldsymbol{\mu = g^{-1}(\vartheta)}$
Gaussian	μ	ϑ
Gamma[15]	$\log(\mu)$	$\exp(\vartheta)$

Table 5.5: Link functions for common distribution from the natural exponential family

Notice that the Lognormal model is replaced by a Gaussian distribution where $y_i^* = \ln y_i$. This is because the Lognormal distribution is not a member of the natural exponential family; thus to employ it along with GLM theory, a log-transformation need to be applied in the dependent variable. Therefore, the component *s* density for observation *i* is given by Equation 5.4:

$$f_s(y_i^*; \mu_{si,}\sigma_s^2) = \frac{1}{\sigma_s\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_s^2}(y_i^* - \mu_{si})^2\right) \tag{5.4}$$

where $\mu_{si} = \vartheta_{si} = \beta_{s0} + \beta_{s1}x_{i1} + \cdots + \beta_{sp,}x_{ip,}$

The component *s* density for the ISSEMyM hospital in the gamma mixture is given by Equation 5.5[16]:

$$f_s(y_i; \mu_{si}, \sigma_s^2) = \frac{\left({}^{\mu_{si}}/_{\sigma_s^2}\right)^{\frac{\mu_{si}}{\sigma_s^2}} y_i^{\left(\frac{\mu_{si}}{\sigma_s^2}-1\right)} \exp\left(-\frac{\mu_{si}}{\sigma_s^2}\right)}{\Gamma\left({}^{\mu_{si}}/_{\sigma_s^2}\right)} \tag{5.5}$$

where $\mu_{si} = \exp(\vartheta_{si}) = \exp(\beta_{s0} + \beta_{s1}x_{i1} + \cdots + \beta_{sp,}x_{ip})$ and Γ is the gamma function[17].

[14] The natural exponential family (NEF) is a subclass of the exponential family such as Gaussian, Poisson, Gamma, etc. The interested reader is referred to Jørgensen (1997)

[15] There are other common choices for the link function (i.e. inverse link, log link, linear link). However the log link used here is recommended when the effect of the predictors is suspected to be multiplicative of the mean (Faraway, 2006).

[16] The notation used for the gamma mixture is changed from the one used in the previous section, for a more convenient form for the purposes of glm theory, where $\mu = \alpha\beta$, $\sigma^2 = \alpha\beta^2$. However the results will be reported with the original notation α for the shape parameter and β for the scale parameter.

The estimation of the generalised linear model was carried out in the same way that the finite mixture models were fitted in Section 4.1: using the STATA command `fmm` and adding the variables (i.e. covariates) selected during the variable selection process in Section 5.1.

On the other hand, the patient membership of one of the components (LoS category) was determined using posterior probabilities (see Equation 4.3 in Section4.1.1). The estimation of the Lognormal regression mixture model was performed using 2/3 of the dataset (named training set) and the remaining (named validation set) was used for testing purposes (Dobbin and Simon, 2011). Conversely, the estimation of the Gamma regression mixture model was performed using 9/10 of the dataset and 1/10 for validation: this was due to small size of the sample.

It is important to mention that although GLM relaxes most of the assumptions of linear models, it still assumes statistical independence of the observations. However, in both datasets, there is a significant percentage of the patients that had repeated admissions to hospital during the period of time being considered here[18] (e.g. patients with renal failure requiring dialysis have in average 10 admissions per year). Moreover, one cannot ignore that the data obtained from a patient's admission to hospital is highly correlated with data collected in previous admissions. Therefore, the assumption of independence across observations cannot be guaranteed, and the specifications of the STATA command to estimate the mixture regression model had to be modified to account for possible correlation between observations within the same patient file number. This is formally known in statistics as robust estimation, where the standard errors allow for intragroup correlation, relaxing the usual requirement of observations being independent. In other words, the observations are independent across groups (patient file number) but not necessarily within groups.

Table 5.6 summarises the parameters estimates for the mixture models. In the Lognormal model, the mixing proportions were 0.41 and 0.59, with a mean LoS per component of 2.2 and 5.06 days[19] respectively; whereas in the Gamma mixture model, two components were identified in proportions 0.92 and 0.08, with a mean LoS per component of 4.0 and 44.4 days[20] respectively.

Table 5.7 summarises the values of AIC and BIC before and after adding the independent variables (covariates) to the mixture model: both AIC and BIC values agree that the Lognormal mixture model with covariates is a better model to explain the LoS data at MRC. In contrast, for

[17] The gamma function is described by $\Gamma(\delta) = \int_0^\infty t^{\delta-1}e^{-t}dt,\ where\ t = {}^{y_i}/_{\beta}\ \ and\ \delta \in (0,\infty)$

[18] Every patient admission was considered as a single observation or data entry

[19] The mean for each component of the log-normal mixture model is equal to $e^{(\mu_s+\sigma_s^2/2)}$

[20] The mean for each component of the gamma mixture model is equal to $\alpha_s\beta_s$

the Gamma mixture model the AIC value indicates a better fit to the data after adding the covariates to the model; however the value of BIC indicates a better model in the one with the intercept only. This should be taken cautiously since a lower value of BIC does not necessarily imply a better fit but either fewer explanatory variables or both (i.e. fewer variables and better fit). It is recommended to take BIC value in conjunction with the AIC.

Gamma mixture (ISSEMyM)			**Lognormal mixture (MRC)**		
Parameter	**1st component**	**2nd component**	**Parameter**	**1st component**	**2nd component**
α	2.17	2.23	μ	0.46	1.50
β	1.85	19.90	σ	0.82	0.49
π	0.92	0.08	π	0.41	0.59

Table 5.6: Parameters estimates for the mixture regression models

		Model with intercept only	**Model with covariates**
Gamma mixture (ISSEMyM)	**AIC**	6857.40	6725.637
	BIC	6881.39	6941.494
Lognormal mixture (MRC)	**AIC**	59404.2	55196.19
	BIC	59440.08	55373.18

Table 5.7: Comparison of AIC and BIC values

Table 5.9 and Table 5.8 summarise the parameters estimates for the regression models[21]: for the MRC hospital, some variables predict membership in component 2: those patients who are older and have few previous admissions to hospital were significantly more likely to have a medium-long LoS (i.e. more than 3 days at hospital). On the other hand, for the ISSEMYN hospital, those who are older, have a diagnosis from category 2 (i.e. diabetes mellitus, stroke, hepatic failure, cirrhosis, gastrointestinal haemorrhage, etc.) or underwent a surgical procedure category 2 (i.e. appendectomy, bowel endoscopy, laparoscopic cholecystectomy, etc.) were more likely to be in component 2, with a long LoS (i.e. more than 12 days at hospital).

[21]Note that the results of the linear model are slightly different from those depicted in Section5.1.1. The regression model presented here includes the final selection of variables after bootstrapping.

	Linear Model		first component		2nd component	
X	β	Std. Error	$\beta.$	Std. Error	β	Std. Error
Age	0.0037*	0.0005	0.0035*	0.0009	0.0039*	0.0006
Gender (female)	0.0417*	0.0188	0.0748*	0.0336	0.0303	0.0238
Previous adm.	-0.0069*	0.0008	-0.0108*	0.0013	-0.0027*	0.0008
Outpatient clinic	-0.3283*	0.0240	-0.5801*	0.0507	-0.1164*	0.0330
General surgery ward	-0.1642*	0.0235	-0.1783*	0.0544	-0.1363*	0.0334
Diagnosis_category2	0.2624*	0.0241	0.2309*	0.0514	0.24*	0.0309
Diagnosis_category3	-0.3908*	0.0278	0.7032*	0.1635	-1.2348*	0.0391
Sp_category 1	-0.1021*	0.0285	-0.1244*	0.0493	-0.0553*	0.0321
Sp_category 2	-0.0216	0.0266	0.7247*	0.1043	-0.5624*	0.0349
Sp_category 3	0.3176*	0.0264	1.1745*	0.1122	-0.3797*	0.0412
cons	1.0395	0.0316	0.4566	0.1234	1.4998	0.0426

Table 5.8: Regression model and lognormal mixture model for MRC. For a full description of the variables the reader is referred to Appendix D. *indicates significant coefficients (i.e. $p \geq 0.5$).

The results of the linear model for the MRC hospital (Table 5.8) do not differ dramatically from the Lognormal mixture regression; with the exception of surgical procedures category 2, which is not significant in the linear model but it is statistically significant in both components of the Lognormal mixture regression.

To understand better the variables effects is easier to use either the exponentiated parameter estimates[22] ($i.e.$ $\exp(\beta)$), where the effect of the variables on the mean LoS is given by the exponential of their coefficients (Dunteman and Ho, 2005) or in terms of a percentage of the mean ($i.e.$ $100 * (\exp(\beta) - 1)$. For example, for patients in the first component, the multiplicative effect of surgical procedure category 2 is 2.06 ($\exp(0.724)$), indicating that the LoS for patients undergoing a surgical procedure under category 2 (i.e. appendectomy, laparoscopic cholecystectomy, endoscopy, etc.) is 106.4% ($100 * (\exp(0.724) - 1)$ higher

[22]The parameter estimates of the mixture regression model should not be interpreted in the way linear model parameters estimates are understood; because the relationship between the independent variable and outcome variable is expressed through the nonlinear link function.

than that for patients not undergoing surgical procedures. However, the multiplicative effect of surgical procedure category 2 in the second component is 0.56, indicating that the LoS for these patients is 43.2% shorter than that for patients not undergoing surgical procedures. The evidence points out that these type of procedures are not very common in the first component as it seems that most of the patients undergoing a surgical procedure of category 2 are associated with a medium-long LoS.

	Linear Model		**first component**		**2nd component**	
X	***β***	**Std. Error**	***β.***	**Std. Error**	***β***	**Std. Error**
Age	0.0034*	0.0015	0.0051*	0.0018	0.0243*	0.0096
1stDiagnosis_category1	0.0615	0.0769	0.0701	0.1001	-0.6770	0.4666
1stDiagnosis_category2	-0.2155*	0.1048	-0.2531*	0.1199	-1.512*	0.3237
1stDiagnosis_category3	-0.0285	0.1089	0.1146	0.1202	-3.7271*	1.1802
Previous adm.	-0.0157*	0.0083	-0.0095	0.0065	-0.1048*	0.0213
Outpatient clinic	-0.1043	0.0650	-0.1897*	0.0706	0.4382	0.3787
Other origin	-0.3734*	0.0624	-0.4258*	0.0684	-0.3536	0.2671
General surgery ward	-0.2978*	0.0817	-0.329*	0.0912	-1.2906*	0.6698
Trauma ward	-0.4182*	0.1033	-0.5088*	0.1149	0.2698	0.5348
Diagnosis_category1	-0.1737	0.1186	-0.1470	0.1274	-0.2071	0.5456
Diagnosis_category2	0.222*	0.1325	0.0773	0.1356	0.9495*	0.4758
Diagnosis_category3	-0.3597*	0.1404	-0.4572*	0.1493	0.1874	0.4732
Num diagnoses	0.156*	0.0517	0.1784*	0.0541	-0.557*	0.2685
Sp_category 1	0.2862*	0.1545	0.3658*	0.1741	0.7915	0.9531
Sp_category 2	0.3315*	0.0820	0.2924*	0.0851	1.1647*	0.5110
Sp_category 3	0.4507*	0.1005	0.5471*	0.1102	0.2301	0.6519
Transfusions	-0.2039*	0.0877	-0.2102*	0.0761	0.5044	0.3784
Num comobidities	0.1152*	0.0494	0.1198*	0.0571	0.0199	0.3268
Drinking/smoking	-0.0734	0.0681	-0.0570	0.0717	-0.1262	0.3411
Drinking&smoking	-0.1028*	0.0623	-0.0549	0.0639	-0.8811*	0.2770
cons	1.2920	0.1545	0.6173	0.1783	2.9906	0.7271

Table 5.9: Regression model and gamma mixture model for ISSEMyM. For a full description of the variables the reader is referred to Appendix D.

On the other hand, according to the results of the linear model in Table 5.9 for ISSEMyM, patient's age (which was statistically significant in the linear model), was significant in both components of the mixture regression. However this is not the case with other variables: the number of previous admissions to hospital had a small but significant effect on LoS, but in the Gamma mixture model, one can see that this small effect is just for patients who are members of the second component (medium-long LoS).

Other special case is the number of diagnoses: the linear model suggested a small but significant positive effect. However this positive effect is just valid for patients who belong to the first component: when it comes to patients in the second component, the number of diagnoses has a negative effect. The multiplicative effect of the number of diagnoses for the first component is 1.19 ($exp(0.178)$), indicating that the addition of one extra diagnosed condition to the initial diagnosis of the patient has the effect of increasing the patient LoS by 19% ($100 * (\exp(0.178) - 1)$ respect to mean of the first component[23]. On the contrary, if the patient belongs to the second component, the effect of number of diagnosis is 0.57, indicating that the addition of one extra diagnosis to the initial condition of the patient has the effect of decreasing the patient LoS by 42.71% respect to the component mean[23]. When more conditions are added to the patient initial diagnosis, the outlook of such a patient may be more complicated than usual and he or she may require medium term care rather than short term, although the majority of these cases do not require very long-term care.

For patients in the first component, the multiplicative effect of surgical procedure category 3 is 1.72, indicating that the LoS for patients undergoing a surgical procedure of this category (i.e. laparoscopy, subtotal hysterectomy, prostatectomy or cholecystectomy) is 72.83% higher than that for patients not undergoing surgical procedures[23]. However, this effect is not significant in patients belonging to component 2 (i.e. patients with longer LoS), indicating that surgical procedures under category 3 have influence just on patients with short-medium LoS rather than with long LoS.

Finally, the accuracy rates on the testing set were estimated to give an idea of how well the models are in predicting new patients into one of the LoS categories (i.e. defined as the components of finite mixture). The LoS category of a new patient (on the testing set) was determined using the highest posterior probability that a patient belongs to a component s (i.e. $max(\Pr[y_i \in components \mid y_i; \theta])$) and then compared it against the category to which the patient should belong according to its observed LoS value. The accuracy rates are displayed in Table 5.10. Both models perform exceptionally well in predicting membership in their largest component. However, they do not achieve such success in predicting patients belonging to their

[23]holding constant the rest of the variables

smaller component. The overall accuracy rate of the Gamma mixture model is not seriously affected since the proportion of patients in the second component is just 8%. However for the MRC hospital, the proportion of patients having a short-medium LoS (i.e. belonging to the first component) is 41%, affecting severely the overall accuracy rate of the Lognormal mixture model.

	Gamma mixture (ISSEMyM)	**Lognormal mixture (MRC)**
First component	99.0%	30.09%
Second component	36.6%	74.82%
Overall accuracy	92.8%	53.76%

Table 5.10: Accuracy rates 10 trials for mixture regression models.

These results indicate that the finite mixture regression have certain limitations in successfully discriminating between the two components or categories when it comes to classify patients. However, the advantages of such classification are the creation of homogenous groups of patients according to their LoS, and the identification and understanding of the factors and their effects that influence each group. Therefore, the researcher still recommends that one performs such classification to get a general idea of what type of LoS the patient is more likely to have (i.e. short or medium-long)[24] and the role that patient attributes play in such classification.

However when it comes to the estimation of the patient LoS distribution (or expected LoS), it is better to use the mixture regression model density equation $\left(i.e. f(y_i;\varphi) = \sum_{s=1}^{S} \pi_s f_s(y_i;\theta_s)\right)$, rather than estimating the density of the component s ($i.e. f_s(y_i;\theta_s)$) with the highest posterior probability. This minimises the risk of incorrect estimations of LoS, because the estimated LoS probabilistic curve (based on the mixture regression density equation) would contain an element from both categories (components) but in different proportions. Furthermore, any model should always account for uncertainty: even when a patient has been classified in certain LoS category, there is always a chance that he or she may have a longer or shorter LoS outside of the intervals that their LoS category considers (due to countless numbers of reasons).

5.3. Summary

Having completed the preparation for both datasets in Chapter 4, a formal methodology to select the significant variables for the LoS was applied using multiple stepwise regression with

[24] Short-medium or long for the ISSEMyN hospital

the backward method. The results for ISSEMyM indicate that 11 variables are significant in explaining 14% of the variance in LoS: origin of the patient, surgical procedure to undergo, total number of current illnesses, ward where it is treated, patient age, previous blood transfusions, number of comorbidities, drinking/smoking status, diagnosis, first diagnosis and number of previous hospital admissions. Conversely, 7 variables in the MRC dataset were significant in explaining 21% of the LoS variance: patient age, number of previous hospital admissions, origin of the patient, treatment ward, diagnosis, surgical procedure and patient gender.

In addition, bootstrapping was performed as an approach to validate the results of the variable selection process, by resampling each dataset and executing multiple regression 100 times. Although the results changed arbitrarily from one replication to other (in the case of the ISSEMyM dataset only), the general outcome coincided with the results derived from multiple regression models.

The final part of this chapter was devoted to the individual-based approach, where all patients were treated as different independent entities. Their individual characteristics predicted firstly the membership to one of the two LoS categories and secondly those same characteristics defined the shape of their LoS probabilistic curve and its associated expected LoS. In this context, the finite mixture model defined in the previous chapter was extended to accommodate the patient characteristics. This broader model is called finite mixture of generalised linear models, where the mean of each component of the finite mixture model is associated with a linear regression model via a canonical link function.

According to the AIC, the Gamma mixture model was a better fit to the data after adding the covariates to the model compared to the model with the intercept only. On the other hand, both AIC and BIC values agree that the Lognormal mixture model with covariates was a better model to explain the LoS data at MRC.

Moreover, for the ISSEMyM hospital, some variables predict membership to the category long LoS: those who are older, have a diagnosis from category 2 (i.e. diabetes mellitus, stroke, hepatic failure, cirrhosis, gastrointestinal haemorrhage, etc.) or underwent a surgical procedure category 2 (i.e. appendectomy, bowel endoscopy, laparoscopic cholecystectomy, etc.) were more likely to have a long LoS. On the other hand, for the MRC hospital, those patients who are older and have few previous admissions to hospital were significantly more likely to have a medium-long LoS.

Finally the accuracy rates when classifying patients into their correct LoS category indicated that both mixture regression models have limited ability to predict accurately membership of the smallest component of the mixture. Therefore, for the estimation of patient LoS distribution, it

is recommended to use the density function of the mixture, rather than the density function for each component.

6 GROUP-BASED APPROACH

The main objective of this chapter is to explore the internal factors associated with LoS from the group-based approach perspective (described in the methodology), where patient attributes or variables will be used to predict the LoS category to which the patient belongs. In this approach, all patients within LoS categories are the same. Although their individual characteristics help to predict the membership of LoS, their length of stay probabilistic curve and associated expected LoS is defined by the parameters of the category itself (see Section 4.1.3).

The group-based approach can be equivalent to Discrete Conditional Survival (DCS) models developed by Cairns and Marshall (2009), which have been used to model skewed distributions for healthcare outcomes. It can be broken down into two components: the conditional component and the process component.

The conditional component comprises a structure that captures the nature of the data by representing the various inter-relationships between variables, and thus can categorise the observations into a number of discrete classes (LoS category). The conditional component precedes the process component.

The process component represents the skewed survival distribution of each discrete class (LoS category) by an appropriate distribution form, which in this case is the LoS category distributions derived from the finite mixture model.

For the conditional component, different data mining models have been selected to explore the relationship between LoS category (the discrete classes on DCS) and the rest of the variables selected previously. Although the variety of techniques within the field is quite broad, the most common and popular techniques, which will be explored in this section are: Logistic regression, classification trees, Naive Bayes and hybrid methods.

The techniques will be evaluated according to how well they classify patients into the correct LoS category. Their performance will be measured through the accuracy rates (per category and overall performance) which express the percentage of times the patient membership (i.e. observed category to which they belong) matches with the membership predicted by the models discussed here.

The chapter is organized as follows: Sections 6.1 to 6.5 explores different techniques from the data mining domain for prediction of LoS category based on patient attributes, within the context of the group-based approach. Section 6.5.1 provides a comparative study of the approaches and models discussed here, with the aim of selecting the "best" approach/model" for each hospital.

6.1. Logit Regression

Let us recall that a new variable "LoS category" was created and defined as a binary variable with two categories, which corresponded to the two components of the finite mixture model: Short/Medium (patients with LoS up to 11 days) and Long (patients with LoS more than 12 days) for the case of the ISSEMyM hospital[25], and for the case of MRC hospital: Short (patients with LoS up to 2 days) and Medium/Long (patients with LoS more than 3 days).

The relationship between the LoS category and covariates can be explored through an especial case of generalised linear models: the binary Logit model or logistic regression model. Logistic regression models are one of the most common and efficient methods for classification and prediction used by statisticians and researchers in a variety of areas such as social sciences, economic research, physical sciences, health and medicine. They analyse the relationship between an explanatory variable and an outcome variable that is categorical. One of the characteristics of generalised linear models is that the dependent variable is specified to be distributed according to one of the members of the exponential family. In this context, LoS category can be specified to be distributed only according to a Bernoulli distribution, since the new dependent variable takes exclusively values of 0 and 1.

The density function of the Bernoulli distribution is given by Equation 6.1

$$f(y;\pi) = \begin{cases} 1-\pi \; for \; y = 0, \\ \pi \; for \; y = 1 \end{cases} \tag{6.1}$$

[25]The categories where coded 0 and 1 respectively

where π is both, the probability of a successful outcome ($y = 1$) and the mean(μ). Recalling that in the generalised linear model μ is associated with a linear regression model, where the explanatory variables affect the observed outcome (i.e. belonging to certain LoS category) via a canonical link function. Equations 6.2 and 6.3 are the canonical and inverse link functions respectively for the Bernoulli distribution.

$$\vartheta = g(\mu) = \ln\frac{\pi}{1-\pi} \tag{6.2}$$

$$\mu = g^{-1}(\vartheta) = \frac{1}{1+e^{-\vartheta}} \tag{6.3}$$

where $\vartheta = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p,} x_{p,}$

Therefore the binary logistic model is defined by:

$$f(y = 1; \boldsymbol{x}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_{p,} x_{p,})}} \tag{6.4}$$

where $\boldsymbol{x}$ is the vector containing the patient attributes. There are different ways to derive and interpret the binary Logit model (e.g. latent variable model, non-linear probability model, discrete choice model). It is the researcher's choice to keep the definition in terms of the generalised linear model, which was discussed in Section 5.2. However the reader is referred to Kleinbaum and Klein (2011) and Long and Freese (2006) for alternative derivations.

The Logit model was fitted on STATA using the `Logit` command, using 2/3 of the MRC dataset (named training set) and the remaining (named validation set) was used for testing purposes(Dobbin and Simon, 2011). Conversely, the estimation of the Gamma regression mixture model for ISSEMyM was performed using 9/10 of the dataset and 1/10 for validation, due to small size of the sample. The covariates selected by bootstrapping (Section 5.1.2) were included in the models as potential predictors. Table 6.1 and Table 6.2show the STATA outputs for both hospitals.

Log pseudo likelihood	-350.672
Wald chi^2(20)	70.51
Prob > chi^2	0.0000
Pseudo R^2	0.1075

	β	**Std. Err.**	***z***	***P>z***	**[95% Conf.**	**Interval]**
Age	0.0125	0.0072	1.730	0.084	-0.0017	0.0267
1stDiagnosis_category1	-0.0585	0.3329	-0.180	0.861	-0.7109	0.5939
1stDiagnosis_category2	-0.9583	0.5458	-1.760	0.079	-2.0281	0.1116
1stDiagnosis_category3	-1.2302	0.4873	-2.520	0.012	-2.1852	-0.2752
Previous admissions	-0.0942	0.0515	-1.830	0.067	-0.1951	0.0066
Outpatient clinic	-0.8124	0.2483	-3.270	0.001	-1.2991	-0.3258
Other origin	-1.5075	0.3007	-5.010	0.000	-2.0968	-0.9182
General surgery ward	-0.1745	0.3061	-0.570	0.569	-0.7745	0.4254
Trauma	-1.1658	0.5587	-2.090	0.037	-2.2608	-0.0708
Diagnosis_category1	-0.3631	0.5386	-0.670	0.500	-1.4188	0.6927
Diagnosis_category2	0.1360	0.5753	0.240	0.813	-0.9916	1.2636
Diagnosis_category3	-0.4026	0.6708	-0.600	0.548	-1.7173	0.9120
Num diagnoses	0.3678	0.1935	1.900	0.057	-0.0115	0.7470
Sp_category 1	1.0626	0.8201	1.300	0.195	-0.5448	2.6700
Sp_category 2	0.4550	0.3700	1.230	0.219	-0.2702	1.1802
Sp_category 3	0.7796	0.4371	1.780	0.074	-0.0771	1.6363
Transfusions	-0.3606	0.4298	-0.840	0.401	-1.2029	0.4818
Num comobidities	0.1642	0.2111	0.780	0.437	-0.2496	0.5781
Drinking/smoking	-0.1633	0.3382	-0.480	0.629	-0.8261	0.4995
Drinking&smoking	-0.0139	0.2581	-0.050	0.957	-0.5198	0.4920
_cons	-2.0638	0.6863	-3.010	0.003	-3.4090	-0.7186

Table 6.1: Binary Logit model for ISSEMyM[26]

[26] For a full description of the variables the reader is referred to Appendix D

Log pseudo likelihood	-4829.49
Wald chi^2(10)	1490.98
Prob > chi^2	0.0000
Pseudo R^2	0.2040

	β	Std. Err.	z	P>z	[95% Conf.	Interval]
Age	0.0100	0.0015	6.470	0.000	0.0070	0.0130
Previous adm	-0.0264	0.0040	-6.630	0.000	-0.0343	-0.0186
Gender (female)	0.0770	0.0580	1.330	0.185	-0.0367	0.1907
Outpatient clinic	-0.6379	0.0937	-6.810	0.000	-0.8216	-0.4542
General surgery ward	-0.4294	0.0736	-5.840	0.000	-0.5736	-0.2852
Diagnosis_category2	0.6565	0.0758	8.670	0.000	0.5080	0.8050
Diagnosis_category3	-1.3574	0.0880	-15.430	0.000	-1.5298	-1.1850
Sp_category 1	-0.2874	0.1011	-2.840	0.004	-0.4855	-0.0893
Sp_category 2	-0.0190	0.0874	-0.220	0.828	-0.1902	0.1522
Sp_category 3	1.0201	0.1088	9.380	0.000	0.8068	1.2333
cons	0.2048	0.1032	1.980	0.047	0.0026	0.4071

Table 6.2: Binary Logit model STATA output for MRC[26]

For both datasets, there are some coefficients that are not significant ($p > 0.1$). The common approach to follow would be to re-run the model without these variables. However some of these variables represent together a specific patient characteristic, for example Sp_cluster 1, 2 and 3 altogether represents the type of surgical procedure that the patient will undergo (let us denominate these groups of variables as variable families). Moreover Sp_cluster 1 and Sp_cluster 2 are not significant but Sp_cluster 3 is. Thus the conclusion that the type of surgical procedure that the patient will undergo does not have a significant effect on LoS_category cannot be based on the lack of significance of two coefficients, but looking at the behaviour of the entire variable family. Therefore, the variables that are not members of a family (i.e. gender, age, etc.) and that are non-significant were discarded. For those variables belonging to a family, they were discarded only when the entire variable family was non-significant.

Table 6.3 and Table 6.4show the STATA outputs after removing the non- significant variables:

Log pseudo likelihood	-329.289
Wald chi^2(13)	82.93
Prob > chi^2	0.0000
Pseudo R^2	0.1209

	β	Std. Err.	z	$P>z$	[95% Conf.	Interval]
Age	0.0135	0.0071	1.890	0.059	-0.0005	0.0275
1stDiagnosis_category1	-0.1435	0.3190	-0.450	0.653	-0.7688	0.4817
1stDiagnosis_category2	-0.8723	0.4793	-1.820	0.069	-1.8117	0.0672
1stDiagnosis_category3	-1.3097	0.5001	-2.620	0.009	-2.2899	-0.3294
Previous admissions	-0.0884	0.0551	-1.600	0.109	-0.1964	0.0196
Outpatient clinic	-0.7576	0.2472	-3.070	0.002	-1.2420	-0.2732
Other origin	-1.7030	0.3138	-5.430	0.000	-2.3179	-1.0880
General Surgery Ward	-0.1843	0.2900	-0.640	0.525	-0.7526	0.3840
Trauma	-1.2026	0.5294	-2.270	0.023	-2.2402	-0.1649
Num diagnoses	0.3885	0.1850	2.100	0.036	0.0259	0.7510
Sp_category 1	0.9384	0.7874	1.190	0.233	-0.6049	2.4818
Sp_category 2	0.4188	0.3482	1.200	0.229	-0.2637	1.1012
Sp_category 3	0.7965	0.4023	1.980	0.048	0.0081	1.5849
_cons	-2.3577	0.5764	-4.090	0.000	-3.4875	-1.2280

Table 6.3: Binary Logit model for ISSEMyM after removing non-significant variables[26]

Log pseudo likelihood	-4803.32
Wald chi^2(10)	1488.94
Prob > chi^2	0.0000
Pseudo R^2	0.2047

	β	Std. Err.	z	P>z	[95% Conf.	Interval]
Age	0.0099	0.0015	6.620	0.000	0.0069	0.0128
Previous adm	-0.0313	0.0040	-7.920	0.000	-0.0391	-0.0236
Outpatient clinic	-0.7202	0.0929	-7.750	0.000	-0.9023	-0.5381
General surgery ward	-0.4631	0.0739	-6.270	0.000	-0.6079	-0.3183
Diagnosis_category2	0.6206	0.0752	8.250	0.000	0.4732	0.7680
Diagnosis_category3	-1.3128	0.0870	-15.090	0.000	-1.4833	-1.1423
Sp_category 1	-0.2845	0.0964	-2.950	0.003	-0.4733	-0.0956
Sp_category 2	-0.0048	0.0865	-0.060	0.956	-0.1744	0.1648
Sp_category 3	1.0046	0.1088	9.230	0.000	0.7913	1.2180
cons	0.2884	0.0969	2.980	0.003	0.0985	0.4783

Table 6.4: Binary Logit model for MRC after removing non-significant variables[26]

To validate the model after removing the non-significant variables, a Likelihood ratio test (LR test)was performed. The LR test works on the null hypothesis that the coefficients of the variables to be tested are equal to zero (i.e. no effect). For example, some of the variables that were discarded in the ISSEMyM dataset are the number of patient's comorbidities, whether the patient drinks or smokes, and the final diagnosis of the patient. Therefore the null hypothesis H_0 states that,

$$H_0: \beta_{comorbidities} = \beta_{drinks/smokes} = \beta_{patient\ diagnosis} = 0$$

The LR test works by comparing the log-likelihood from a full model with that of the restricted model (i.e. after removing the non-significant variables), using the `lrtest` STATA command.

For the ISSEMyM hospital, the LR test result indicates that the effects of five of the discarded variables (i.e. "diagnosis" (any category), "transfusion", "number of comorbidities" "drinking/smoking" and "drinking&smoking) are equal to zero, and this statement cannot be rejected at the 0.1 level($LRX^2 = 10.63, df = 8, p = 0.22$).

In the case of the MRC hospital, the results are similar: the effect of patient gender is not significant at the 0.1 level ($LRX^2 = 1.77, df = 1, p = 0.18$).

In terms of interpretation of the parameters, logit regression compares the likelihood of being at the base category with the likelihood of being at the other category. The interpretation of the model can be more manageable if it comes in terms of odds ratios (e^{β}). The odds of an outcome occurring are defined as the probability of an outcome occurring (e.g. MRC patient having a Medium/Long LoS) divided by the probability of that outcome not occurring (e.g. MRC patient having a Short LoS). The odds ratio compares the change in the odds for different values of an outcome variable: if the value is greater than 1 (i.e. computed from positive *β's*) then it indicates that as the predictor increases the odds of the outcome occurring increase, and it is said that the predictor has a positive effect. Conversely a value lower than 1 (i.e. computed from negative *β's*) indicates that as the predictor increases the odds of the outcome occurring decreases, and it is said that the predictor has a negative effect.

To facilitate the interpretation, when a predictor has a negative effect on the outcome occurring, instead of calculating the odds of the event occurring, the odds of the event not occurring are computed by simply taking the inverse of the effect on the odds of the event occurring $\left(\frac{1}{e^{\beta}}\right)$.

From Table 6.3, the results for the ISSEMyM hospital can be interpreted as follows:

A patient whose "first diagnosis" belongs to "first diagnosis category 3"(e.g. diabetes mellitus, fracture in lower legs, gastrointestinal haemorrhage, etc.) is 3.7 times more likely to have a short-medium LoS than a long LoS[27].

A patient undergoing one of the surgical procedures of category 1(e.g. Cholecystectomy, laparoscopy or hysterectomy) is 2.6 times more likely to have a long LoS than a short-medium one[27].

From Table 6.4, the results for the MRC hospital can be interpreted as follows:

A patient who enters the hospital via the outpatient clinic is 2 times more likely to have a short LoS than a medium-long[27].

A patient whose diagnosis belongs to diagnosis category 2 (e.g. diabetes mellitus, stroke, hepatic failure, gastrointestinal haemorrhage) is 1.9 times more likely to have a medium-long LoS than a short LoS[27].

The next step was to perform an analysis of the residuals, in order to identify the data points for which the logit model fits poorly. Since, it is assumed that the standardised residuals are normally distributed, Field (2009) suggested that 5% of the residuals should have absolute

[27]This interpretation is true only if the effects of the other variables are held constant.

values above 2, and that no more than 1% should have absolute values above 2.5. Any observation with a value above 3 should be taken as cause for concern. Table 6.5 displays the percentages of residuals for both hospitals outside the previous criteria.

	Percentage outside ± 2	**Percentage outside ± 2.5**	**Percentage outside ± 3**
ISSEMyM	5.69%	3.58%	2.05%
MRC	5.05%	1.94%	0.74%

Table 6.5 Percentage of outliers outside the criteria based on normality

The results from the analysis of the residuals for the ISSEMyM data might be a cause for concern, however Long (2000) commented that there is no hard-and-fast rule for identifying large residuals. Moreover Hosmer and Lemeshow (2010) stated in a detailed discussion of residuals that it is impossible to provide any absolute standard as it depends on the type of data involved.

Finally, Table 6.6 describes the accuracy rates for the Logit regression model: both models perform exceptionally well in predicting patients in the first LoS category. However, they do not achieve such success in predicting patients belonging to the second category. In particular, the model for ISSEMyM fails drastically to predict patients with long LoS, although its overall accuracy rate is not seriously affected since the proportion of patients in this category is just 18.0% (more discussion about the low accuracy rates for ISSEMyM on Section 6.6). On the other hand, the model for the MRC has a consistently good performance in both categories.

	LoS category	**Accuracy rate**	**Overall**
ISSEMyM	Short-medium	100%	92.0%
	Long	8.00%	
MRC	Short	72.33%	72.83%
	Medium-long	73.40%	

Table 6.6: Accuracy rates for binary logistic models.

6.2. Decision Trees

Classification trees is one of the main techniques used in Data Mining and Machine Learning, widely used in applied fields e.g. for diagnosis and prognosis in medicine (Ture et al., 2009), species classification in ecology (De'ath and Fabricius, 2000), and market segmentation in marketing (Chen, 2003), among other fields. They are used to predict membership of cases or

observations in the classes of a categorical dependent variable from one or more predictor variables.

The tree is constructed by recursively partitioning a learning sample of data using all predictor variables to create *m* child nodes repeatedly, beginning with the entire dataset; all possible splits for each predictor variable at each node are examined to find the split which maximises the homogeneity within groups. In particular, the node where a discrete predictor *X* is tested has *m* possible splits $X = d_1, X = d_m$, where $d_1, \ldots d_m$ are the known values for predictor *X*. The node where a continuous predictor is tested has two possible splits, $X > t$ and $X \leq t$, where t is a value determined at the node and is called the threshold.

Trees are represented graphically, with the root node at the top, representing the data before division, and the branches and leaves beneath (terminal nodes). Each leaf represents one of the final groups. Most of the software packages include additional information on the tree such as a statistical summary.

In most cases, the interpretation of results is very simple, and it allows a rapid classification of new observations (e.g. the users can follow a few conditions easily) and often yields a much simpler "model" for explaining why observations are classified or predicted in a particular manner than traditional statistical methods.

Decision trees do not hold the implicit assumption that the underlying relationships between the predictor variables and the dependent variable are linear or monotonic[28]. For example, LoS_category could be positively related to a continuous variable (e.g., patient's age), if the latter is less than a certain value, but negatively related if it is more than that value.

Creating trees involves basically three steps:

1. Selecting splitting criteria: The splitting criterion is commonly divided in finding the split independent variable and in selecting the split point for the selected independent variable. Given the hierarchical nature of trees, splits occur one at time starting with the split at the root node, and continuing with splits of resulting child nodes until the stop criterion is met.
2. Selecting stop criteria: If the splitting continues, eventually each terminal node will contain only one class of objects (patients), achieving homogeneity. However in real life data contains noise or measurement error, making unrealistic the achievement of homogeneity in the terminal nodes. In practical terms the splitting should stop at certain point when the objects have been reasonably correctly classified. Some of the stopping

[28] A relation between two variables is said to be monotonic when it is either entirely non-increasing or non-decreasing.

rules or criteria are: minimum number of objects per terminal node, fraction of objects per terminal node, minimum decrease in impurity and minimum change in expected cell frequencies.

3. Finding the best tree: The choice of the optimal stop criterion value could be complex because small values might result in very large trees, with the risk of overfitting the training set and a poor performance in the testing set. Conversely higher values of the stop criteria could result in smaller trees that might not discover important structural information (e.g. interactions between independent variables). There are some procedures which assist on deciding the right size of the tree: cross-validation, V-fold cross-validation and cross-validation pruning.

The trees that will be explored in this research are: Classification and Regression Tree (CART), Quick, Unbiased, Efficient Statistical Trees (QUEST), Commercial Version 4.5 (C4.5) and Chi-squared Automatic Interaction Detection (CHAID). The reader is referred to the Appendix E for full size diagrams of each tree.

6.2.1. Classification and Regression Tree (CART)

CART can be used both for regression and classification. The goal is to produce subsets of the data which are as homogeneous as possible while keeping the size of the reasonably small.

CART, as other trees, uses exhaustive search for univariate splits method, where the selection of the split independent variable at each terminal node and selection of the split point occur at the same point. In other words, all possible splits for each independent variable at each node are evaluated to define the split producing the largest improvement in the homogeneity of the node. The number of possible splits for an independent variable is equal to $(2^{M-1} - 1)$ for a categorical variable with M values and $(n - 1)$for numerical variables with n distinct values.

Missing values are dealt with by CART using "surrogates". For the case when the value of the split predictor is missing, surrogates are other independent variables with high association with the split predictor which are used as alternative predictors.

The homogeneity of the nodes is defined by impurity, a measure which takes the value zero for completely homogeneous nodes and increases as homogeneity decreases. The most common measures of impurity (splitting criteria) for categorical dependent variables are: Gini, information index, towing and ordered towing. In this study, Gini impurity was used for categorical variables. The Gini index takes the form$1 - \sum c^2$, where c are the proportions of responses in each category. At each split the Gini index tends to split off the largest category into a separate group.

The tree continues splitting until the decrease in impurity due to further splits is less than a user-specified stop criterion (i.e. CART in SPSS uses a minimum change of 0.0001 in improvement of the Gini index).

The stop criteria in CART often generates over-large trees. To solve this problem the tree is pruned following two steps. Firstly, the selection of a family of subtrees is carried out according to a penalty function. During the construction of the tree, before starting the splitting, the root node is defined as a subtree (size[29] one) of the fully grown tree (defined as T_{max}). As each split is added to the tree, a subtree of size $|\widetilde{T}|$ is generated and its misclassification rates computed. When the tree is fully grown, a sequence of decreasing misclassification rates are calculated, known as resubstitution costs *R(T)* (or resubstitution estimates errors). Each resubstitution cost corresponds to a specific subtree of size $|\widetilde{T}|$. In other words, there is a misclassification cost for each split of the tree.

From the large number of generated subtrees, a sequence of trees can be found in order of decreasing size:$T_0,\ T_1,\ T_2, \dots, T_L$ where T_0 is the first tree after pruning T_{max} and T_L is the last tree containing only the root node and no splits at all. This sequence should contain the best subtrees of their size, which minimise the cost-complexity function, defined as Equation 6.5:

$$R(T) + \alpha|\tilde{T}| \tag{6.5}$$

where $\alpha \geq 0$ is the complexity parameter. By allowing α to increase, larger subtrees are penalized for their complexity. Initially the complexity function is computed for every subtree of $T_{max}$where$\alpha = 0$. Thenα is increased continuously until the complexity function value of the largest subtree exceeds the complexity function value of a smaller subtree, this smaller subtree is defined asT_0. The complexity parameter is again increased until the value of the complexity function of T_0 exceeds the value of the complexity function of a smaller subtree, which now is T_1, the process continues until T_L (the tree containing the root node) is found. The sequence of nested subtrees $T_0,\ T_1,\ T_2, \dots, T_L$ is defined as the reference parametric family $T(\alpha)_{max}$.

To select the best subtree, cross-validation is used in order to obtain an estimate of the true error of the tree parametric family. *V*-fold cross-validation is useful to get more accurate estimates of the error. The dataset $\mathfrak{R}$, which was used to build the reference parametric family, is split in *V* equally sized folds: $\mathfrak{R}^1, \mathfrak{R}^2, \dots, \mathfrak{R}^V$ (10 folds are often recommended in literature). A new auxiliary tree T_{max}^1 is generated using $\mathfrak{R} - \mathfrak{R}^1$ and its classification costs, now called cross-validation costs (CV costs), are computed on the testing set $\mathfrak{R}^1$. Then $T_{\max}^2$ is generated from

[29] The size of the tree is defined as the number of leaves in the tree

$\mathfrak{R}-\mathfrak{R}^2$ and CV costs calculated from $\mathfrak{R}^2$, and so on, until V auxiliary trees are generated: $T^1_{max}, T^2_{max}, \dots, T^V_{max}$. If $V = 10$, in each turn 90% of the data is used to build the auxiliary tree and 10% is held back for testing. Following the approach in the first step, it is possible to generate V distinct parametric families: $T(\alpha)^1_{max}, T(\alpha)^2_{max}, \dots, T(\alpha)^V_{max}$.The next step is grouping all the subtrees of the V parametric families according to their size (i.e. each group just contains subtrees of the same size). The CV from each group are averaged, enabling more accurate error estimates of a specific subtree of size $|\widetilde{T}|$. The tree size that produces the minimum cross-validation cost is labelled "minimum CV". It is important to note that the auxiliary trees built during the cross-validation process are used only to find the minimum CV. The best tree for classification of the dataset $\mathfrak{R}$ is the smallest tree from the reference parametric family, of which $R(T)$ is within one standard error from the minimum CV. These pruning criteria are formally called "minimal cost-complexity cross-validation pruning".

The first tree generated in SPSS for the ISSEMyM dataset (Figure 6.1) surprisingly contains just one node, the root node, which represents the data before division, and no branches or leaves beneath.

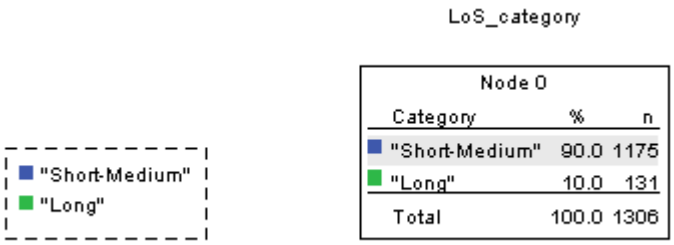

Figure 6.1: CART tree for ISSEMyM

The fact that no variable emerges as a predictor (i.e. compared with the Logit regression model where many variables resulted significant to predict LoS category) rises doubts about whether the tree might be overpruned. Esposito et al. (2002) support that indeed minimal cost-complexity cross-validation pruning has a propensity for overpruning. They argued the 1SE rule is too wide, allowing very small trees to be selected, with *R(T)*s still within one standard error. Therefore, for the ISSEMyM dataset, it was decided to use a 0SE rule instead, in order to find the subtree from the reference parametric family with the smallest *R(T)*; although simplicity may have to be sacrificed (i.e. *R(T)* decreases with more complex trees). However the results for the re-estimated tree using the 0SE rule shows exactly the same outcome: a single node, discarding "partially" the hypothesis of overpruning. It is mentioned "partially" because Esposito et al. (2002) also argues that the cross-validation used in the cost-complexity pruning may provide an error rate estimate whose amount of bias is unpredictable.

Figure 6.2shows an unpruned CART tree for the ISSEMyM hospital with seven nodes (i.e. size 7). The interpretation is very straightforward and intuitively clear: the patient's origin is the best predictor of LoS category. However all patients are more likely to have a short-medium LoS, which explains why the minimal cost-complexity cross-validation pruning generated a tree of size one.

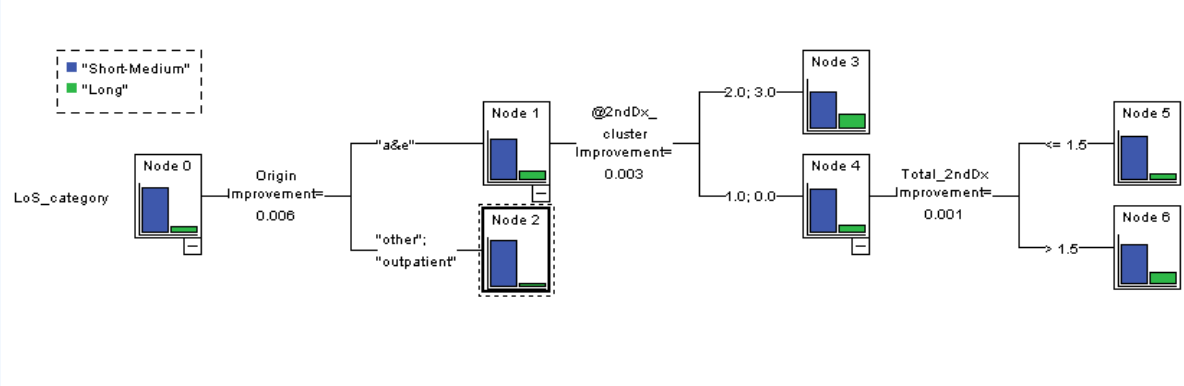

Figure 6.2: CART for ISSEMyM[30]

On the contrary, Figure 6.3 depicts the tree generated for the MRC hospital which discriminated better between the two LoS categories and where patient diagnosis is the best predictor of LoS category.

[30] The key for the variables acronym and full description of them is available at the Appendix D

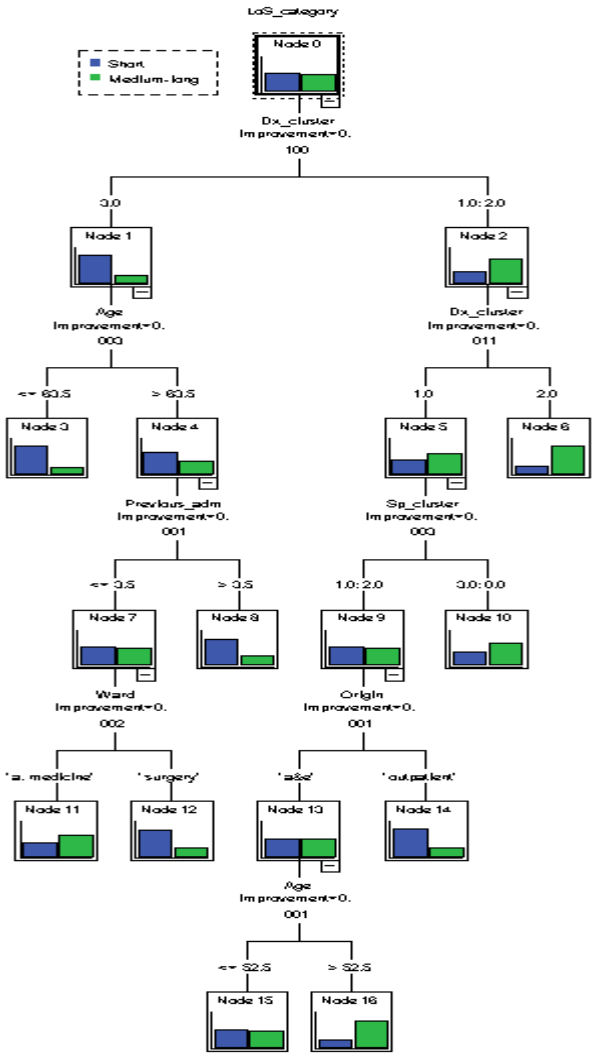

Figure 6.3: CART for the MRC hospital[30]

6.2.2. Quick, Unbiased, Efficient Statistical Trees (QUEST)

QUEST (Loh and Shih, 1997) was developed as a response to the two main drawbacks of CART: computational complexity and bias in variable selection. The latter is due to the splitting criteria in CART, which tends to select variables that have more splits points[31].

QUEST uses discriminant-based univariate splits method as splitting criteria: firstly, the split independent variable in each node is selected by performing an ANOVA (Analysis of variance) F-test for each continuous independent variable (*p-values* of F statistics are recorded). Then, for each categorical independent variable Pearson's χ^2 test is performed to identify associations between the outcome variable and the predictors (*p-values* of theχ^2 test are recorded).The next

[31] For example, when a variable is continuous and the other is discrete with few distinct values, exhaustive search shows a large bias towards the continuous variable because it affords more splits.

step is to select the independent variable with smallest *p-value* and denote it *X**.It is important to remember that smaller *p-values* in an ANOVA test, support the hypothesis that if an independent variable is split into the categories of the outcome variable, each category has a different mean. On the other hand, smaller *p-values* in Pearson's χ^2 test, indicate strong association between the independent and dependent variables. In rough terms, both are indicators of the "strength" of the independent variable's influence on the dependent variable.

If this smallest *p-value* is less than the default Bonferroni adjusted *p-value*[32] for multiple comparisons, *X**is selected as the split predictor for the node. If no *p-value* is smaller than this threshold, then other statistical tests (that are robust to assumptions) are performed for numeric independent variables, such as *Levene's F* statistic. Smaller *p-values* on *Levene's F* statistic support the hypothesis that the variance of the independent variable is not the same in all the categories of the outcome variable. Subsequently all *p-values* are compared again and the independent variable with the smallest p-value is denote as *X***. If this smallest *p-value* is less than $\alpha/K + K1$, where *K* is the number of predictors and $K1$ is the number of continuous predictors, *X*** is selected as the split predictor for the node, otherwise *X** is selected.

To deal with missing values, QUEST uses the same approach as CART, based on selecting surrogates for the missing values.

On the other hand, the procedure to determine the split point for the selected independent variable is rather complicated and the details can be found in (Loh and Shih, 1997). In rough terms, QUEST employs for numeric variables the two-means clustering algorithm of Hartigan and Wong (1979). The aim is to divide the variable into two super classes, using the two most extreme sample means as initial cluster centres. Then using the mean, variance and priors probabilities of each super class a quadratic equation is deduced (quadratic discriminant analysis) and the two solutions or roots of the equation are computed. QUEST chooses the solution which is closer to the mean of each super class as the split point of the numeric independent variable.

In the case of categorical variables QUEST uses CRIMCOORD to transform the categorical variables into numerical values: Suppose *X* is a categorical variable taking values in the set $\{c_1, \dots c_M\}$. Then, for all the *N* categorical variables, each value of *X* is transformed into an *M*-dimensional dummy column vector $\boldsymbol{v} = (v_1, \dots, v_M)'$, where Equation 6.6 applies:

$$v_l = \begin{cases} 1, & if\ X = c_l, \\ 0, & otherwise. \end{cases} \tag{6.6}$$

[32] The Bonferroni adjusted p-value is equal to α/K where $\alpha \in (0,1)$ is a user-specified level of significance and K is the total number of independent variables.

Let V be the $N \times M$ matrix consisting of the $v - values$.

Then principal components analysis is used to reduce the dimensionality of the matrix *V*. The reduced dimensional data is projected onto the largest discriminant coordinate, whose values are used as the numerical value of the categorical variable. The next step consists of applying the same algorithm used for numerical variables to find the split point. Finally the split point is re-expressed into the form of categorical values.

To find the best tree, QUEST uses the same approach as CART: minimal cost-complexity cross-validation pruning.

The first tree generated in SPSS for the ISSEMyM dataset contains again just one node. This is because QUEST uses the same pruning method as CART: minimal cost-complexity cross-validation pruning. Therefore it was decided to run an unpruned version depicted in Figure 6.4, where patient's origin emerged again as the best predictor; and, as was observed in the CART results, there is no discrimination for patients with long LoS.

Conversely the QUEST tree generated for MRC in Figure 6.5, defines patient's diagnosis as the best predictor: patients with a diagnosis in category 1 (e.g. cholecystitis, appendicitis, and gastroenteritis) and category 2 (e.g. diabetes, stroke, hepatic failure, gastrointestinal haemorrhage, etc.) are more likely to have a medium-long LoS, whereas patients with a diagnosis in category 3 (e.g. inguinal hernia, umbilical hernia, renal failure, etc.) are more likely to have a short LoS.

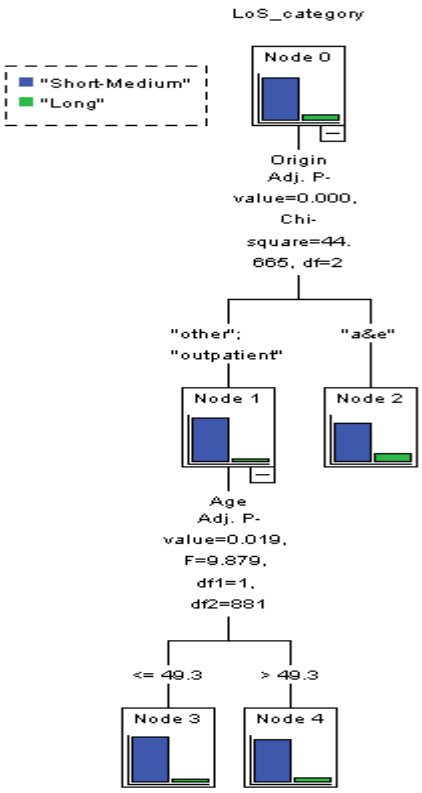

Figure 6.4: QUEST for ISSEMyM[30]

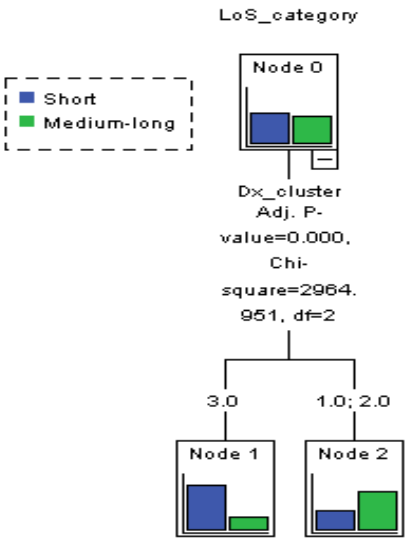

Figure 6.5: QUEST for MRC hospital[30]

6.2.3. Commercial Version 4.5 (C4.5)

C4.5 developed by Quinlan (1993) has always been taken as the point of reference for the development more sophisticated trees. C4.5 constructs the classification tree under the divide and conquer algorithm.

The splitting criterion is based on the information gain ratio[33] of each independent variable. The independent variable with the highest value is selected as the split predictor. If a variable is selected as the split predictor the number of child nodes will be *s*. In this context, for categorical variables $s = h$ where *h* is the number of known categories and $s = 2$ for continuous variables.

Let Z be the set of observations in every node and $C_1, \dots, C_{NClass}$ the categories of the outcome variable. For categorical variables the information gain ratio is calculated as follows (see Equations 6.7-6.10):

$$Info(Z) = -\sum_{j=1}^{NClass} \frac{freq(C_j,\ Z)}{|Z|} \times log_2\left(\frac{freq(C_j,\ Z)}{|Z|}\right) \tag{6.7}$$

$$Gain(Z) = Info(Z) - \sum_{i=1}^{s} \frac{|Z_i|}{Z} \times Info(Z_i) \tag{6.8}$$

$$Split(Z) = -\sum_{i=1}^{s} \frac{|Z_i|}{|Z|} \times log_2\left(\frac{|Z_i|}{|Z|}\right) \tag{6.9}$$

$$Information\ gain\ ratio = \frac{Gain(Z)}{Split(Z)} \tag{6.10}$$

Equation (6.7) measures the average amount of information needed to identify the number of classes in *Z*. It is called as well as the entropy function[34]. Equation (6.8) measures the information that is gained by the partition of Z according to the split predictor. Moreover, Equation (6.9) represents the potential information generated by dividing Z into *n* subsets and Equation (6.10) represents the proportion of information generated by the split that is useful for classification.

For continuous variables the threshold *t* that maximizes the information gain ratio is found by the following steps: the values of the continuous variable X are ordered $v_1, v_2, \dots v_N$. Every pair of adjacent values suggests a potential threshold value $t = \frac{v_i + v_{i+1}}{2}$, and a corresponding split of Z into two nodes: $X \leq t$ and $X > t$. For each value *t*, the information gain ratio $gain_t$ is

[33] Information gain ratio is based on the "Information theory" which establishes that the information conveyed by message depends on its probability and can be measured in bits as minus the logarithm to base 2 of that probability (Quinlan, J.R., 1993)

[34] Entropy is a measure of the uncertainty associated with the independent variable. Lower values of entropy of the class indicate that the class is more predictable. The amount by which the entropy of the class decreases reflects the additional information about the class provided by the independent variable and is called information gain.

calculated. The value t^*, which $gain_t$ is maximum, is set to be the local threshold (Quinlan, 1996). Next, if the variable is selected as the split predictor, the split point is calculated by a linear search in all the values of the continuous variable that best approximates the local threshold *t** from below.

To deal with missing values C4.5 modifies Equation (6.8) to Equation (6.11)

$$Gain(Z) = F\left(Info(Z) - \sum_{i=1}^{s} \frac{|Z_i|}{Z} \times Info(Z_i)\right) \tag{6.11}$$

where *F* is the proportion of known cases $(i.e. 1 - \%\ of\ missing\ observations)$. In addition Equation (6.9)(i.e. $split(Z)$) is modified by adding a new predictor category, which includes the missing observations. During the partitioning, if an observation has a missing value, it is divided into pieces or fragments, resulting in a single observation belonging to more than one child node within the same parent. The fragmentation is done according a weighting system representing the probability that the observation belongs to each node (Quinlan, 1996)

The splitting process continues until all the observations belong to the same class C_j or the number of observations is less than a user specified value. Then, the splitting process stops and the node is converted into a leaf with an associated class C_j (the most frequent class).

To find the best tree, C4.5 uses error-based pruning (EBP) as a pruning approach (Quinlan, 1987): the fully grown tree is defined as T_{max}, T_t is the branch containing the node *t* and all its children and $T_{t'}$ is defined as a sub-branch rooted in a child *t'* of *t*. Every leaf (terminal node) contains *K* cases from which ones *J* are incorrectly classified. Using a testing set, for every *t*, EBP compares the misclassification error of T_t against the misclassification error of the two simplified trees, one where *t* is turned into a leaf and other one where $T_{t'}$ is grafted onto the place of *t*. If the performance of one of the simplified trees is better, the branch is pruned. The operation is repeated until the misclassification error increases (Esposito et al., 2002).

Quinlan (1987)introduced a more "pessimistic" view of misclassification error, which is arguably a more realistic error rate, where *J* is replaced with the upper 75% confidence bound *J',* assuming a binomial distribution.

For the ISSEMyM dataset, the C4.5 algorithm, like its predecessors, produced a tree with just a single node. In this case, two alternatives were explored: first WEKA[35], which is the software used to implement the C4.5 tree, allows the user to select another pruning approach called reduced error pruning (REP). Unfortunately, after running the modified algorithm, there was no

[35] WEKA acronym stands for Waikato Environment for Knowledge Analysis.

improvement, REP produced also a tree of size one. The second alternative was to leave the tree unpruned, as was done previously with CART and QUEST. Figure 6.6 pictures the unpruned C4.5 tree for the ISSEMyM dataset, which unlike their predecessors successfully classified a very small proportion of patient as long LoS, i.e. those who entered the hospital via the A&E department, had a diagnosis in category 3 (e.g. chronicle renal failure, hernia, benign lipomatous etc.) and underwent a surgical procedure category 3 (e.g. laparoscopy, subtotal hysterectomy, prostatectomy, cholecystectomy, etc.) were more likely to have a long LoS.

Figure 6.7 shows the C4.5 tree for the MRC dataset, which is considerably bigger that the CART and QUEST versions. The tree points patient's diagnosis as the best predictor for LoS category.

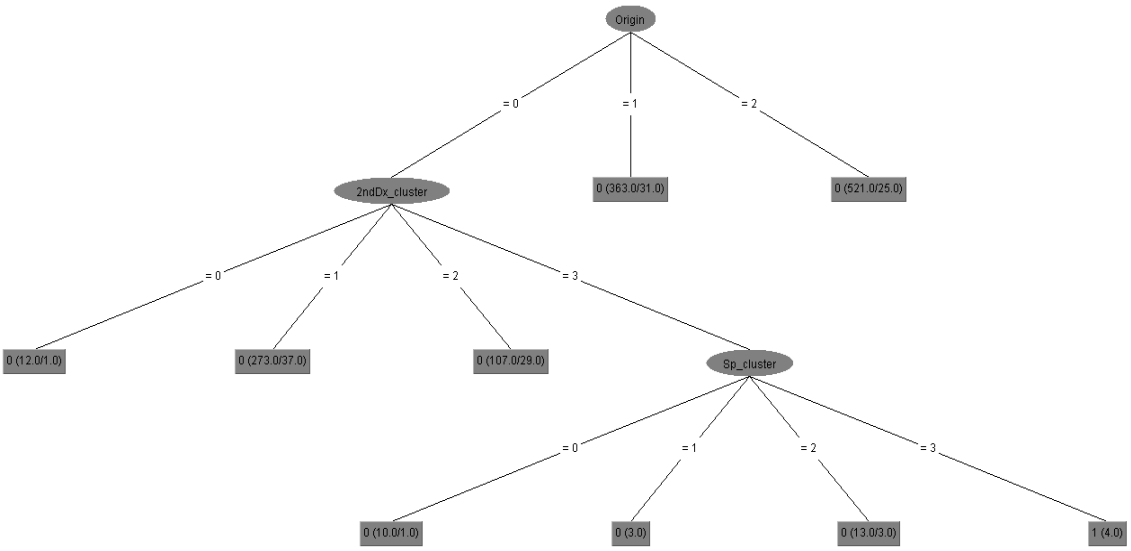

Figure 6.6: C4.5 generated for ISSEMyM[30]

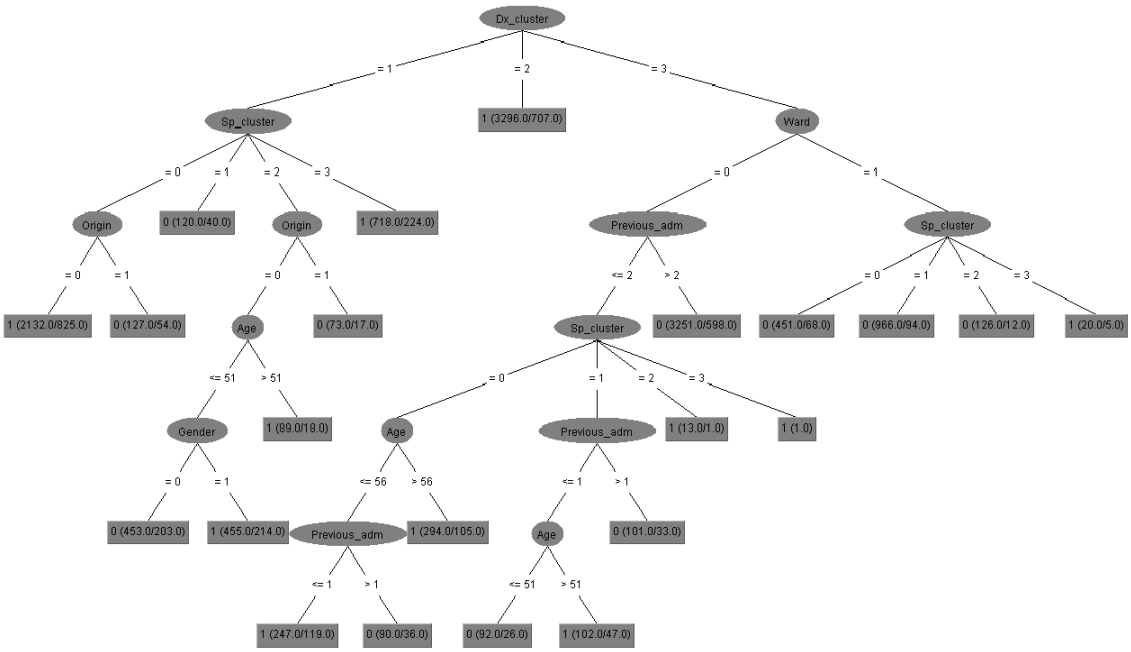

Figure 6.7: C4.5 generated for MRC[30]

6.2.4. Chi-squared Automatic Interaction Detection (CHAID)

CHAID tree developed by Kass (1980) is based on the χ^2 test of association. The best split at any node is defined by merging any allowable pair of categories of the predictor variables until there is no statistically significant difference within the pair with respect to the target variable. This CHAID method naturally deals with interactions between the independent variables that are directly available from an examination of the tree.

The CHAID algorithm only accepts nominal or ordinal categorical independent variables. When variables are continuous, they are transformed into ordinal predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations.

For every predictor, firstly a $m \times k$ two-way table is created where m represents the categories of the predictor variable and k refers to the categories of the outcome variable is created. The pair of categories of the predictor whose *p-values* of the χ^2 test are least significant are merged into one compound category if the *p-value* is less the Bonferroni adjustment (discussed in Section 6.2.2); this operation is repeated until all the categories have significant *p-values*. It is important to note that for ordinal variables just contiguous categories may be grouped together. In addition, any category having less than user specified number of cases should be merged with the most similar category, judged by the largest of the *p-values.*

For the case of missing values, CHAID considers the "unknown" value as another possible value (or category) for each independent variable and deals with it as it does with the other variables (Liu et al., 1997).

Next, the independent variable whose χ^2 is the largest is selected as the split predictor and the data is then subdivided into $m \leq l$ subsets, where l is the number of categories resulting from the merging process. The process from the first step (i.e. merging process) is repeated until non-significant χ^2 values are available.

To find the best tree, CHAID does not use any pruning method as do the other methods, however SPSS allows performing *V*-fold cross-validation to get a better estimate of the misclassification error. The dataset $\mathfrak{R}$ is split into *V* equally size folds $\mathfrak{R}^1, \mathfrak{R}^2, \ldots, \mathfrak{R}^V$. Next, a new auxiliary tree is generated via the CHAID algorithm using $\mathfrak{R} - \mathfrak{R}^1$ as the training set and its misclassification error is computed on the testing set $\mathfrak{R}^1$. Afterwards, a second tree is generated from $\mathfrak{R} - \mathfrak{R}^2$ and the CV costs (see Section 6.2.1) are calculated from $\mathfrak{R}^2$, and so on, until *V* auxiliary trees are generated: Then the misclassification errors of the *V* auxiliary trees are averaged to get the *V-fold* classification error. As happens in CART, the auxiliary trees built during the cross-validation process are used only to find the *V-fold* classification error. The final tree presented is built using the whole dataset $\mathfrak{R}$.

Figure 6.8 and Figure 6.9 illustrate the CHAID trees for ISSEMyM and MRC respectively. The same variables that emerged as best predictors for LoS category previously (patient's origin and diagnosis respectively) have the same role in the CHAID trees. Unsurprisingly, the CHAID tree for ISSEMyM repeats an inability to discriminate patients with long LoS.

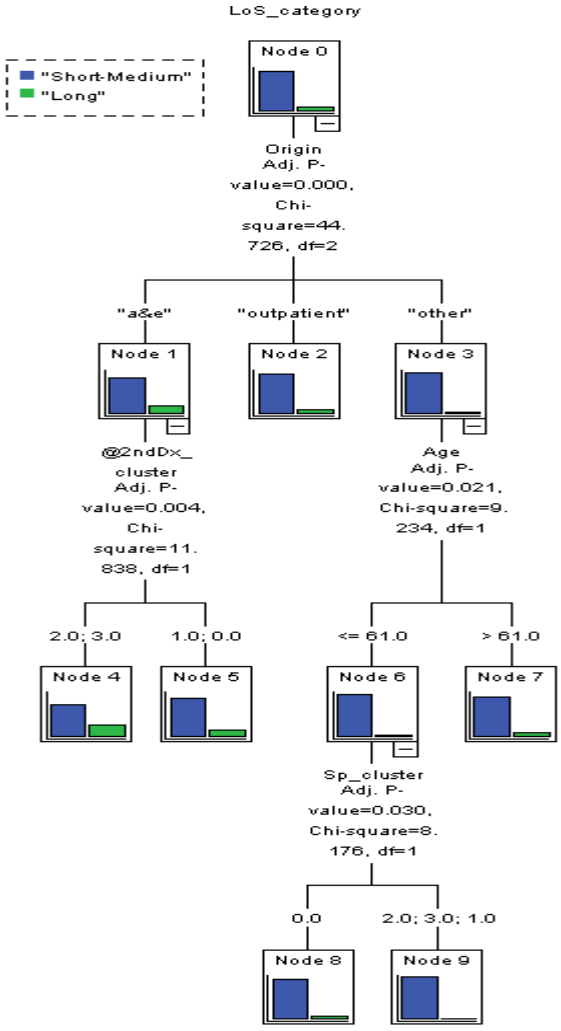

Figure 6.8: CHAID for ISSEMyM[30]

On the other hand, the CHAID tree generated for MRC is a lot more difficult to interpret due to its size. However in rough terms, one can identify the same pattern as in the QUEST version, where patients with a diagnosis under category 1 (e.g. cholecystitis, appendicitis, and gastroenteritis) or under category 2 (e.g. diabetes, stroke, hepatic failure, gastrointestinal haemorrhage, etc.) are more likely to have a medium-long LoS, with just the exemption of those patients who additionally undergo a surgical procedure under category 1 (e.g. dialysis, open hernia repair, aspiration skin, etc.) who are more likely to have short LoS. On the other hand, patients with a diagnosis in category 3 (e.g. inguinal hernia, umbilical hernia, renal failure, etc.) are more likely to have a short LoS.

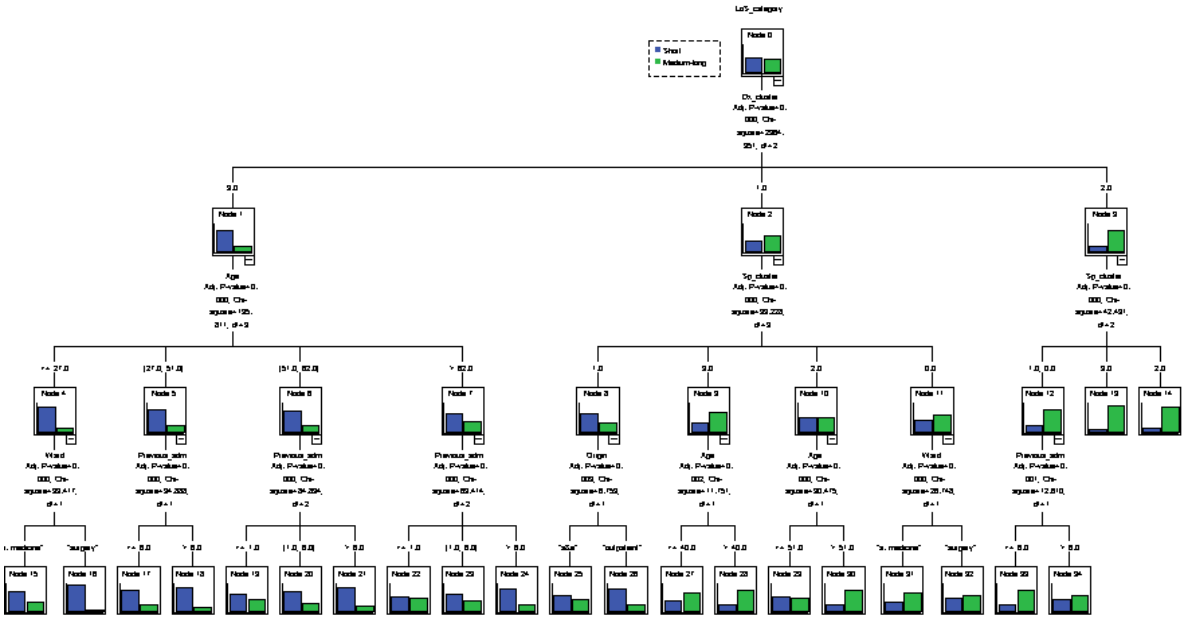

Figure 6.9: CHAID for the MRC hospital[30]

6.2.5. Discussion

The following tables summarise the main characteristics and accuracy rates of the four tree algorithms.

	Variables used in the model	Terminal nodes	Tree size	Short-medium	Long	Overall accuracy
CART	Origin (A&E, outpatient clinic and other), diagnosis (four categories) and number of diagnoses	4	7	100.00%	0%	90.00%
QUEST	Origin (A&E, outpatient clinic and other), and First diagnosis (four categories)	3	5	100.00%	0%	90.00%
C4.5	Origin (A&E, outpatient clinic and other), diagnosis (four categories) and surgical procedures (four categories)	9	12	98.90%	5.3%	89.30%
CHAID	Origin (A&E, outpatient clinic and other), diagnosis(four categories), age and surgical procedures (four categories)	6	10	100.00%	0%	90.00%

Table 6.7: Comparative table for the ISSEMyM hospital

	Variables used in the model	Terminal nodes	Tree size	Short	Medium-Long	Overall accuracy
CART	Diagnosis (three categories), surgical procedure (four categories), origin (A&E and outpatient clinic), age and gender	7	13	68.80%	76.89%	72.64%
QUEST	Diagnosis (three categories),	2	3	64.70%	79.79%	71.90%
C4.5	Diagnosis (three categories), previous admissions, surgical procedures (four categories), ward (adult medicine and g. surgery), origin(A&E and outpatient clinic), age and gender	22	36	66.80%	80.41%	73.33%
CHAID	Diagnosis (three categories), previous admissions, surgical procedures (four categories), ward (adult medicine and g. surgery), origin (A&E and outpatient clinic), and age	22	35	73.03%	71.91%	72.8%

Table 6.8: Comparative table for MRC hospital

The accuracy rates are almost identical for the four methods and show the same tendency as the Logit regression model. Unfortunately for the ISSEMyM hospital, the trees generated by CART, QUEST and CHAID algorithms, apart from highlighting the significant variables, were useless for classification purposes.

It has been mentioned widely in literature that no single tree algorithm dominates over all others (Ture et al., 2005). Any tree algorithm's performance might change from one data structure to another. The QUEST algorithm produced the simplest trees for both datasets followed by CART. According to Loh and Shih (1997), the QUEST algorithm tends to yield shorter trees than the exhaustive search method uses by CART. In this context, the difference could be explained by the fact that most of the variables selected by QUEST can afford many splits (e.g. diagnosis, surgical procedures, age, previous admissions, etc.) whereas CART has demonstrated to be biased on selecting precisely those ones with more splits.

On the other hand, some authors such as Breiman (1996b), Ting and Zheng (1999) and Dietterich (2000a) reported that classification trees are unstable methods, where small changes in the training dataset can yield to large changes in the resulting classification tree. Dietterich (2000a) explains thoroughly why the C4.5 is one of the most highly unstable methods.

Quinlan (1996) quoted the general complain that C4.5's performance is weaker in datasets with a preponderance of continuous variables than in datasets containing mainly discrete variables. It

seems that C4.5 algorithm does not take full advantage of the information that continuous predictors supply (Wang et al., 2006).

In terms of how the models handle missing values, it seems that the method used by CART and QUEST to find surrogates is more efficient than the methods used by C4.5 and CHAID: the latter methods, especially CHAID's, (i.e. where all the observations with missing values are put apart in separate nodes), lead to the loss of valuable information. Therefore finding surrogates is more efficient, provided that it is possible to find suitable substitutes (i.e. variables with high association).

6.3. Naïve Bayes

Recursive graphical models or probabilistic expert systems are considered as an important and sophisticated tool for prediction purposes in Data mining. In addition, these methods can also be used to classify observations, when they are more commonly known as Bayesian networks. One of the simplest and most useful Bayesian network is the Naïve Bayes model (Jiawei and Kamber, 2001). The Naïve Bayes classifier works as follows:

Each observation on the data can be defined as a vector $\boldsymbol{x} = (x_1, x_2, \dots, x_s)$ here x_s is the value of the independent variable A_s. Then the classifier assigns an unlabelled observation X to the class (or outcome category) C_i if and only if (see Equation 6.12 and 6.13)

$$P(C_i|\boldsymbol{x}) > P(C_j|\boldsymbol{x}) \text{ for } 1 \leq j \leq m, j \neq i. \tag{6.12}$$

By Bayes theorem,

$$P(C_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_i)P(C_i)}{P(\boldsymbol{x})} \tag{6.13}$$

The prior probability of $\boldsymbol{x}$, $P(\boldsymbol{x})$ is constant for all classes, so just $P(\boldsymbol{x}|C_i)P(C_i)$ needs to be estimated. The class prior probabilities are estimated by $P(C_i) = n_i/n$ where n_i is the number of observations belonging to class C_i and n is the total number of observations.

This classifier is called "naïve" because it makes two important assumptions. The first one, called the class conditional independence assumption, states that the values of the independent variables are conditionally independent from one to another, in other words there are no dependence relationships among predictors. Therefore, $P(\boldsymbol{x}|C_i)$ can be estimated as follows:

$$P(\boldsymbol{x}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{6.14}$$

For categorical variables $P(x_k|C_i) = \frac{n_{ik}}{n_i}$, where n_{ik} is the number of observations belonging to class C_i whose value in the predictor A_k is x_k.

The second important assumption is regarding numeric variables: it is assumed they follow a normal distribution:

$$P(x_k|C_i) = g(x_k, \mu_{Ci}, \sigma_{Ci}) = \frac{1}{\sqrt{2\pi\sigma_{Ci}}} e^{(x_k-\mu_{Ci})^2 / 2\sigma_{Ci}^2} \tag{6.15}$$

where $g(x_k, \mu_{C_i}, \sigma_{C_i})$ is the normal density function for the numeric predictor $A_k$whereas $\mu_{C_i}$and σ_{C_i} are the mean and standard deviation respectively for the observations belonging to class C_i.

Despite the simplifying assumptions that underlie the Naïve Bayesian classifier, experiment on real world data have repeatedly shown it as a good competitor with much more sophisticated classification algorithms(John and Langley, 1995).

Naïve Bayes was implemented for the ISSEMyM dataset using WEKA: Table 6.9 summarises the predictors used to build the models for both datasets. For numeric variables, it displays the associated parameters of the normal distribution fitted to those variables. For the rest of the variables, the table displays the counts in each category (i.e. the number of patients belonging to class $C_i$whose value in the predictor A_k is x_k) The results of ten-fold cross-validation show a lower overall accuracy rate of 84.67% compared with the logistic model and trees. This is mainly because the accuracy rates are 92.8% and 12.8% for short-medium and long LoS respectively. The overall accuracy rate for the MRC dataset was 72.4%, very similar to the values obtained with previous models. The accuracy rate for the short category was 61.9% and for the medium-long category 84%, slightly below and above the rates obtained with the other models respectively.

ISSEMyM		
Predictor	**Short/medium**	**Long**
	0.9	0.1
Age		
mean	46.725	51.7976
std. dev.	15.48	15.21
Previous admissions		
mean	1.086	0.7023
std. dev.	3.3265	1.4233
Num. diagnoses		
mean	1.1472	1.2901
std. dev.	0.5215	0.6823
Num. comorbidities		
mean	0.2204	0.2672
std. dev.	0.5218	0.5501
First diagnosis		
none	189	31
Category 1	641	83
Category 2	241	14
Category 3	108	7
Total	1179	135
Origin		
A&E	348	76
Outpatient	333	32
Transfer	497	26
Total	1178	134
Ward		
Adult Med.	444	66
G. Surgery	587	61
Trauma	147	7
Total	1178	134
Diagnosis		
none	71	8
Category 1	714	73
Category 2	231	42
Category 3	163	12
Total	1179	135
Surgical Procedure		
none	535	72
Category 1	52	5
Category 2	433	36
Category 3	159	22
Total	1179	135
Transfusions		
yes	1070	123
no	107	10

MRC		
Predictor	**Short**	**Medium/long**
	0.52	0.48
Age		
mean	42.6749	47.877
std. dev.	19.4496	19.9894
Previous admissions		
mean	9.4863	2.9346
std. dev.	15.8978	7.551
Gender		
Female	3635	3241
Male	3301	3044
[total]	6936	6285
Origin		
A&E	5768	5904
Outpatient	1168	381
[total]	6936	6285
Ward		
Adult Med.	4020	3717
G. Surgery	2916	2568
[total]	6936	6285
Surgical Procedure		
none	3771	4369
Category 1	2251	523
Category 2	668	632
Category 3	248	763
Diagnosis		
Category 1	1741	2428
Category 2	708	2590
Category 3	4488	1268
[total]	6937	6286

[total]	1177	133
addiction		
none	663	73
Drinking/smoking	212	23
Drinker&smoker	303	38
[total]	1178	134

Table 6.9: WEKA output using Naive Bayes algorithm. For the numeric variables, the associated parameters of the normal distribution fitted is displayed. For the rest of the variables, the counts on each category are displayed

However the assumption of normality of the numeric attributes seems quite unrealistic. Although P-P plots (not displayed here) showed that patient's age follows reasonably well the normality assumption for both datasets (with some heavy tails in the case of ISSEMyM though). However, this is not the case for the rest of the numeric variables in both datasets (i.e. previous admissions, number of diagnosis and number of comorbidities). It is important to remember that when statistical assumptions are broken, the model cannot be accurately applied to the whole population.

Naive Bayes on WEKA presents two alternatives to deal with non-normally distributed numeric attributes: Kernel estimation [36] and discretisation. Kernel estimation is a non-parametric technique for estimating the unknown probability density function of a random variable, more details of its estimation can be found in John and Langley (1995). On the other hand, discretisation converts numeric variables into categorical ones (C4.5 and CHAID use other forms of discretisation). The two most common methods of discretisation are "equal interval width" and "equal frequency interval": the former one merely divides the range of observed values for a variable into *k* equal sized bins, where *k* is user specified, whereas the latter divides the range of observed values into *k* bins where each bin contains *m/k* adjacent values (Dougherty et al., 1995).

Naïve Bayes was then re-estimated using different combinations of methods to handle numeric predictor (i.e. using kernel estimator for patient age and equal frequency interval for the rest). However, the results did not show a significant improvement of the overall accuracy rates using any of the techniques mentioned above in both datasets. Therefore, the Naive Bayes model version that will be recalled in the future is the one without discretisation and/or kernel estimation of the numeric attributes.

One of the main advantages of Naïve Bayes is its simple computation which does not required necessarily specialised software (i.e. it could be implemented in a spread sheet), and as new data becomes available, it can included to the model easily for re estimation. Another advantage is

[36] When Naives Bayes uses kernel estimation it is often referred to as Flexible Naïve Bayes

that missing values do not represent a problem: if the value of an independent variable is missing, such a variable is simply omitted from Equation (6.14) and the probability ratios are based on the number of values that actually occur rather than on the total number of observations (Witten and Frank, 2005).This makes it very robust to irrelevant attributes since the final classification is based on evidence from many variables to make the final prediction, a property that is very useful where there is doubt whether the variables have an effect on the dependent variable (Kohavi, 1996).

On the other hand, the Naives Bayes output is not more than a table of prior and conditional probabilities derived from the dataset. It provides useful information but it does not have the same graphical interpretation and straightforward prediction of new observations compared with classification trees. Although there have been some attempts to make the visualisation of the Naïve Bayes more accessible to non-experts (i.e. Becker et al., 2000; Kononenko, 1993 and Možina et al., 2004), this required extra calculations and sometimes the use of additional software.

Another reason of concern is the conditional independence assumption. During the variable selection process (Section 5.1) some variables in both datasets showed some signs of collinearity. Therefore we cannot reject completely the presence of dependency among predictors. In this context, other Bayesian networks might be considered since they model causal relationships between variables.

Finally let us exemplify how the Naïve Bayesian classification works by creating a virtual patient named “Eva” with the following characteristics: Eva is a female patient, 57 years old, treated in the adult medicine ward at MRC hospital with a diagnosis of hepatic failure. She entered to hospital via A&E department and has a record of 30 previous admissions to this hospital.

Variable	Observed value	Short $P(Short) = 0.52$	Medium/Long LoS $P(Medium/Long) = 0.48$
Age	57	0.016	0.018
Gender	Female	0.524	0.516
Previous adm	30	0.011	0.000
Origin	A&E	0.832	0.940
Ward	Adult medicine	0.421	0.409
Surgical p. category	0	0.544	0.695
Diagnosis category	3	0.647	0.202
$\prod_{k=1}^{n} P(x_k \mid C_i)$		1.10E-05	4.29E-08

Table 6.10: Posterior probabilities of X conditioned to C_i

Table 6.10 summarises the posterior probabilities based on the results of Table 6.9 and Eva's attributes. Notice that the highest posterior probability occurs when "Eva" is classified to the short LoS category

6.4. Hybrid Methods

Hybrid methods have emerged in the context of data mining in order to boost classification and prediction performance (among other objectives) by a coupling of two or more algorithms. Such approaches have experienced a revival over the last few years, connected with the fact that no one algorithm is best for all problems, depending on the nature and quality of the data. The concurrent development of technology which allows faster computations and more friendly platforms for implementation has opened up a new future for hybrid methods.

6.4.1. Naïve Bayes Trees (NBTree)

NBTree is similar to the recursive schemes observed in classification trees, except that the terminal nodes contain Naïve Bayes models instead of simple nodes predicting a single class (Kohavi, 1996). The purpose of its development was to improve the accuracy of the Naïve Bayesian classifier by relaxing the assumption of conditional independence of the Naïve Bayes (Ting and Zheng, 1999). In this context, there are more chances that some of the leaves of the decision trees (from where the NB model is derived) would satisfy the conditional independency assumption since they contain very few observations (Wang et al., 2006).

First, let us define the utility of the node *t* equal to the accuracy rate of using a local Naïve Bayes to predict membership at node t[37](after a 5-fold cross-validation). Next, the algorithm described in (Kohavi, 1996) works as follows: for each predictor X_i, the utility of the split on *t* denoted by $u_t(X_i)$ is calculated by assuming that X_i is used as the split predictor and $u_t(X_i)$ is equal to the weighted sum of the utility of all its nodes (child nodes), where the weight given to child nodes is proportional to number of observations that reach the node. For numeric variables a threshold is previously computed using the same procedure as in Section 6.2.3 for a C4.5 tree.

The variable with the highest utility is denoted as X^*. If $u_t(X^*)$ is significant better than the utility of the node[38], the node *t* is split into *s* child nodes, where $s = 2$ for numeric variables and $s = number\ known\ categories$ for categorical variables; otherwise a local Naïve Bayes model is created for the current node.

For each child, the steps are called recursively until the relative reduction in error on a determined split is greater than 5% and there are at least 30 observations in the node.

The NB trees generated for both datasets trees in Table 6.11 and Table 6.12 are size one (i.e. the root node only) indicating that the utility of their root nodes was higher than the utility of X^*(i.e. the variable with the highest calculated utility of the split at the root node). In other words the accuracy of a single Naive Bayes model at the root node is higher than a Naive Bayes model at each leaf. Kohavi (1996) had already demonstrated than, in general, the NBtree algorithm induces smaller trees than C4.5, especially in large datasets like the MRC dataset.

In terms of classification performance, The ISSEMyM accuracy rates for short-medium and long LoS were 99.7% and .09% respectively with an overall accuracy rate of 90.72%. Conversely, the MRC accuracy rates for short and medium-long LoS are 68.5% and 77.6% respectively, with an overall accuracy rate of 72.84%.

[37] Using discretized data

[38] This is trying to determine whether the accuracy for a Naïve Bayes model at each leaf is higher than a single Naïve Bayes model at the current node.

Predictor	Short/medium (0.9)	Long (0.1)
Age	1173	134
First Diagnosis		
none	189	31
Category 1	640	83
Category 2	239	16
Category 3	108	7
Total	1176	137
Previous admissions	1173	134
Origin		
A&E	348	76
outpatient	332	32
Transfer	495	28
Total	1175	136
Ward		
Adult Medicine	443	66
G. Surgery	585	63
Trauma	147	7
Total	1175	136
Diagnosis		
none	71	8
Category 1	714	73
Category 2	230	42
Category 3	161	14
Total	1176	137
Num. diagnoses	1173	134
Surgical procedure		
none	534	72
Category 1	50	7
Category 2	433	36
Category 3	159	22
Total	1176	137
Transfusions		
no	1067	125
yes	107	10
Total	1174	135
Number comorbidities	1173	134
addictions		
none	211	23
drinking/smoking	661	75
both	303	38
Total	1175	136

Table 6.11: NBtree of a single node and its respective Naive Bayes model for ISSEMyM. For each variable the counts on each category are displayed

Predictor	Short (0.52)	Medium/Long (0.48)
Age		
(-inf-10.5]	519	176
(10.5-66.5]	5784	4893
(66.5-inf)	634	1217
Total	6937	6286
Gender		
Female	3635	3241
Male	3301	3044
Total	6936	6285
Previous admissions		
(-inf-2.5]	4176	5552
(2.5-3.5]	186	128
(3.5-11.5]	811	276
(11.5-inf)	1765	331
Total	6938	6287
Origin		
A&E	5768	5904
outpatient	1168	381
Total	6936	6285
Ward		
Adult Med.	4020	3717
G. Surgery	2916	2568
Total	6936	6285
Surgical procedure		
none	3771	4369
Category 1	2251	523
Category 2	668	632
Category 3	248	763
Total	6938	6287
Diagnosis		
Category 1	1741	2428
Category 2	708	2590
Category 3	4488	1268
Total	6937	6286

Table 6.12: NBtree of a single node and its respective Naive Bayes model for MRC. For each variable the counts on each category are displayed

The algorithm for both hospitals did not perform significantly better than Naive Bayes. One can see how numeric variables are causing problems, especially with the ISSEMyM dataset, since they were not split (unlike the NB tree for MRC hospital). In fact, Wang et al. (2006) demonstrated that NBtrees tends to perform poorly in datasets where numeric variables are involved. This is because the discretisation method used in the algorithm may affect the selection of split predictors.

Moreover, Naïve Bayes is relatively stable with respect to small changes to training data, but decision trees are not. Combining these two techniques might result in a method that produces more unstable classifiers than Naïve Bayes alone (Ting and Zheng, 1999), which might be the case here.

6.4.2. Logistic Model Trees (LMTs)

LMTs use the same recursive partitioning scheme as the other classification trees mentioned in this chapter to create a tree structure. In addition, the terminal nodes are logistic regression models that compute class probability estimates rather than a single classification. As it happens with most of the hybrid methods, the purpose of combining these two techniques is to boost the advantages of the two methods while trying to minimise the disadvantages. Logistic regression models are quite stable methods but with potentially high bias. On the other hand, as was previously mentioned in Section 6.2.5, classification trees are highly unstable but often with low bias (Landwehr et al., 2005). LMTs therefore aim to be a trade-off between variance and bias.

The steps for creating a LMT follow generally the same steps as the other classification trees: selecting the splitting criteria, selecting the stop criteria and finding the best tree (pruning).

A logistic regression model at the root node is constructed using the LogitBoost algorithm designed by Friedman et al. (2000). Next, the node is split according the criteria used by C4.5 (Section 6.2.3), where the split predictor and split point (for continuous variables) are computed based on information gain ratios. Logistic regression models are fitted next at the child nodes using again the LogitBoost algorithm, with the parameters of the last LogitBoost interaction at the parent node as initial parameters for the new estimation. Because a logistic regression model has been fitted at the parent node, it is reasonable to use it as a base for fitting a logistic regression at the child node. Landwehr et al. (2005) argues that parameters of the model in the parent node can conceal global influences of some of the predictors on the output variable and at child nodes, and thus, the same model can refined by taking into account local influences of the predictors.

The stopping criteria consider three aspects: firstly, only nodes with more than 15 observations are split, secondly, a split is only considered if it generates at least two child nodes with two observations each and thirdly a logistic regression model is only built if the node contains at least five observations (i.e. this a requirement of the LogitBoost algorithm).

Finally, LMTs use minimal cost-complexity cross-validation pruning, described in Section 6.2.1, to find the best tree. The pruning process becomes more important in LMTs because the LogitBoost algorithm can fit only numeric variables. Nominal variables need to be converted to binary form, increasing considerably the dimensionality of the data and it is well-known that high dimensionality increases the danger of overfitting (Landwehr et al., 2005).

For a detailed description of LogitBoost, the interested reader is referred to Friedman et al. (2000). In general terms, LogitBoost was designed to fit additive logistic regression models. The basic idea of these models is replace the linear component of the logistic regression model$\tilde{x}\beta_j$[39]by an additive component (see Equation 6.16)

$$F_j(x) = \sum_{m=1}^{M} f_{mj}(x) \tag{6.16}$$

where $f_{mj}(x)$ is a unspecified function (Hastie and Tibshirani, 1990), often referred to in Data Mining and Machine Learning domains as "weak learner"[40](Landwehr et al., 2005).

The following process is repeated *m* times (the optimal value of *M* is estimated by cross-validation): for each one of the *J* outcome categories, LogitBoost computes a response variable defined asz_{ij} that encodes the error of the current fit model on the training set, and then it fits the function $f_{mj}(x)$ by weighted least squares regression of z_{ij}. Next, the new function $f_{mj}(x)$ is added to $F_j(x)$. Finally an observation is assigned to the category with the highest $F_j(x)$value.

Table 6.13 shows the results of ten-fold cross-validation on WEKA: a tree with just the root node and its associated $F_j(x)$ functions. It seems that rather than a more complicated tree substructure, a simpler model pruned back to the root is more appropriate for the ISSEMyM dataset (i.e. better bias-variance trade-off). The accuracy rates for short-medium and long LoS are 99.7% and 0.08% respectively, with an overall accuracy rate of 89.5%. With just one node, the model can be read as simple multiple regressions. For example, the patients that are more likely to have a short-medium LoS (than long LoS) are those ones that have a first diagnosis

[39] Where $\tilde{x}$ is the vector of independent variables and β_j its associated coefficients for category *j*

[40] Trees are usually the first option as weak learners. Some examples are Lutz, (2006) and Dettling, and Bühlmann, (2003). However the LogitBoost algorithm used in LMT uses simple linear regression function as the weak learner.

under category 2 or under category 3 (i.e. cholecystitis, appendicitis, gastro reflux disease, fracture in lower legs, diabetes mellitus, gastrointestinal haemorrhage, etc.), were transferred from another healthcare facility, are being treated in the trauma ward, and have a drinking or smoking behaviour.

Predictor	Short/Medium	Long
Age	-0.01	0.01
No 1st diagnosis	-0.03	0.03
1stDiagnosis_category1	-0.05	0.05
1stDiagnosis_category2	0.31	-0.31
1stDiagnosis_category3	0.57	-0.57
Previous admissions	0.05	-0.05
A&E	-0.43	0.43
Other origin	0.3	-0.3
Adult medicine ward	-0.13	0.13
Trauma ward	0.54	-0.54
No diagnosis	-0.18	0.18
Diagnosis_category2	0.06	-0.06
Diagnosis_category3	-0.2	0.2
Num diagnoses	-0.14	0.14
No surgical procedure	0.18	-0.18
Sp_category 2	-0.42	0.42
Sp_category 3	-0.17	0.17
Transfusions	0.2	-0.2
Num comobidities	-0.04	0.04
Drinking/smoking	0.1	-0.1
cons	1.69	-1.69

Table 6.13: WEKA output for ISSEMyM using LMT algorithm

Figure 6.10 shows the tree of size 8 and its associated $F_j(x)$ functions are depicted in Table 6.15. Although LMT methodology is based on how C4.5 trees are built, the resulting LMT for

MRC dataset is a simpler model that its C4.5 predecessor (Section 6.2.3).The accuracy rates for short, and medium-long LoS are 68.6% and 78.5% respectively and the overall accuracy rate is 73.27%.

From the first terminal node containing the linear model 1 (Table 6.15), it is clear that those patients who have a diagnosis from category 1 (i.e. appendicitis, cholecystitis, gastroenteritis etc.), are young, female, treated in the general surgery ward, undergoing a surgical procedure of category 1 or 2 (peritoneal dialysis, cholecystectomy, appendectomy, open repair of hernia, etc.) and entered to the hospital via the outpatient clinic are more likely to have a short LoS.

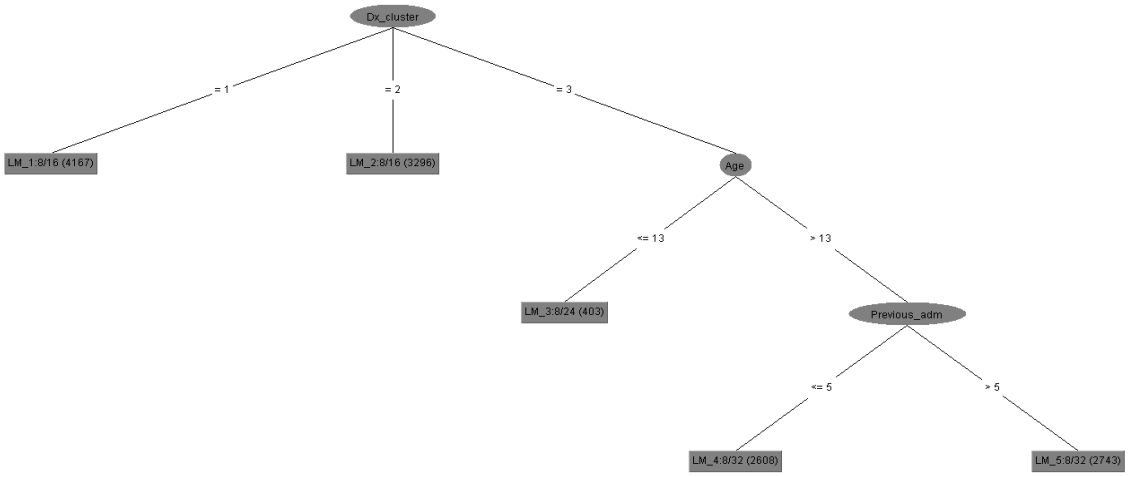

Figure 6.10: Logistic model tree for MRC dataset

	Predictor	Short	Medium-long
Linear Model 1 (LM_1)	cons	-0.08	0.08
	Age	-0.01	0.01
	Female	-0.04	0.04
	Previous admissions	0	0
	Outpatient clinic	0.28	-0.28
	G. surgery ward	0.15	-0.15
	Sp_category1	0.35	-0.35
	Sp_category2	0.07	-0.07
	Sp_category3	-0.33	0.33
	Diagnosis_category2	-0.39	0.39
	Diagnosis_category3	0.8	-0.8
Linear Model 2 (LM_2)	cons	-0.29	0.29
	Age	0	0
	Previous admissions	0.02	-0.02
	Outpatient clinic	0.28	-0.28
	G. surgery ward	0.01	-0.01
	No surgical procedure	0.04	-0.04
	Sp_category1	0.24	-0.24
	Sp_category2	-0.24	0.24
	Sp_category3	-0.71	0.71
	Diagnosis_category2	-0.39	0.39
	Diagnosis_category3	0.8	-0.8
Linear Model 3 (LM_3)	cons	-0.69	0.69
	Age	0.05	-0.05
	Previous admissions	0.02	-0.02
	Outpatient clinic	0.28	-0.28
	G. surgery ward	1.13	-1.13
	Sp_category1	-0.03	0.03
	Sp_category2	1.54	-1.54
	Sp_category3	-1.25	1.25
	Diagnosis_category2	-0.39	0.39
	Diagnosis_category3	0.8	-0.8

Linear Model 4 (LM_4)	cons	-0.62	0.62
	Age	-0.01	0.01
	Female	0.01	-0.01
	Previous admissions	0.13	-0.13
	Outpatient clinic	0.27	-0.27
	G. surgery ward	0.65	-0.65
	No surgical procedure	-0.07	0.07
	Sp_category1	0.16	-0.16
	Sp_category2	0	0
	Sp_category3	-1.25	1.25
	Diagnosis_category2	-0.39	0.39
	Diagnosis_category3	0.8	-0.8
Linear Model 5 (LM_5)	cons	0.17	-0.17
	Age	0	0
	Female	-0.09	0.09
	Previous admissions	0	0
	Outpatient clinic	0.82	-0.82
	G. surgery ward	0.9	-0.9
	No surgical procedure	0.02	-0.02
	Sp_category1	0.03	-0.03
	Sp_category2	0.6	-0.6
	Sp_category3	-1.25	1.25
	Diagnosis_category2	-0.39	0.39
	Diagnosis_category3	0.8	-0.8

Table 6.14: Linear model functions associated to the logistic model tree for MRC

6.5. Ensemble Methods

Although all the previous algorithms have shown a moderately good performance in predicting LoS category at the MRC hospital, there are serious concerns about the performance on the ISSEMyM dataset where all the algorithms have failed drastically to predict long LoS.

Aiming to increase the accuracy rates, especially the long LoS category, a different type of classification algorithms named ensemble methods are presented in this section. Ensemble methods are a collection of individual classification algorithms which have been proved to be

more accurate than the individual algorithms that compound them (Dietterich, 2000a). There are many methods to construct ensembles, but the most popular is via manipulation of the training dataset and two main techniques called Bagging and Boosting. Both techniques take a base algorithm and invoke it many times in different training datasets in order to increase the algorithm's accuracy.

The algorithms presented so far can be referred to as white box systems because their inner structure and variables that are involved are available for inspection and interpretation. Bagging and Boosting are more complex in that aspect, although they cannot be referred completely as a black box system. However they are used in this research as a last resort where interpretation will be sacrificed on behalf of higher accuracy.

The name of Bagging stands for "Bootstrap aggregating" developed by Breiman (1996a).It uses bootstrapping to resample the dataset. For each of the *t* interactions, bagging samples *n* observations with replacement from training data. The result is a training set of the same size as the original, but some of its observations may not appear in it while others appear more than once. Next, the classification algorithm is applied to each of the *t* datasets and the resulting models stored. Subsequently, the class or outcome category of each observation in the original dataset is predicted using each one of the *t* models. Finally, the outcome category which was predicted most often for a specific observation is used to classify that observation.

Whereas in Bagging individual models are built separately, in Boosting, each new model is influenced by the performance of those built previously. Boosting implemented in WEKA uses the AdaBoost algorithm: equal weights are assigned to each training observation i. For each of *t* interactions the classification algorithm is applied to weighted dataset and the resulting model stored. The error *e* of the model on the weighted dataset is computed and if e=0 or $e \geq 0.5$ the model generation terminates; otherwise each correctly classified observation in the dataset multiplies has its weight multiplied by$\frac{e}{1-e}$. Next, the weights of all observations are normalised (i.e. each observation's weight is divided by the sum of the new weights and multiplied by the sum of the old ones). This automatically increases the weight of each misclassified observation and reduces it relative to the correctly classified ones. In every interaction, Boosting focuses on the observations with the higher weights (i.e. the incorrectly classified observations) enabling later models to become experts for observations classified incorrectly by earlier ones.

When the model generation finalises (i.e. e=0 or $e \geq 0.5$) the weights of all the observations are set to zero. Afterwards, the outcome category of observation$i$ is predicted using each one of the *t* models and a new weight, equal to $-log\left(\frac{e}{1-e}\right)$, is assigned to each observation, where *e* is the error of model *t*. Finally, for each observation i, all the weights across the *t* models that were

predicted with a particular class C_j are summed up and the class with the total highest weight is selected as the predicted outcome category for observation i. This algorithm is known as "Boosting by weighting". Another approach is "Boosting by sampling" which selects observations with replacement from the dataset with probability proportional to their weights (Dietterich, 2000b) .

Moreover, Bagging and Boosting have very different profiles: Breiman (1996a) comments that a critical factor in whether Bagging will improve accuracy is the stability of the base algorithm: *"Improvement will occur on unstable methods. On the other hand it can slightly degrade the performance of stable procedures"*. As we mentioned before trees are very unstable methods, therefore Bagging will be applied to C4.5 and CART. Conversely Boosting requires some instability too, but to a lesser extent, because if the error on *t* is equal or greater than 0.5, the model generation will stop earlier (Quinlan, 1996). On the other hand, Boosting requires that the individual performance of the base algorithm should have an accuracy rate higher than 50% (Freund and Schapire, 1996). Therefore Boosting was applied to Naives Bayes, NBtrees and logistic regression which are more stable methods. The number of interactions *t* was set to 10 according to Breiman (1996a) who states that most of the improvement is reached with that number.

Table 6.15 shows accuracy rates using 10-fold cross-validation. The accuracy rates of both datasets were compared against those generated by the base algorithm alone (e.g. Boosted CART vs. CART) using a T-test to compare the means (Table 6.16). In most of the cases for ISSEMyM, there was no significant difference between the accuracy rates generated by the base algorithm alone and the ensemble method. The exceptions were NB Tree and C4.5: on average the Boosted NB Tree performs worse (86.56%) than the NB tree alone (89.81%), and this difference was significant $t(198) = 14.18, p < 0.001$. Conversely, the Bagged C4.5 has a higher accuracy rate (89.76%) than the C4.5 tree alone (89.27%), and this difference was significant $t(198) = -4.13, p < 0.001$. It seems that the high instability of C4.5 boosted the performance of the bagged version.

In the case of MRC, none of the ensemble methods had a statistically significant better performance than their counterparts.

Ensemble method	Base algorithm	Accuracy rates for ISSEMyM			Accuracy rates for MRC		
		Short-medium	Long	Overall	Short	Medium-long	Overall
Boosting	MNLM	100%	0%	90.07%	71.46%	75.00%	73.14%
	NBTree	96.61%	3.38%	87.78%	65.41%	80.09%	72.39%
	Naïve Bayes	92.37%	7.62%	83.20%	59.94%	86.46%	72.54%
Bagging	C4.5	99.15	0.84%	89.31	65.70%	81.21%	73.07%
	CART	100.0%	0%	90.07	68.15%	78.34%	72.99%

Table 6.15: Accuracy rates for ensemble methods

Algorithms to compare			t	df	Sig.(2-tailed)
Logit	vs.	Boosting Logit	.000	198	1.000
Naive Bayes	vs.	Boosting Naive Bayes	-1.742	198	.083
NBTree	vs.	Boosting NBTree	14.187	198	.000
CART	vs.	Bagging CART	1.304	198	.194
C4.5	vs.	Bagging C4.5	-4.138	198	.000

Table 6.16: T-test results to compare means of base algorithms and ensemble methods for ISSEMyM

Although Boosting was a promising method to increase the accuracy of long LoS since it specialises in the observations that are more difficult to categorise, it failed on the ISSEMyM dataset. Freund and Schapire (1999) mentioned that when the number of outliers (i.e. observations that are atypical, incorrectly recorded or ambiguous) is very large, the emphasis placed on the hard examples can be detrimental to the performance of Boosting. A quick glance at the data revealed that both datasets contains a considerable number of ambiguous observations: for example, it was found in the MRC dataset, a considerable percentage of patient records had equal values on their 9 predictors, however they were assigned to different LoS category: 464 (out of 705) patient records classified as medium-long LoS have at least one "twin" record (same values in all the predictors) classified in the short category. Moreover it is very common to find "triplets" records (i.e. three identical records assigned to different LoS). In total approximately 47% of the patient records are ambiguous outliers, restraining the inference process and causing the classification algorithms to be unable to discriminate categories correctly. In fact, the relative good performance of the previous algorithms is quite unexpected

given the high percentage of ambiguous observations; especially Boosting which is particular sensitive to large amounts of outliers.

A possible solution would be to collect more data. There may be other variables that have not been recorded, but have an influence on LoS category. It might be the case that the information contained in the datasets is not enough to predict accurately LoS category. On the other hand, it is quite plausible that those patients, who represent this type of outliers, indeed present with the same characteristics as other patients, until something unpredictable and immeasurable happens to them that affects their length of stay. More on how to deal with ambiguous outliers is described in Section 9.2.2.

Finally, there are more complicated ensemble models available such as Multiboosting and Stacking, which can be employed. MultiBoosting combines AdaBoost with wagging, a variant of Bagging. On the other hand, Stacking is an ensemble method which uses different types of classification algorithms (unlike Boosting and Bagging that combine models of the same type). However a necessary condition for an ensemble of classification algorithms to be more accurate than any of its individual members is that the classifiers are accurate and diverse. Two classifiers are diverse if they make different errors on new observations (Dietterich, 2000a)and it has become evident that the classification algorithms presented in this research cannot be considered diverse (i.e. accuracy rates per category are almost identical). Therefore it is doubtful that any benefit can be obtained by trying these more complicated techniques.

6.5.1. Discussion

In this chapter focusing on the group-based approach, thirteen different data mining techniques were explored: one logistic regression model, four classification trees, one Bayesian network, two hybrid methods and five ensemble algorithms.

Before deciding which model is more appropriate for each hospital, it is important to ensure that the accuracy rates, in which the choice of model will be based, are not just product of the process of how the data is split for training (testing) purposes. Table 6.17 summarizes the accuracy rates of 10 x 10 cross-validations:

Algorithm	Accuracy rates for ISSEMyM			Accuracy rates for MRC		
	Short-medium	Long	Overall accuracy rate	Short	Medium-long	Overall accuracy rate
Naive Bayes	92.96%	14.07%	84.92%	61.69%	84.07%	72.33%
C4.5	99.23%	1.51%	89.27%	66.65%	80.25%	73.11%
NBTree	99.99%	0.15%	89.82%	68.55%	77.65%	72.88%
Bagged C4.5	99.83%	0.98%	89.76%	67.57%	78.94%	72.97%
Bagged Cart	99.91%	0.83%	89.81%	69.32%	76.38%	72.67%
Boosted NBTree	94.98%	12.51%	86.57%	68.71%	77.40%	72.84%
Boosted Logit	99.88%	1.80%	89.89%	72.92%	72.16%	72.56%
Boosted Naive Bayes	93.69%	13.85%	85.56%	61.69%	84.07%	72.33%
LMT	99.94%	0.43%	89.80%	68.57%	78.60%	73.34%
CHAID	99.67%	0.02%	89.82%	73.03%	71.91%	72.8%
QUEST	99.99%	0.01%	89.82%	64.80%	79.60%	71.90%
Logit	99.88%	1.80%	89.89%	72.92%	72.16%	72.56%
CART	99.79%	0.90%	89.71%	68.53%	77.43%	72.76%

Table 6.17: Classification algorithms accuracies for ISSEMyM and MRC

The results for ISSEMyM show that in general the accuracy rates are sufficiently similar that their differences are in practical terms insignificant. The exception are Naive Bayes, Boosted Naive Bayes and Boosted NBtree, whose overall accuracy rates seem to be behind the other methods; however, when using those algorithms, there was a significant improvement on the accuracy rate of long LoS (at the expense of the accuracy rate in short-medium LoS). Similar behaviour was found with the MRC dataset, where the overall accuracy rates are similar among the algorithms. Nevertheless, an ANOVA analysis confirmed statistically significant differences. These differences are very small in practical terms, since any of them represents an increment (or decrement) of more than 1% in respect to the mean LoS. However there are important differences (in statistical and practical senses) within categories. Naive Bayes and Boosted Naive Bayes were the best classifiers of medium-long LoS category, but at the same time they were the worst on classifying short LoS.

It seems that all the data mining algorithms explored here have reached the limit of information that can be predicted and there is no room for future improvement. Lim et al. (2000)demonstrated in his experiment comparing 33 classification algorithms on 32 different datasets, that the performance of many of the algorithms are very similar among them with differences statically insignificant.

On the other hand, as the reader might have noticed, this chapter started with the purpose of finding the best technique to classify patients into two groups based on their attributes, emphasising the role that such attributes hold on LoS. However as the chapter was developed and in response to the poor performance of most of the algorithms to predict long LoS patients (ISSEMyM hospital only), other more complicated models were explored aiming to increase accuracy rates, while sacrificing simplicity and valuable understanding of how patient attributes influence LoS[41].

Nevertheless, the main objective should not be ignored and simplest models should be always preferred over hybrids or ensemble counterparts. Therefore, for the MRC dataset, Logit and CHAID were the techniques which combine high accuracy rates per category, high overall performances (all over 70%) and ease of interpretation. Although CHAID presents the advantage of being a friendlier model for non-experts, Logit is a simpler and more stable option, which makes it the preferred model for the MRC hospital. Contrary to what happened in the individual-based approach, the models discussed here (for the MRC hospital only) were by far more successful in discriminating between the two LoS homogenous categories. In this context, it is possible to make an estimation about the expected LoS or about the LoS distribution for a specific cohort of patients (LoS category) using the density function of the category s i.e. $f_s(y_i)$ (more on this in Chapter8).

Conversely for the ISSEMyM hospital dataset the best option is Naive Bayes, although its performance was mediocre. In addition to the presence of ambiguous outliers (discussed in Section 6.5), the size of the ISSEMyM dataset (i.e. 1300 records) and the small proportion of patients with Long LoS (i.e. 18.0%) may be possible causes of the poor performance of the classification algorithms. According toMorgan et al. (2003),in general any algorithm's accuracy tends to increase (at a decreasing rate) with increments in sample size. Moreover, the size of the sample needed to achieve the maximum potential performance of an algorithm is likely to vary across datasets and the nature of the outcome variable. In this context, it is possible to calculate the optimal size of the sample needed for each algorithm of this research(Bull, 1993), however the reader might remember that accessing and collecting data from the ISSEMyM hospital is a

[41] These types of models are usually known as black box models, where the models can be viewed solely in terms of input and output without any knowledge of internal working.

very expensive procedure(see Section 3.3), which complicates any attempt to increase the sample size.

Therefore, if the group-based approach is applied to the ISSEMyM hospital, there is a very high risk of misclassifying a patient. However if the ISSEMYM hospital acknowledges that approximately 18.0% of its admissions would be long LoS and it allocates the right number of resources to treat those patients without affecting the resources allocated for short-medium patients, then misclassifying a long LoS patient as short-medium should not represent a serious problem for the hospital in terms of bed management related issues. However for estimation of LoS distribution or expected LoS for ISSEMyM patients, an individual-based approach using finite mixture regression is recommended. As was mentioned previously, the finite mixture regression would minimise the risk of incorrect estimations of LoS, because the estimated LoS probabilistic curve (based on the mixture regression density equation) would contain an element from both categories (components) but in different proportions.

6.6. Summary

In the group-based approach all patients within LoS categories are treated as equal entities. Although their individual characteristics helped to predict the LoS category, their LoS probabilistic curve and associated expected value of LoS was defined by the parameters of the category itself (defined previously in the last chapter).

Some of the most common data mining prediction techniques, including Logit regression, decision trees (CART, QUEST, C4.5 and CHAID), Naive Bayes and hybrid methods (Naive Bayes trees and Logistic Trees) were evaluated to find the best method to predict LoS category based on patient characteristics.

Later, in response to a poor performance of all the models mentioned above to predict long LoS category at ISSEMyM hospital, other more sophisticated data mining techniques were incorporate to the approach, such as Bagging and Boosting, which are commonly referred as ensemble methods. However the results did not indicate a significance difference on their performances from the previous group of algorithms.

In order to provide a comparative point, the accuracy rates per category and overall performances of all the models were verified using 10×10 cross-validations. The results for ISSEMyM showed that in general the accuracy rates are sufficiently similar among all the models that their differences are in practical terms insignificant. Similar behaviour was found on the MRC dataset where the overall accuracy rates are similar among the algorithms, although

some important differences within categories were found in the Naive Bayes and Boosted Naive Bayes.

Finally, the good performance of Logit regression, in addition to its characteristics of simplicity and stability, made it the preferred model for the MRC hospital. On the other hand, the least bad option for the ISSEMyM hospital was the Naive Bayes. However in this case, the finite mixture regression seems to be a more appropriate choice.

7 MULTILEVEL GROUP-BASED APPROACH

So far in Chapters 5 and 6, just internal factors and patient attributes have been analysed in order to identify their relationship and influence on LoS. Moreover the models used for this task have allowed making estimations of the LoS distribution for individual patients and cohort (groups) of patients within hospitals.

On the other hand, in Chapter 1 the complex structure of Mexican healthcare systems was exposed as the cause of enormous heterogeneity between the major services providers, complicating the development of a common plan of action to improve the performance of the services. Such heterogeneity suggests that beyond patient characteristics and internal factors LoS can vary from one hospital to other or from one healthcare provider to other.

In this context, Leyland and Goldstein (2001) state that health outcomes (like patient LoS) vary between different institutions for at least three reasons:

- Differences may be attributable to random variations
- Institutions may differ systematically with respect to the care they provide
- The health of their respective patient population may differ prior the admission(e.g. the socio-economic level of the population treated in SSA hospitals is lower than in other institutions

Frick et al. (1999) argues that patient LoS is influenced by factors related to the patient, disease treatment, hospital and the health system. Also, in a large study conducted in England by Martin and Smith (1996), it was concluded that LoS variation is driven by patient characteristics, the local supply of National Health Service (NHS) care, the local pressure on NHS resources, other supply factors and local policy variations.

Consequently there is a need for a tool to help in the decision-making process, which effectively incorporates the heterogeneity of the Mexican healthcare system and vice versa.

The aim of this chapter is to provide an extension of the models built in previous sections, in order to understand the environment in which the patient is treated and how this affects LoS, while providing a model that adapts itself from a local level (hospital) to a regional or institutional level. Specifically, the group-based approach models, which cluster patients into LoS homogenous groups, will be extended to account for external and environmental factors that might influence the patient classification between hospitals into homogenous groups.

7.1. Extended Logit Model

Let us consider both hospitals that have been studied so far: ISSEMyM and MRC. The hospitals are sited in different geographical regions of Mexico, they are controlled by different healthcare providers (ISSEMyM belongs to the State Services and MRC belongs to Secretariat of Health) and they treat different sectors of the population (ISSEMyM treats mainly social security holders whereas MRC is open to the population that cannot afford private care). This is why during initial analysis of the LoS in Section 3.5, it was clear that they have a very distinct nature, and therefore separated models were developed for each hospital in Section 4.1.3

The next table depicts a logistic model applied now to a dataset containing both hospitals and including an extra variable to account for the hospital where the patient is treated. Let us recall that the logistic regression model was a relatively successful tool to predict the LoS category for both hospitals separately (see Section 6.4.2 and 6.5.1). From Table 7.1 it can be noticed that the hospital variable was highly significant to predict LoS category, just after diagnosis and surgical procedure variables. The odds indicates that a patient admitted at ISSEMyM hospital is more likely (2.1 times more) to have a medium-long LoS than a short one[42].

[42]This interpretation is true only if the effects of the other variables are held constant.

	β	Std. Err.	z	P>z	[95% Conf.	Interval]
Age	0.0099	0.0012	8.250	0.000	0.0076	0.0123
Previous admissions	-0.0285	0.0034	-8.430	0.000	-0.0351	-0.0218
Outpatient clinic	-0.5825	0.0695	-8.380	0.000	-0.7187	-0.4463
Other origin (Transfer)	-0.9321	0.1400	-6.660	0.000	-1.2065	-0.6577
General surgery ward	-0.4477	0.0575	-7.790	0.000	-0.5605	-0.3350
Diagnosis_category1	-0.2703	0.3767	-0.720	0.473	-1.0086	0.4679
Diagnosis_category2	0.3883	0.3804	1.020	0.307	-0.3573	1.1340
Diagnosis_category3	-1.5694	0.3823	-4.100	0.000	-2.3188	-0.8200
Sp_category 1	-0.3815	0.0862	-4.430	0.000	-0.5504	-0.2127
Sp_category 2	-0.0132	0.0664	-0.200	0.842	-0.1434	0.1170
Sp_category 3	1.0135	0.0828	12.250	0.000	0.8513	1.1758
Hospital_ISSEMyM	0.7503	0.1040	7.210	0.000	0.5464	0.9542
cons	0.5134	0.3870	1.330	0.185	-0.2452	1.2720

Table 7.1:Logistic model for MRC and ISSEMyM hospitals

So far just two hospitals have been considered; but when a strategic level of planning is being conducted the sample of hospitals to be considered should be larger because the aim is to make an inference about the population. It may be of interest to compare the hospitals of an entire region, or all the hospitals of a specific healthcare provider. In that case the logistic regression model like the one above is unsuitable for mainly four reasons:

1. It was mentioned before that logistic regression models are based on the assumption of independency across observations, and previous models (logistic and mixture regression model) were designed to relax the assumption with observations with the same patient file number. However there might be other forms of dependency: patient records within a specific hospital could have more similarities than those from a different hospital.
2. Contextual variables cannot be included in the logistic model. A contextual variable describes the environment, for example, the level of poverty of the region where the hospital is placed. According to Mexican office of National Statistics[43], patients treated in hospitals placed in poorer regions tend to have longer LoS. Other contextual

[43] Instituto Nacional de Estadistica y Geografia, website www.inegi.org.mx

variables could be type of policy, level of education in the region, level of urbanisation of the region, number of beds in the hospital, number of operating rooms, etc. With just two hospitals included in this research, these variables would be perfectly correlated, breaking the assumption of no multicollinearity.

3. If the subjects of study are *n* hospitals, $n - 1$ dummy variables need to be included in the logistic model making it very complex as *n* increases.
4. The hospital data has actually what in statistics is called "hierarchical structure" where there are patients clustered in hospitals and hospitals clustered in regions or healthcare providers and so on. Therefore the outcome variable has both an individual and a cluster aspect to its variability; in addition, the explanatory variables may also contain a cluster aspect (Channon, 2010), and therefore the full variability of the system cannot be fully captured by including explanatory variables.

In this context, multilevel models, developed in the 1980's, were found in this research suitable to overcome some of these limitations.

7.2. Multilevel Analysis

Hierarchical analysis or multilevel analysis is a methodology to model the dynamics between clustered data, such as pupils in classes, employees in companies, individuals in communities, patients at hospitals, etc.

In the past 20 years, a considerable number of multilevel models have been developed using healthcare data to analyse situations such as geographic variations in practice patterns, contextual variation in health behaviours, variation in the performance of healthcare institutions, etc. In the context of LoS, Martin and Smith (1996) analysed variations in LoS for acute inpatient care across 4,585 small areas in England using hierarchical analysis. Leung et al. (1998) aimed to identify the factors related to pregnancy and childbirth that predict LoS after delivery using the same technique. At almost the same time, Frick et al. (1999) used multilevel analysis to compare the influence of individual and organizational variables on LoS in a large psychiatric hospital in Germany. Carey (2002) explored the effects of decreasing LoS on hospital costs. Urbach and Austin (2005) used a three level model to explain length of postoperative hospital stay for three major surgical procedures in Ontario, Canada. Jong et al. (2006) used multilevel analysis to test whether physicians working in different hospitals adapt their LoS decisions to what it is the common practice in the hospital under consideration, etc.

A hierarchical structure of clustered data consists of units grouped at different levels: in a two-level model, at level-1, the units are patients, and each patient's outcome (LoS or LoS category)

is represented as a function of a set of individual characteristics. At level-2, the units are organizations (healthcare providers, hospitals, etc.) This type of analysis takes into account the fact that the variability associated with each level of clustering, which if ignored, as happens with classical methods, may lead to wrong conclusions (Snijders and Bosker, 1999). Moreover, the outcome variable has both an individual and a cluster effect to its variability. In this context, there are two ways in which the cluster aspect influences the level-1 unit's outcome. In the first option, some characteristics of the level-2 units exert a common influence on each person within it. Such a cluster effect modifies only the mean level of the outcome for a specific level-2 unit, leaving unchanged the distributions of effects among persons within the level-2 units. In statistical terms, only the intercept, $\beta_{0j,}$ varies across level-2 units, all other level-1 parameters remaining fixed. This basic model is referred to as the random intercept model.

In the second possibility, the cluster effect may modify both the mean level of outcomes and how the effects of other variables are distributed among individuals. In statistical terms, both the intercept and slopes (explanatory variable parameters) vary among the units. This is the full hierarchical linear model with random intercepts and slopes.

The multilevel model discussed in this section is an extension of the logistic regression model. For the basic linear multilevel model, the reader is referred to Snijders and Bosker (1999) and Goldstein (2011)

Recalling that the binary logistic model is defined by Equation 7.1 and 7.2:

$$f(y = 1; \vartheta) = \frac{1}{1 + e^{-\vartheta}} \tag{7.1}$$

$$\vartheta = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p,} x_p \tag{7.2}$$

The extension to a two-level random intercept model is described by the outcome for patient *i* in hospital *j* as y_{ij} (see Equations 7.3-7.6):

$$f(y_{ij} = 1; \vartheta_{ij}) = \frac{1}{1 + e^{-\vartheta_{ij}}} \tag{7.3}$$

$$\vartheta_{ij} = \beta_{0j} + \beta_1 x_{1ij} + \cdots + \beta_{p,} x_{pij} \tag{7.4}$$

$$\beta_{0j} = \beta_0 + u_{0j} \tag{7.5}$$

$$u_{0j} \sim N(0, \sigma_{u0}^2) \tag{7.6}$$

Notice that the intercept β_{0j} consists of two terms: a fixed component and a hospital-specific component called the random effect (residual at level-2) u_{0j} which is normally distributed with a mean of zero and a variance of σ_{u0}^2.

σ_{u0}^2 is also known as between-group variance, it is the variance of β_{0j}, when $\boldsymbol{x_{ij}} = 0$, and it can be used to calculated the variance partition coefficient (VPN), which is the fraction of total variability explained by the cluster structure (more on this later).

Notice that unlike the standard logistic regression model (as the one described in Section 7.1), the two-level model requires the estimation of only two extra parameters (u_{0j} and σ_{u0}^2) regardless of the number of units *j* under consideration.

To extend the model to a random intercept and random slope is shown in Equations 7.7-7.13:,

$$\vartheta_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \cdots + \beta_{p,} x_{pij,} \tag{7.7}$$

$$\beta_{0j} = \beta_0 + u_{0j} \tag{7.8}$$

$$\beta_{1j} = \beta_1 + u_{1j} \tag{7.9}$$

$$u_{0j} \sim N(0, \sigma_{u0}^2) \tag{7.10}$$

$$u_{1j} \sim N(0, \sigma_{u1}^2) \tag{7.11}$$

$$cov(u_{0j}, u_{1j}) = \sigma_{u01} \tag{7.12}$$

$$var(u_{0j} + u_{1j}) = \sigma_{u0}^2 + 2\sigma_{u01} x_{ij} + \sigma_{u1}^2 x_{ij}^2 \tag{7.13}$$

Notice that the slope β_{1j}, along with the intercept, consists of a fixed component and a hospital-specific component. On the other hand, Equations (7.10) and (7.11) state the same normality assumption must be met for the level 2 residuals; Equation (7.12) confirms that the slope and the intercept are non-independent and their relationship is described by their covarianceσ_{u01}; and finally, Equation (7.13) gives the total variability at level 2, which now depends on the value of $\boldsymbol{x_{ij}}$, which is the vector containing the attributes of patient *i* at hospital *j*.

Equations (7.4) and (7.7) can be rewritten to accommodate contextual variables, defined as $z_{1j},\ z_{2j}, \ldots, z_{Qj}$ which describes the environment of the clustering aspect. Notice that contextual variables have only subscript *j*, indicating that each individual *i* in group *j*, has the same value ofz_{qj}.

At level 1, the model is based on the same assumptions as the basic logistic regression model: no multicollinearity. At level 2, the model is based on the assumptions as the linear regression model, discussed in Section 5.1.1, independent, normally distributed standard residuals and homogeneity of the variance. Therefore, a standard analysis of the residuals needs to be carried out at level 2 to check the adequacy of the model.

The estimation of the model is done by quasi-likehood methods which firstly linearise the non-linear model and then apply iterative algorithms to obtain maximum likelihood estimates of the coefficients and variance. For more details, the reader is referred to Goldstein (2011)

7.2.1. Building the Model

Unfortunately it is not possible to perform multilevel analysis with the current dataset containing information from two hospitals only. According to Snijders and Bosker (1999), the requirements for the sample size at the highest level of the model are at least as stringent as requirements on the sample size in a single level model. Instead, the multilevel analysis will be carried out on a different dataset containing patient records of 19 Mexican hospitals from year 2005.The 19 hospitals of the new data set are very similar to the MRC hospital: they are located in the in State of Mexico, they belong to the Secretariat of Health. In addition, they offer second-level type of care (see Section 1.4.1) and they are open to the general population.

Because, the variable of interest remains LoS category which corresponds to a two-component Lognormal finite mixture model (according to the results in Section 4.1.2); it would be desirable if the new dataset is accurately described by the same finite mixture model. Let us start with the hypothesis that the new dataset, named regional dataset, and the dataset from ISSEMyM and MRC come from the same population and therefore their corresponding LoS distributions can be accurately modelled by the two-component Lognormal finite mixture model described earlier. To test the hypothesis, a Kruskal-Wallis test was computed, which is a non-parametric test, usually presented in literature as an alternative to ANOVA when the assumption of normality or equality of variance is not met. It tests the null hypothesis that the samples of two or more groups come from identical populations. The results of a Kruskal-Wallis test carried out using SPSS indicates that the null hypothesis cannot be rejected: $H(1) = -585654,\ p = 1.00$, supporting the use of the existing LoS category parameterisation (a two-component Lognormal finite mixture model) on the regional dataset.

The regional dataset contains at level 1 the variables that are already familiar to the reader: patient age, number of previous hospital admissions, origin of the patient, ward where the patient is treated, diagnosis, surgical procedure and patient gender. In addition, Table 7.2 describes the contextual variables available in the regional dataset.

Contextual variables	Description
High level poverty	Indicates if the hospital is located in a municipality with high level of marginalization
Number of consultation rooms	Number of consultation rooms available on the outpatient clinics at hospital
Hospital size according to the number of beds	The size of the hospital is determined by its number of beds, e.g. <60 beds, 60-120 and >120 beds
Ratio medical staff-patient	Number of medical staff per patient
Ratio nursing staff-patient	Number of nurses per patient
Number of operating theatres	Number of operating rooms available at hospitals
Bank of blood	Indicates whether the hospital has its own bank of blood
High technology medical equipment	Indicates whether the hospital have at least one unit of high technology equipment (MRT, MRI, electroencephalogram, mammography equipment, lithotripter, etc.)
Hospital municipality	Municipality where the hospital is geographical located

Table 7.2: Contextual variables added to the multilevel analysis of the regional dataset

All the numerical variables were centred on their means; this procedure was adopted because multilevel models allow drawing conclusions for a specific baseline scenario, where all the variables, except the ones with random effects, have a value equal to zero. However, there are variables for which the value of zero makes little sense, such as a patient age or a hospital without any operating theatres.

The estimation of the model was carried out using MLwiN® software. The first step is to build a one-level logistic regression model, which is the standard logistic model. Table 7.3 depicts the parameter estimates of the model and their corresponding standard errors in brackets. Unlike STATA, MLwiN does not indicate whether the variables in the model are significant. This can be tested by calculating the *z-ratios*, e.g. $\beta/SE(\beta)$, which can be compared with a standard normal distribution where values greater than |1.96| yields to a $p-value < 0.05$, indicating significance of the parameter. Notice that all the variables in the model have *z-ratios* with magnitude greater than 1.96.

	One-level model		
	Predictor	$\boldsymbol{\beta}$	**Std. Err.**
Fixed part	**Age**	0.01	0.001
	Female	0.059	0.023
	Previous admissions	-0.055	0.003
	Outpatient clinic	-0.56	0.036
	Other origin (Transfer)	-0.294	0.062
	General surgery ward	-0.646	0.029
	Diagnosis_category2	0.686	0.032
	Diagnosis_category3	-0.908	0.031
	Sp_category 1	-0.535	0.042
	Sp_category 2	0.398	0.029
	Sp_category 3	1.384	0.05
	cons	0.255	0.04

Table 7.3: Parameter estimates of the one-level model for the regional dataset

Notice that the results of the model can be interpreted in terms of odds ratios (see Section 6.1). For the time being, the interpretation of the parameters will be reserved for the final model. The next step is to extend the standard logistic model to allow for hospital effects on the dependent variable. The random intercept model allows the intercept of the model across hospitals to vary. Table displays the parameters estimates and standard errors. Notice that the estimates are slightly different that the one-level model depicted on Table 7.3, as they are now accounting for the hierarchical structure. Although the z-ratios of all the variables confirm their significance; the concern is to test if the additional parameter is significant, i.e. if the hospital level variance (σ_{u0}^2) is significant. This can be done by calculating the Wald test, which works in a similar way that likelihood ratio test[44] defined in Section 6.1, it works on the null hypothesis that the parameters to be tested are equal to zero (i.e. no effect). The result of the test indicated that the effect of the hospital level variance is significant at the 0.05 level ($Wald\ X^2 = 8.13, df =$

[44]Since the multilevel logistic model is estimated by quasi-likelihood methods, the likelihood ratio test is not appropriate to test the parameters significance.

$1, p < 0.05$)[45].This suggests that there are significant differences across hospitals in the way in which the probabilities of LoS category=1 (i.e medium-long) are assigned by the logistic model, once the other variables are held constant.

	Two-level model with random intercept		
Fixed part	**Predictor**	$\boldsymbol{\beta}$	**Std. Err.**
	Age	0.009	0.001
	Female	0.046	0.023
	Previous admissions	-0.052	0.003
	Outpatient clinic	-0.635	0.042
	Other origin (Transfer)	-0.725	0.074
	General surgery ward	-0.589	0.03
	Diagnosis_category2	0.708	0.032
	Diagnosis_category3	-0.966	0.033
	Sp_category 1	-0.541	0.044
	Sp_category 2	0.222	0.034
	Sp_category 3	1.291	0.051
	cons	0.31	0.103
Random part		**Coeff.**	**Std. Err.**
	$\boldsymbol{\sigma^2_{u0}}$	0.152	0.053

Table 7.4: Parameter estimates for two-level model with random intercept. Where σ^2_{u0} is the hospital-level variance.

Now it is possible to add the environmental factors. Therefore, the next step is to add one by one the contextual variables and to test its significance with the Wald test. Table 7.5 shows the parameters estimates of the model with random intercept and the variable "number of operating theatres" as the only contextual variable from Table 7.2 that was significant ($Wald\ X^2 = 30.05, df = 1, p < 0.0001$).

[45]When testing random effects, a one-sided test should be used because the variance term per definition is always positive.

	Two-level model with random intercept and contextual variables		
Fixed part	**Predictor**	$\boldsymbol{\beta}$	**Std. Err.**
	Age	0.009	0.001
	Female	0.047	0.023
	Previous admissions	-0.052	0.003
	Outpatient clinic	-0.637	0.041
	Other origin (Transfer)	-0.724	0.074
	General surgery ward	-0.59	0.03
	Diagnosis_category2	0.707	0.032
	Diagnosis_category3	-0.966	0.033
	Sp_category 1	-0.537	0.044
	Sp_category 2	0.228	0.034
	Sp_category 3	1.294	0.051
	Number of operating theatres	0.306	0.056
	cons	-0.616	0.182
Random part		**Coeff.**	**Std. Err.**
	$\boldsymbol{\sigma^2_{u0}}$	0.052	0.019

Table 7.5: Parameter estimates for two-level model with random intercept and contextual variables

So far it has been assume that the only variation between hospitals is their intercepts, and the coefficients of the rest of the variables (slopes) remain constant between hospitals. However different hospitals might have different ways of running their wards. The management of general surgery or adult medicine wards might be different from one hospital to another and this might have an effect on the patient LoS. Table 7.6 shows the parameters estimates of the model with random intercept and slope. Let us remember that the intercept and the coefficient of general surgery ward are allowed to vary across hospitals. They are both comprised by a fixed and a random part. The random effects follow normal distributions with mean zero and variance σ^2_{u0} and σ^2_{u6}, respectively. Within a hospital, the relationship between slope and intercept is explained by the covariance σ^2_{u16}. The results indicate that the addition of the extra parameter was significant ($Wald\ X^2 = 7.17, df = 2, p < 0.05$), concluding that the effect of the ward where the patient is treated does indeed differ across hospitals.

	Two-level model with random intercept and slope		
Fixed part	**Predictor**	$\boldsymbol{\beta}$	**Std. Err.**
	Age	0.01	0.001
	Female	0.051	0.023
	Previous admissions	-0.049	0.003
	Outpatient clinic	-0.658	0.042
	Other origin (Transfer)	-0.691	0.075
	General surgery ward	-0.535	0.102
	Diagnosis_category2	0.721	0.033
	Diagnosis_category3	-0.981	0.034
	Sp_category 1	-0.547	0.045
	Sp_category 2	0.22	0.036
	Sp_category 3	1.297	0.052
	Number of operating theatres	0.273	0.051
	cons	-0.556	0.184
Random part		**Coeff.**	**Std. Err.**
	$\boldsymbol{\sigma^2_{u0}}$	0.141	0.051
	$\boldsymbol{\sigma_{u06}}$	-0.125	0.05
	$\boldsymbol{\sigma^2_{u6}}$	0.157	0.059

Table 7.6: Parameter estimates for the two-level model with random intercept and slope. Where σ^2_{u0} and σ^2_{u6} are the intercept and slope variances respectively and σ_{u06} is the covariance.

Before interpreting the parameters estimates of the final model depicted on Table 7.6, let us recall that the multilevel model, like all statistical models, is based on a number of assumptions which need to be evaluated, in order to validate the results of the model. These assumptions state that the group residuals u_j are independent between groups (i.e. level 2 units) and that they are normally distributed with a constant variance (which is often referred to as homoscedasticity).

Figure 7.1 depicts the residuals u_{0j} and u_{6j} (i.e. random part of intercept and slope respectively), which do not show any particular pattern, supporting the assumption of independency of the residuals. However, Figure 7.2 shows some important deviations from normality for both types of residuals, and in Figure 7.3 the points seem to be more spread out at

the centre and right side of the graphs, which could indicate violation of the homogeneity of variance assumption. The problems could be being caused due to the small size of the sample at level 2, however, when statistical assumptions are broken, the model cannot be accurately applied to the whole population (i.e. the parameters of the model are said to be biased). In other words it is not possible to draw conclusions about the population, although valid estimates of a multilevel model for the sample were generated. Therefore the results of the model generated in this section should be interpreted cautiously.

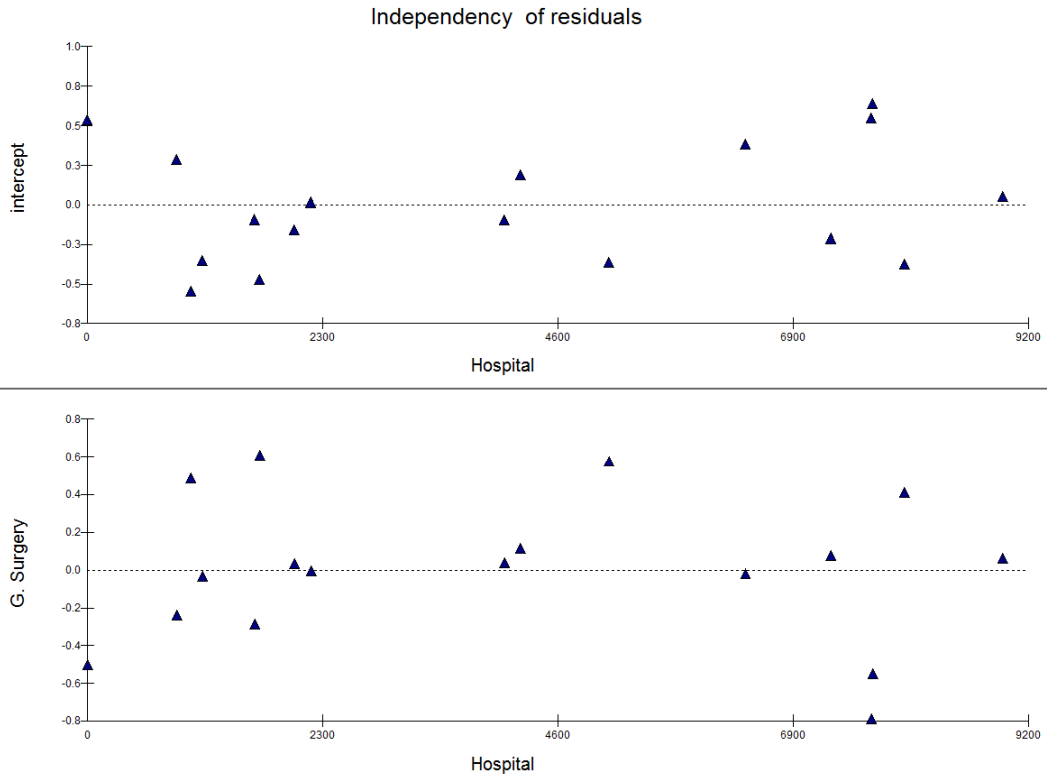

Figure 7.1: Checking the assumption of the independence of the residuals

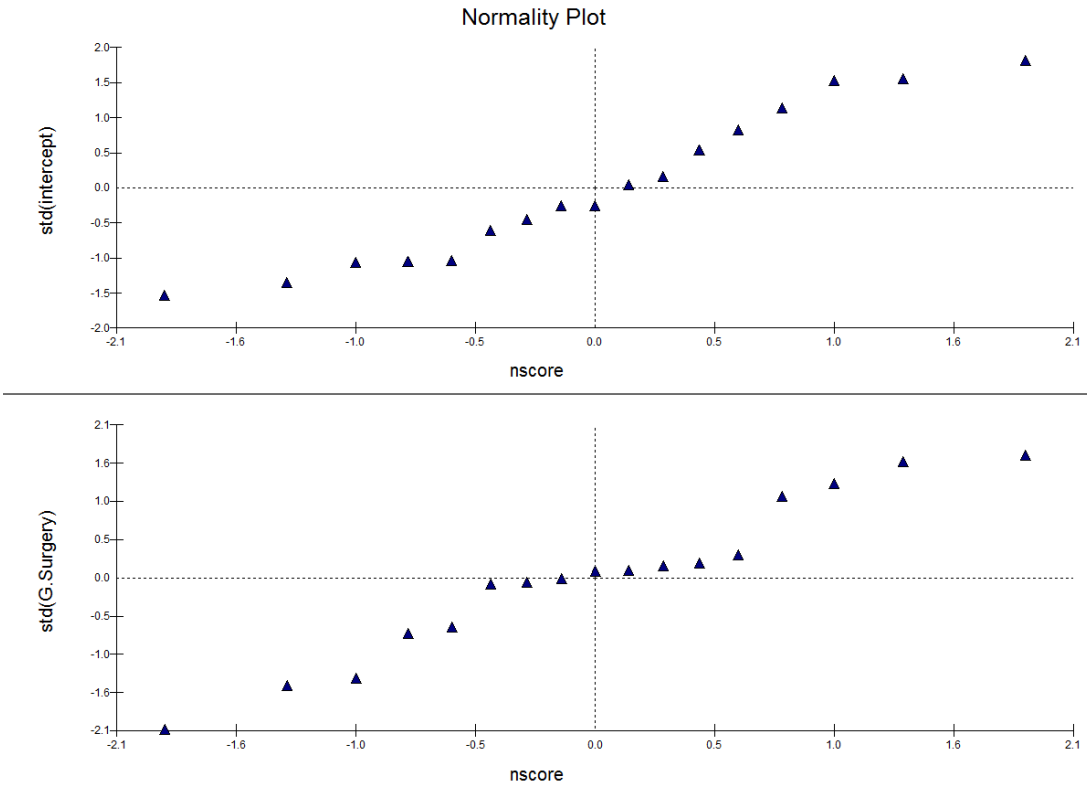

Figure 7.2: Checking normality assumption on residuals

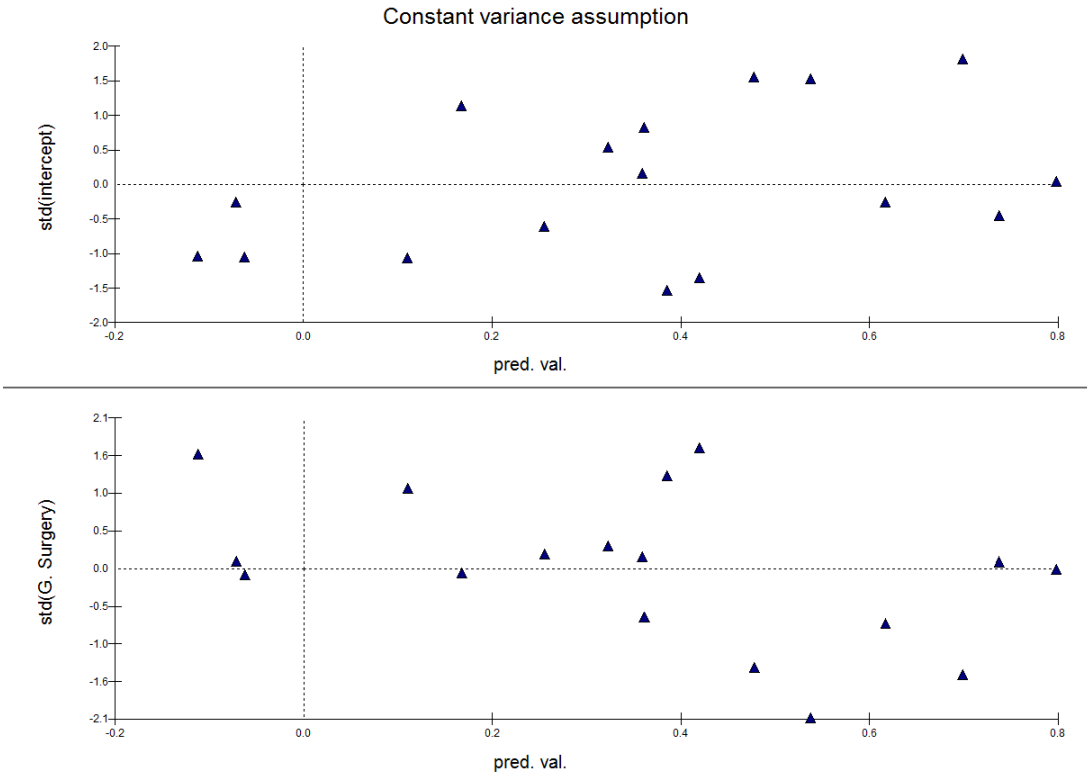

Figure 7.3: Checking constant variance assumption

In terms of interpretation of the model, the fixed coefficients in the model have the usual interpretation (see Section 6.1). For example, a patient whose diagnosis belongs to category 2 (e.g. diabetes mellitus, stroke, hepatic failure, gastrointestinal haemorrhage) is 2.05 $(e^{0.721})$ times more likely to have a medium-long LoS than a short LoS[46].

However, the interpretation of the intercept and slope parameters is a little bit more complex since they have both a fixed component and a random component. The fixed component of the intercept tells that on average a patient is 1.74 $(1/e^{-0.556})$ times more likely to have a short LoS than a medium-long[46]. In other words, for a patient on the baseline categories admitted at an "average" hospital ($u_{0j} = 0$), the probability to have a short LoS is equal to:

$$f(y_{ij} = 0) = \left(1 - \frac{1}{1 + e^{-\beta_0}}\right) = 0.63$$

The fixed component of the slope can be interpreted similarly, however an alternative interpretation tells that on average the odds of having a medium-long LoS decreases 41.4% $(e^{\beta_6} - 1)$ for a patient admitted to the general surgery ward[46].

On the other hand, the random part of the intercept tells that since the hospital effect u_{0j} follows a normal distribution, it would be expected that 95% of the hospitals have a value of u_{0j} between $\pm 1.96\sigma_{u0}$. Therefore, for 95% of the hospitals, the probability to have a short LoS for a patient on the baseline categories lies between:

$$1 - \left[1 + e^{-(\beta_0 + 1.96\sigma_{u0})}\right]^{-1} = 0.32 \text{ and } 1 - \left[1 + e^{-(\beta_0 - 1.96\sigma_{u0})}\right]^{-1} = 0.67$$

This type of interval is sometimes called a coverage interval (Steel, 2009).

Similarly, the random part of the slope suggests that for 95% of the hospitals, the probability to have a medium-long LoS for a patient treated in the general surgery ward and the rest of the characteristics on the baseline categories, lies between $\left[1 + e^{-(\beta_0 + \beta_6 + 1.96(\sigma_{u0} + \sigma_{u6}))}\right]^{-1} = 0.06$ and $\left[1 + e^{-(\beta_0 + \beta_6 - 1.96(\sigma_{u0} + \sigma_{u6})}\right]^{-1} = 0.60$.

Notice that coverage intervals are based on the normality (of the residuals) assumption. However, this assumption has been already questioned in the previous analysis. Moreover the caterpillar graphs depicted on Figure 7.4 show that around 10 hospitals at the lower and upper end of the graph have confidence intervals which do not include the mean zero, indicating that these hospitals differ significantly from the average at the 5% level. Therefore the interpretation of the model-based on the assumption of normality of the residuals (e.g. coverage intervals) should be taken cautiously.

[46]This interpretation is true only if the effects of the other variables and random effects are held constant (Urbach D.R. and Austin P.C., 2005)

For patients in the baseline categories the residual variance is $var(u_{0j}) = \sigma_{0j}^2 = 0.141$, and for patients admitted to the general surgery ward, the residual variance is calculated using Equation 7.13: $var(u_{0j} + u_{6j}Ward\_1) = 0.121$. This indicates that there is a slightly greater variation in the probability of having a medium-long for patients admitted at the adult medicine ward than those admitted at general surgery ward.

Figure 7.5 depicts a negative linear relationship between the intercept and slope residuals which it is supported by the negative covariance value (-0.125), indicating that the higher the intercept the less steep the slope. In other words, the hospitals with above-average probabilities of having a medium-long LoS tend also to have below-average effect of general surgery ward. This negative linear relationship is very strong as its correlation coefficient of 0.84 demonstrates.

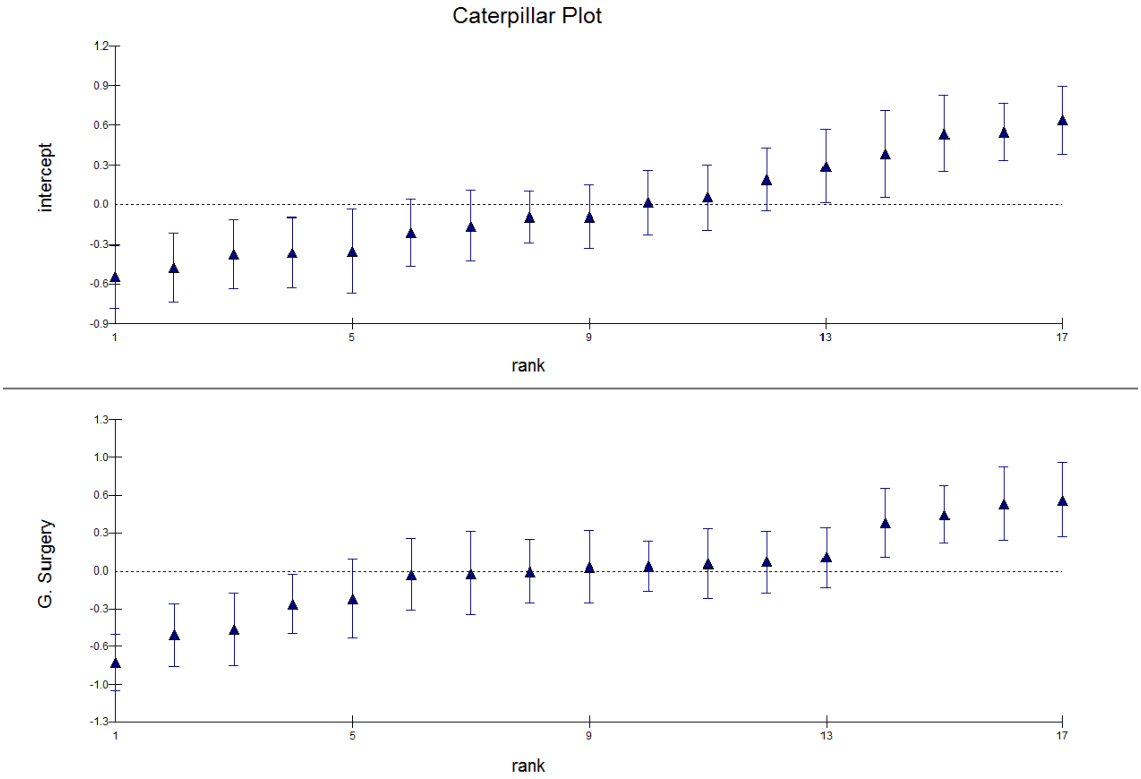

Figure 7.4: Caterpillar plot

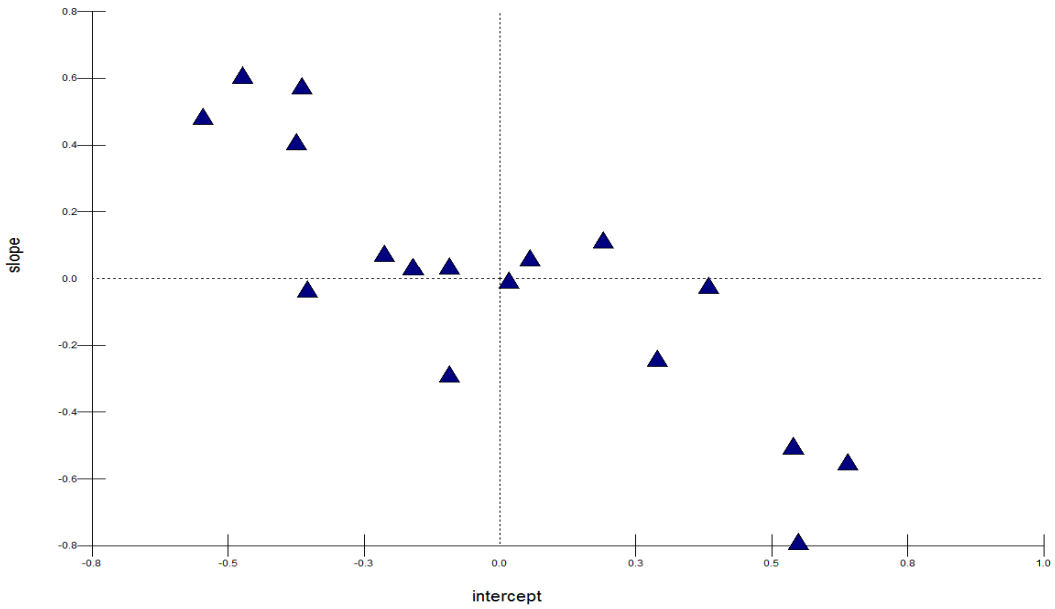

Figure 7.5: Relation between slope and intercept

On the other hand, the estimate of the contextual variable indicates that the odds of having a medium-long LoS are 1.31 higher for each extra operating room in the hospital *j* where the patient is treated.

Previously, it was mentioned that the residual variation at level 2 for a patient in the baseline categories is 0.14; however, it might be of interested to the user to know how much of the total variance is attributable to the hospital only. In this context, when the outcome variable is continuous this can be obtained by dividing the variance at level two by the total variance (i.e. variance at level two plus the variance at level 1) and is called the variance partition coefficient (VPC) or intra class correlation. However, when the outcome variable is binary there is not an estimate of the variance at level 1[47], and the level 2 variance is measured on a logistic scale so it is not directly comparable to level 1 variance (Goldstein et al., 2002). In their paper, Goldstein et al. (2002) discusses four methods to calculate a VPN for the multilevel logistic model: one of them called the simulation method is implemented on MLwiN. It consists of drawing *M* values of u_j from $N(0, \sigma_u^2)$; next, for given values of $\boldsymbol{x_{ij}}$, *m* corresponding values of π_{ij} are computed using equations (7.3) to (7.5). Finally, the level 1 variance is equal to $\frac{1}{M}\sum_{m=1}^{M}\pi_{ij}^{m}\left(1-\pi_{ij}^{m}\right)$ and the level 2 variance is equal to $var(\pi_{ij})$. The results of 5000 simulations ($M = 5000$) of the model described on this chapter, with the baseline scenario (i.e. $\boldsymbol{x_{ij}} = 0$), yield a level 1 variance of 0.2107 and a level 2 variance of 0.0065, generating a VPC equal to 0.03. This

[47] In logistic regression models, level 1 variance is a function of the mean, which depends on the values of the explanatory variables in the model $var(y_{ij}) = \pi_{ij}(1-\pi_{ij})$

means that among patients in the baseline categories 3% of the residual variation is attributable to difference between hospitals. This is small effect since it is common to find VPC ranging from 0.05 to 0.20 according to (Snijders and Bosker, 1999).

Finally, the model could be evaluated according to how well they classify patients into the correct LoS category. The results of ten-fold cross-validation show an overall accuracy rate of 71.12%, and accuracy rates per category of 80.7% and 59.0% for short-medium and long LoS respectively. Recalling that the accuracy rates express the percentage of times the patient observed LoS category matches with the predicted LoS category.

7.3. Summary

The aim of this chapter was to provide an extension of the models built in previous sections, in order to understand the environment in which the patient is treated and how this affects LoS. As a result a model is provided, that adapts itself from a local level (hospital) to a regional or institutional level and vice versa.

In a first direction to address the objective, the logistic regression model built in previous chapters for the MRC and ISSEMyM hospitals was extended to account for the hospital where the patient is treated. The results confirmed the statistical significance of the extra variable, highlighting the importance of counting for such external factors in any model to predict LoS.

However this extended logistic regression model is inadequate to make any inference about the population, as explained in Section 7.2. In consequence, a multilevel logistic regression model was suggested to address the new challenges. Using data from 19 hospitals the model was developed step by step, following a specific methodology, described in Section 7.2.1.

The results highlighted a fact that was previously hidden: the fact that there are actually significant differences across hospitals in the way that the probabilities of LoS category=1 (i.e. long LoS) are assigned by the logistic model and that the effect of the ward (i.e. general surgery ward) where the patient is treated differs across hospitals. Further, the interpretation of the estimates highlighted the utility of the model for a decision-making process. In addition, to the information provided by the standard logistic model (i.e. understanding the effects of patient characteristics on their LoS), multilevel models allow the calculation of probabilities for a patient on a baseline scenario in an "average" hospital, and in this context, it is possible to identify which hospitals differ considerable from that average.

Although valid estimates of a multilevel model for a sample of 19 hospitals were generated, the analysis of the residuals suggested that the model does not meet all the assumptions on which it is based and therefore it is advisable to take the results of this analysis cautiously. Nevertheless,

when the data is sufficient and the assumptions are met, the use of the multilevel model is highly recommended in healthcare modelling for the various reasons mentioned in previous sections. Being perhaps the simplest argument that because so much of what it is studied in healthcare sciences is multilevel in nature, researchers then should use theories and analytic techniques that are also multilevel. If they do not, there is a chance to run into serious problems of inference, like inappropriately assuming that relationships discovered at one particular level occur in the same fashion at some other (higher or lower) level (Luke, 2004).

8 APPLICATIONS TO THE DECISION-MAKING PROCESS

The aim of this chapter is to explore the application of the models developed in previous sections to the decision-making process in healthcare. Section 8.1, first discusses the individual and group-based approaches that have been extended to contribute for a better understanding of patient flow in a bed management context. Next, it describes current methodology to calculate bed requirements and recommends the use of two other methods based on the finite mixture models defined in Chapter 4 as an alternative approach. Finally Section 8.2 is devoted to discuss the applications of the multilevel model discussed in Chapter 7 and how it can be extended to provide predictions for institution, hospital and patient levels.

8.1. Bed Management

Length of stay is certainly a direct determinant on bed management practices in practically any hospital around the world. Bed management consists on the tactical/operational day to day allocation of beds and the strategic planning task of ensuring beds are available for admission whilst not restricting elective work by keeping beds idle (Boaden et al., 1999). Moreover, good bed management is a bi-fold task consisting of having the appropriate number of beds that a hospital/ward requires and the efficient utilisation of those beds.

8.1.1. Understanding Patient Flow

One of the most common challenges at any healthcare facility is the efficient utilisation of beds. Because expanding their infrastructure is not possible or limited, hospitals usually are forced to work with their resources.

The use of beds in an efficient way is a process of matching the minute by minute changes of supply and demand (i.e. the new need for beds and current bed status). Bed managers, who are usually the chief nurses, need to have constantly updated information about patients unlikely (and likely) to leave in the next 24hr or in the upcoming days, number of vacant beds and information on emergency/programmed admissions (Boaden et al., 1999).

The first two points are related to the patient flow. In this context, the models developed through this thesis can provide valuable information for a better understanding of those flows. For example, Table 8.1 lists the characteristics of five patients admitted to ISSEMyM hospital. Since the individual-based approach is used for this hospital, every patient has an associated density curve[48] (Figure 8.1) and an expected value of LoS[49] (Figure 8.2) that were estimated based on the patient characteristics. For example, according to the finite mixture regression patients who have some of the following characteristics are more likely to have a long LoS: those who are older, have a diagnosis from category 2 (i.e. diabetes mellitus, stroke, hepatic failure, cirrhosis, gastrointestinal haemorrhage, etc.) or underwent a surgical procedure category 2 (i.e. appendectomy, bowel endoscopy, laparoscopic cholecystectomy, etc.).

Notice that the expected LoS is similar to the observed value for most of the patients, although the models developed in this thesis do not aim to give crude estimations of LoS (but to ascertain the probability with which specific values of the variable LoS will occur). Moreover the model for ISSEMyM successfully predicted the observed LoS category.

[48] Using the density function $f(y_i;\varphi) = \sum_{s=1}^{S} \pi_s f_s(y_i;\theta_s)$

[49] The mean or expected LoS was calculated as a linear combination of the means of each component: $E(y_i|\boldsymbol{x_i}) = \sum_{s=1}^{S} \pi_s E_s(y_i|\boldsymbol{x_i})$

	Patient A	Patient B	Patient C	Patient D	Patient E
Age (years)	36	26	26	44	69
Category first diagnosis	1	0	3	2	0
Previous admissions	2	0	0	2	0
Origin	A&E	A&E	Transfer	Transfer	A&E
Ward	G. Surgery	G. Surgery	Trauma	G. Surgery	A. Medicine
Category main diagnosis	2	1	1	1	2
Number of diagnoses	1	1	1	1	1
Category surgical Procedure	2	0	2	2	0
Transfusions in the past	None	None	None	None	None
Total number of comorbidities	0	0	0	0	0
Addictions	Drinking & smoking	None	Drinking or smoking	None	Drinking or smoking
Observed LoS	4	1	4	1	22
Observed LoS category	Short-medium	Short-medium	Short-medium	Short-medium	Long
Predicted LoS category	Short-medium	Short-medium	Short-medium	Short-medium	Long
Expected LoS[50]	3.4	1.8	1.2	1.3	15.8

Table 8.1: Characteristics of five selected patients at ISSEMyM hospital

[50] In the current version of the STATA program used to estimate Gamma finite mixture models for the ISSEMyM hospital, the individual mean per component $E_s(y_i|\boldsymbol{x_i})$ is equal to $\alpha_s \exp(\vartheta_{si})$, where $\vartheta_{si} = (\beta_{s0} + \beta_{s1}x_{i1} + \cdots + \beta_{sp,}x_{ip})$ is the vector containing the random variables corresponding to the measurements of the patient features under study.

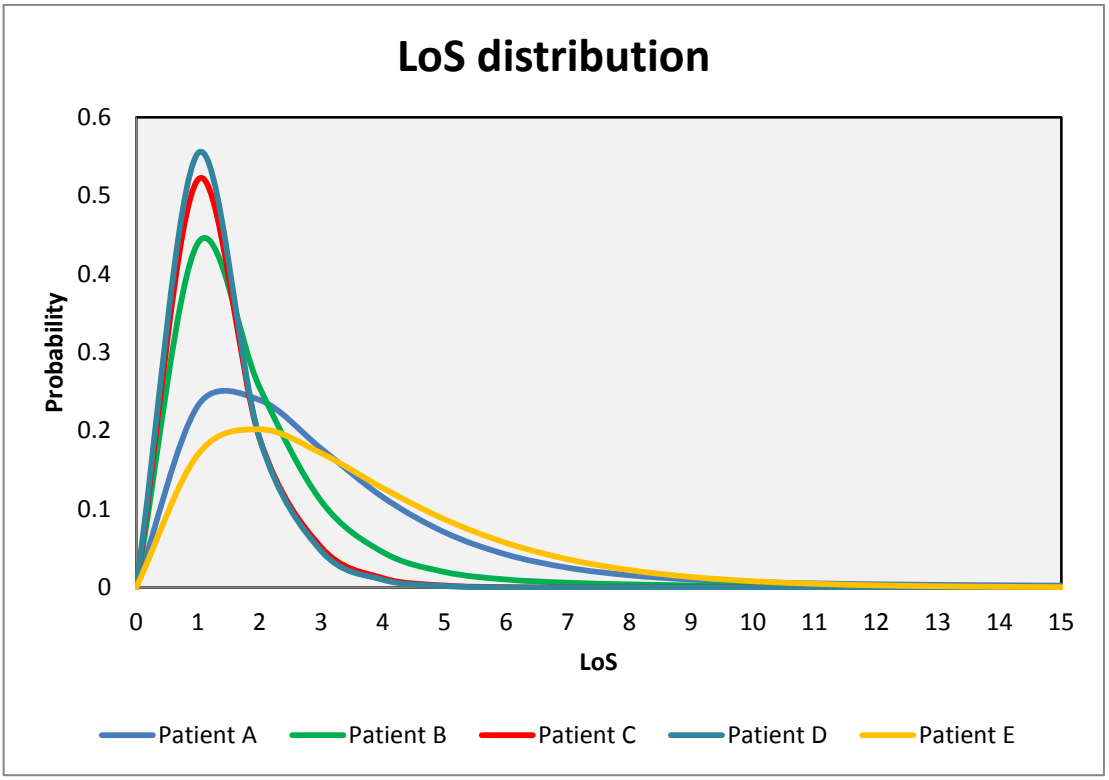

Figure 8.1: LoS density curves for five selected patients admitted at ISSEMyM

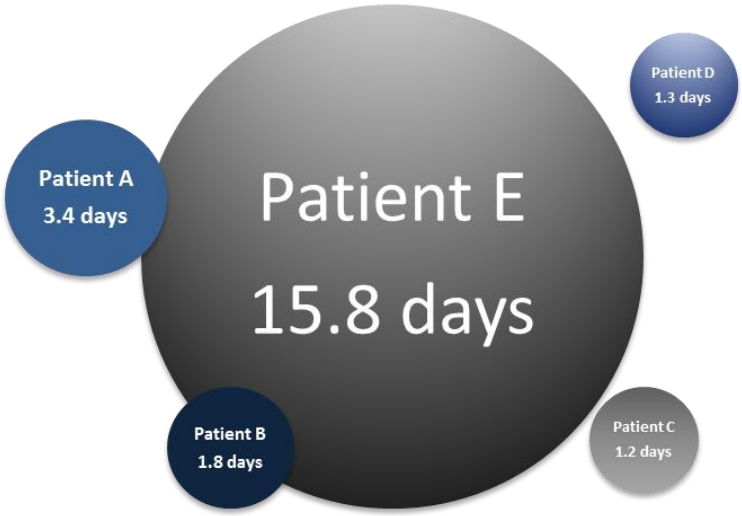

Figure 8.2: Expected length of stay for five selected patients admitted at ISSEMyM

Figure 8.1 exposes how patient characteristics can shape the LoS distribution: Patient C and D exhibit a similar behaviour, where their probability of having a very short LoS (< 2days) is a lot higher than for the rest of the patients. Patient B has a high likelihood of having a very short LoS (as the prominent peak at the very beginning of the graph points out) but its slope decreases at a slower rate than patients C and D, indicating a higher likelihood of having a medium LoS compared to C and D. Finally for patients A and E, their probabilistic curves indicate that the

likelihood of having a medium or long LoS for these patients is much higher than for the rest of the selected patients.

Moreover, in Table 8.2, the probability of a patient having a LoS equal to *y* days or less is displayed, which was calculated using the cumulative density function of the finite mixture model with *S* components (see Equation 8.1),

$$F(y_i|\boldsymbol{x_i}) = \sum_{s=1}^{S} \pi_s \int_0^{y_i} f_s(y_i|\boldsymbol{x_i}) dy_i \tag{8.1}$$

where $\boldsymbol{x_i}$ is the vector containing the attributes of patient *i*.

However, rather than referring to LoS probability density function or cumulative density function, it is really more practical to understand the LoS density in terms of survival analysis. Survival analysis is concerned with the distribution of the time to the occurrence of some event (i.e. patient discharge) or events (Kleinbaum, 1998).Survival and hazard functions are the heart of survival analysis, the former provides the probability of surviving beyond time *y*, whereas the latter gives the limiting probability that the event occurs in a given interval of time provided that the subject has survived after time *y*(Cleves et al., 2008).

McLachlan and McGiffin (1994) reviewed the role of finite mixture models in the field of survival analysis, where the survival function $R(y_i|\boldsymbol{x_i})$ of the patient *i*, corresponding to the finite mixture model with *S* components, has the form of Equation 8.2:

$$R(y_i|\boldsymbol{x_i}) = \sum_{s=1}^{S} \pi_s R_s(y_i|\boldsymbol{x_i}), \tag{8.2}$$

where $R_s(y_i|\boldsymbol{x_i}) = \int_{yi}^{\infty} f_s(y_i|\boldsymbol{x_i}) dy_i$, and the corresponding hazard function $h(y_i|\boldsymbol{x_i})$ is (see Equation 8.3)

$$h(y_i|\boldsymbol{x_i}) = \sum_{s=1}^{S} \pi_s h_s(y_i|\boldsymbol{x_i}) \frac{R_s(y_i|\boldsymbol{x_i})}{R(y_i|\boldsymbol{x_i})} \tag{8.3}$$

since $h_s(y_i|\boldsymbol{x_i}) = \frac{f_s(y_i|\boldsymbol{x_i})}{R_s(y_i|\boldsymbol{x_i})}$, the hazard function can be written as Equation 8.4:

$$h(y_i|\boldsymbol{x_i}) = \frac{f(y_i|\boldsymbol{x_i})}{R(y_i|\boldsymbol{x_i})} \tag{8.4}$$

Table 8.3 and Table 8.4 show the survival and hazard functions of five patients admitted to ISSEMyM applied to the context of this research.

LoS	Patient A	Patient B	Patient C	Patient D	Patient E
1 day	0.14	0.33	0.52	0.51	0.09
5 days	0.84	0.97	0.99	0.99	0.73
10 days	0.97	0.99	1	1	0.91

Table 8.2: The probabilities that selected patients will be discharged by day *y* or before.

LoS	Patient A	Patient B	Patient C	Patient D	Patient E
1 day	0.86	0.67	0.48	0.49	0.91
5 days	0.16	0.03	0.01	0.01	0.27
10 days	0.03	.01	0.0	0.0	.09

Table 8.3: The probabilities that selected patients will still be retained in hospital after *y* days.

LoS	Patient A	Patient B	Patient C	Patient D	Patient E
1 day	0.27	0.67	1.07	1.1	0.18
5 days	0.42	0.51	1.5	1.6	0.33
10 days	0.19	.36	1.6	1.7	0.09

Table 8.4: The probabilities that selected patients who have been in the hospital for y days will be discharged in the next 24 hours. (Notice some values are higher than 1; more on this follows).

Conversely, Table 8.5 lists the characteristics of five patients admitted to the MRC hospital. Because the group-based approach is used for this hospital, every patient is assigned to a LoS category based on their characteristics. For example, according to the logistic regression model, patients who have some of the following characteristics are more likely to have a short LoS: those who enter to the hospital via A&E, are treated in the adult medicine ward have a diagnosis from category 2 (i.e. diabetes mellitus, stroke, hepatic failure, cirrhosis, gastrointestinal haemorrhage, etc.) or underwent a surgical procedure category 3 (i.e. exploratory laparoscopy, prostatectomy, cholecystectomy, etc.). All the sampled patients belong to the short LoS category, except for patient D who clearly belongs to the medium-long category. Finally, an expected value of LoS[51] (Table 8.5), density curve[52] (Figure 8.1) and survival curve[53] (Figure

[51] $E(y_i|\boldsymbol{x_i}) = E_s(y_i)$

8.4) are appointed according to the category they belong (based on the two-component Lognormal mixture model). As it was mentioned before, the models developed in this thesis do not aim to provide crude estimations of LoS. However, notice that the expected LoS (and expected LoS category) are very similar to the observed values.

	Patient A	**Patient B**	**Patient C**	**Patient D**	**Patient E**
Age (years)	20	63	39	30	37
Previous admissions	5	48	1	1	7
Origin	A&E	A&E	A&E	A&E	A&E
Ward	A. Medicine	A. Medicine	G. Surgery	G. Surgery	A. Medicine
Category main diagnosis	3	3	3	1	3
Category surgical procedure	1	1	1	3	1
Observed LoS	2	2	1	4	1
Observed LoS category	Short	Short	Short	Medium-long	Short
Predicted LoS_Category	Short	Short	Short	Medium-long	Short
Expected LoS	1.9	1.9	1.9	6.5	1.9

Table 8.5: Characteristics of five selected patients at MRC hospital

The reason why the survival curve is displayed here is because it is important to highlight that, although the survival function is derived from the density curve itself, it does have a very different application as Harrison (1994) stated. The LoS density curves depicted in Figure 8.3 can be understood alternatively as the fraction of short LoS (or medium-long LoS) patients who leave the hospital on day *y* after admission, whereas the survival curves can be alternatively understood as the fraction of short LoS (or medium-long LoS) patients who will be in the hospital for at least *s* days (and therefore will occupy beds). The LoS density function describes the duration of treatment, whereas the LoS survival function describes the bed occupancy distribution. Therefore, the survival curve should exhibit theoretically the same behaviour that a curve generated from a census of all the patients in the hospital at a particular time and record how long each patient has been in the hospital at that time.

[52] $f(y_i|\boldsymbol{x_i}) = f_s(y_i)$
[53] $R(y_i|\boldsymbol{x_i}) = R_s(y_i)$

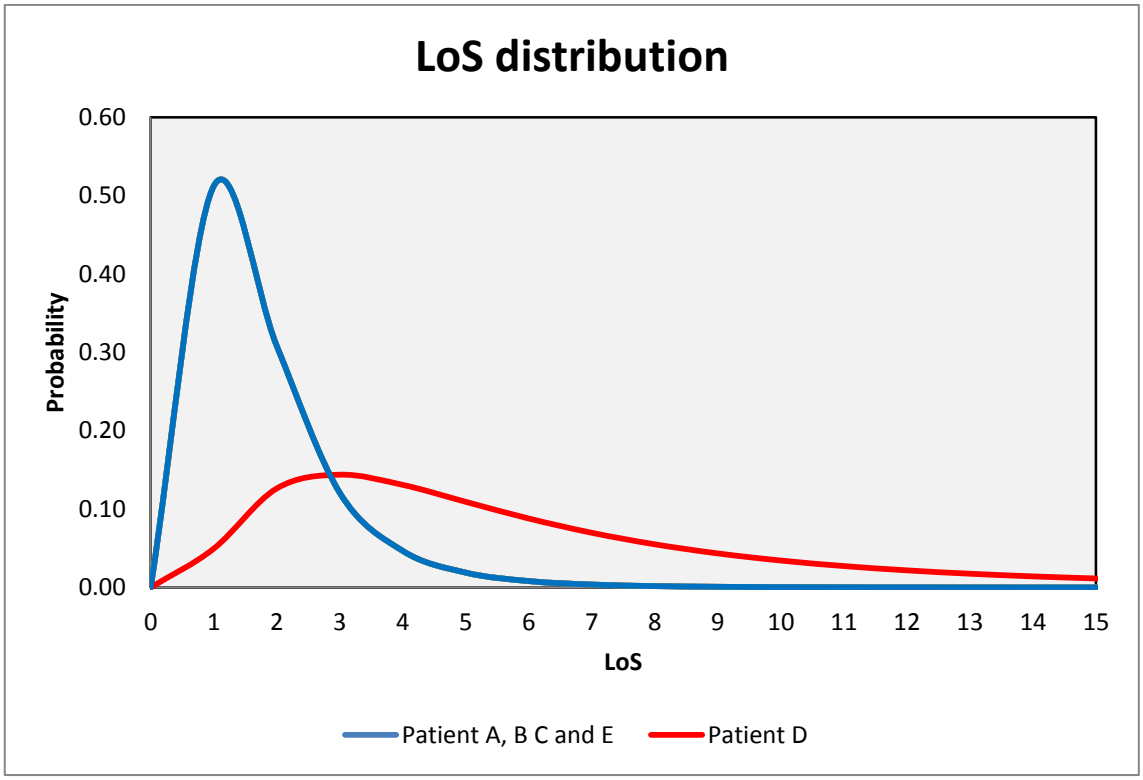

Figure 8.3: LoS density curves for five selected patients admitted at MRC

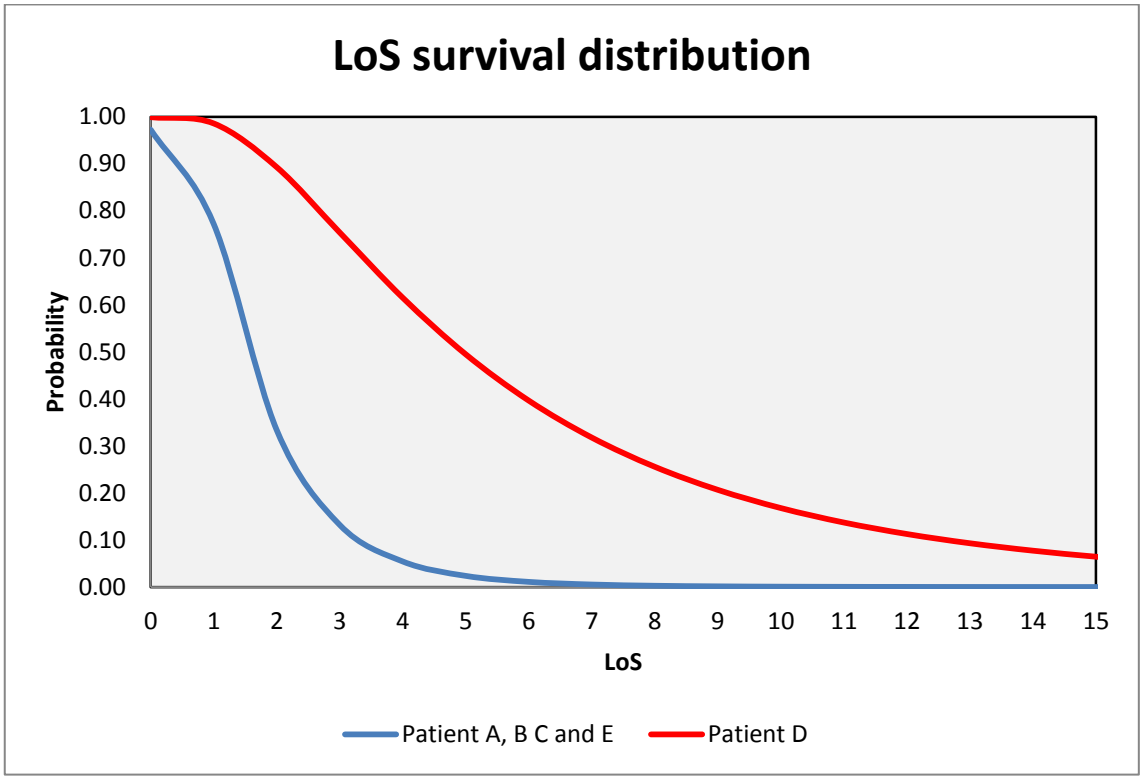

Figure 8.4: LoS survival curves for five selected patients admitted at MRC

Table 8.6 lists the cumulative density function ($F_s(y_i)$).Table 8.7 describes in more detail the survival curve depicted in Figure 8.4.Table 8.8shows the hazard function $h_s(y_i)$.

	Short LoS	**Medium-long LoS**
LoS	**Patient A, B, C and E**	**Patient D**
1 day	0.23	0.02
5 days	0.97	0.50
10 days	0.999	0.83

Table 8.6: The probabilities that selected patients will be discharged by day *y* or before.

	Short LoS	**Medium-long LoS**
LoS	**Patient A, B, C and E**	**Patient D**
1 day	0.77	0.98
5 days	0.03	0.50
10 days	0.001	0.17

Table 8.7: The probabilities that selected patients will still be retained in hospital after *y* days.

	Short LoS	**Medium-long LoS**
LoS	**Patient A, B, C and E**	**Patient D**
1 day	0.66	0.05
5 days	0.80	0.22
10 days	0.58	0.20

Table 8.8: The probabilities that selected patients who have been in the hospital for y days will be discharged in the next 24 hours.

Notice that, as the patient LoS increases, the cumulative probability function in Table 8.2 and Table 8.6 increases towards one, and the survivor function in Table 8.3 and Table 8.7 decreases towards zero. Even so, the hazard functions in Table 8.4 and Table 8.8 present a different behaviour. The hazard function is the probability that the discharge of a patient will occur in a given interval of time; therefore, it is not coherent that some patients on Table 8.4 exhibit a value greater than one. However, the hazard function can actually vary from 0 to infinity because it is a rate measured in $^1/_t$ units, or more specifically it is an instantaneous rate of

failure. For example, patient C has a hazard function (or rate of failure) of $1.8/day$ when LoS= 5 days. This means that if this rate to continue for an entire day, it would be expected that the patient C is discharged 1.8 times. In practical terms, it is evident that a patient cannot be discharged more than one time, and therefore, a hazard function greater than one means that, if the rate continue for an entire day, it would be expected that the patient is definitely discharged that same day.

Another characteristic of the hazard function is the one observed in ISSEMyM patients A, B and E on Table 8.4 and MRC patients A, B and C on Table 8.7; over time, the hazard rates can increase, decrease, remain constant or even take on a more serpentine shape[54]. There is a one-to-one relationship between the probability of staying in hospital after a certain time (i.e. survival function) and the amount of risk of been discharged that has been accumulated up to that time. The hazard function measures the rate at which the risk of discharge is accumulated (Cleves et al., 2008). Therefore since the probability for ISSEMyM patients A and B, and MRC patients A, B and C of staying at hospital after 10 days is close to zero, the associated accumulated risk of been discharged is very small likewise. Similarly the survival function for the ISSEMyM patient E decreases towards zero at a very small rate (it seems to be almost constant for a while), so it is expected that the risk of being discharged is equally accumulated at a very small rate.

Hazard rates can be of especial interest for the bed manager at the hospital because it gives an idea of which patients are unlikely to leave in next 24 hours (i.e. hazard function less than 0.5) and which patients are likely to be discharged on the next visit by the consultant (i.e. hazard function greater than 0.5). Moreover summing the hazard rates for an entire hospital (or ward) would give the expected number of beds that would be available in the next 24 hours. (Harrison and Escobar, 2010)

Finally, the following conditional probabilities could be calculated for both approaches and can bring valuable information for the bed manager:

The (conditional) probability that a patient will still remain at hospital *m* days later, given that he or she has already been at hospital for *n* days (see Equation 8.5). For example the probability patient A at will still remain at ISSEMyM three days later, given that he or she has already been at hospital for five days is 0.35.

$$r(m|n) = \frac{R(y_i = m + n)}{R(y_i = n)} \tag{8.5}$$

[54] Cleves, M. A., Gould, W. and Gutierrez, R. (2008) provides a nice example: the human mortality generates a falling hazard for a while after birth, and then a long flat plateau, and thereafter constantly rising until eventually it reaches infinity at about 100years.

The (conditional) probability that a patient will be discharged *m* days later, given that they have been at hospital for *n* days. For example: the probability that patient A from ISSEMyM will be discharged exactly three days later, given that he or she has been at hospital for five days is 0.11

$$d(m|n) = \frac{F(y_i = m + n) - F(y_i = m + n - 1)}{R(y_i = n)} \tag{8.6}$$

8.1.2. Bed Requirements

In the case of planning for a new hospital or expanding/reducing the capacity of an existing hospital, determining the optimal number of beds becomes a high priority target. Determining the appropriate number of inpatient beds is a strategic decision over the long term range that needs to be regularly reviewed. Nowadays, developed countries can take advantage of the so-called bed management systems, which, using an element of forecasting the supply and demand of beds over a given time frame, have been designed to improve the quality of decision-making in bed management on both a short and long-term basis. Two examples of these systems used by British hospitals are BOMPS (McClean and Millard, 1995) and PROMPT (Harper, 2002).

However the reality in Mexico, as happens frequently in other developing countries, is different whereby the ISSEMyM calculates the appropriate number of beds based on the very common method of ratios (see Equation 8.7):

$$N = \frac{Average\ LoS \times number\ of\ patients}{number\ of\ days} \tag{8.7}$$

This method is based on ALoS (See Equation 2.1) whose drawback and inappropriateness when high variability is present was already described in Chapters 1 and 2. In this context, one of the most important flaws of ALoS is that it leads to an overestimation of LoS when a hospital has clearly different types of workloads (i.e. a high proportion of short-stay patients vs. a small proportion of long-stay patients). In addition, the traditional ratio method tends to overestimate the number of beds required by departments (or hospitals) when LoS are usually higher (Nguyen et al., 2005), as it is the case of ISSEMyM.

On the other hand, the MRC hospital does not has any method to calculate the appropriate number of beds, although there is increasing interest for adopting planning practices.

The implementation of any bed management system like the ones mentioned before would imply lengthy investment of time, human and financial resources for both hospitals. In the meantime, a small modification of the Equation (8.8) is suggested:

$$N_s = \frac{\pi_s E_s(LoS) \times numberof\ patients}{number\ of\ days} \tag{8.8}$$

According to the group-based approach, $E_s(LoS)$ is the mean of component s, π_s is its relative size and N_s is the number of beds required for component s-type patients. Equation (8.8) is suggested to make a distinction between the different types of beds that are needed according to type of stay. Since beds at public hospitals in Mexico usually have the dual purpose of short stay and long stay care[55], this valuable information about specific requirements of beds by type of workload could possibly prevent bed blocking and reduce pressure on public hospitals by making them more efficient in their bed management

	ISSEMyM	**MRC**
Average monthly admission	784	277
Observed ALoS	5.73	3.97
	Gamma mixture	**Lognormal mixture**
$\boldsymbol{E_1(LoS)}$	3.76	1.87
$\boldsymbol{E_2(LoS)}$	14.61	6.54
$\boldsymbol{\pi_1}$	0.82	0.55
$\boldsymbol{\pi_2}$	0.18	0.45

Table 8.9: Average admission per month, oberved ALoS and finite mixture extimates for both hospitals (based on the results of Section 4.1.3). $E_s(LoS)$ is equal to $e^{(\mu_s+\sigma_s^2/2)}$ for the Lognormal mixture and $\alpha_s\beta_s$ for the Gamma mixture.

Using data from Table 8.9, it can be calculated that the appropriate numbers of beds for ISSEMyM and MRC for a month (30 days) using Equation (8.7): ISSEMyM hospital requires 81 short-medium stay beds $(3.76 \times 0.82 \times 784)/30$ and 69 long stay beds $(14.61 \times 0.18 \times 784)/30$; whereas MRC requires 10 short stay beds and 27 medium-long stay beds.

However one of the main drawbacks of both methods is that they do not take into account the fluctuation of admissions over time. Taking into account this information, a descriptive statistical analysis of admission rates per day of the week, week and month of the year was performed to explore such fluctuations. The results (not shown here) highlighted variation per

[55]Healthcare facilities in Mexico for long term or rehabilitation patients are extremely scarce, putting public hospitals (designed originally for acute care only) under too much pressure (Secretaria de Salud, 2007).

day of the week, unsurprisingly more obvious during the weekends, but consistently present as well during the working days. Furthermore an ANOVA test confirmed that these variations are significant in statistical terms.[56]

To overcome this limitation, it is suggested another method inspired by Harrison (2001)'s display of the midnight bed occupancy data[57]. Table 8.10, which is defined here as "survival bed occupancy table", is based on Harrison's arrangement of bed occupancy data, where a number of MRC patients (from the adult medicine ward) with different lengths of occupancy on seven consecutive days[58] is displayed. A diagonal, such as the number shown in bold, represents a cohort of patients admitted in the same day. For example on the second day sixteen patients were admitted. On the third day it is predicted that around fourteen patients would be still in the hospital (with occupancy time *y=1*); on the fourth day nine of them would be still there (with occupancy time *y=2*), etc. On the other hand a column represents all the patients currently in the hospital on one day. The example provided by Harrison contains real data from the midnight bed occupancy census and it represents cohorts of patients through a number of consecutive days. The method proposed here predicts the behaviour of those cohorts of patients based on their survival probabilities, using just the average admission rates per day[59] (first row in grey) and the total number of patients who are in the hospital at a particular time sorted according to how long they had been in the hospital (column day 0). This method makes distinction between two different types of patients: the first type is patients who are already in the hospital and the length of time that they have been occupying a bed is known and the second type represents patients that will be admitted to hospital during the time to consider in the analysis.

For the first type of patients, the patients that are already in the hospital, conditional probabilities are employed using Equation (9.5) since information about their current occupancy patterns is available. For example on the day of the census, there were four patients who had been in the hospital for 1 day. Their probability to still remain in the hospital one day later given that they had already been at hospital for one day is equal to $r(m = 1|n = 1) = 0.67$. This multiply by four, suggests that, by the end of the first day, around 2.70 patients of the four patients on day 0 will still remain at hospital. From those 2.7 patients, just 1.9 will remain in hospital by the end of the second day ($2.7 \times r(m = 2|n = 1)$) and so on.

The rest of table (highlighted in light grey) describes the occupancy patterns of the second type of patients (i.e. the new arrivals). That part of the table is based on the average admission rate

[56] No significant differences were found by week or month of the year. It seems that seasonal effects are not as common in Mexico as they are in United Kingdom.
[57] It was later fitted to a mixed exponential distribution
[58] For practicality Day 0 is assumed to be Sunday, Day 1 Monday and so on.
[59] Average admission rates per day were calculated using admission data from each hospital for 2005-2009.

per day of the week; and the MRC survival function $R(y_i)$ derived from the finite mixture model formulated in Section4.1.2. For example, in the second day 16 new patients are admitted; however, one day later just 13.96 of them will still remain in the hospital $(16 \times R(y = 1))$.

Next, by taking the average of the totals per column, it is possible to determine the expected number of beds that will be required for the time frame under study. Table 8.10 represents just one week of bed occupancy activity; however, the results of extending the table to account for 4 weeks (see Appendix G) suggests that an average of 37 beds are required for the adult medicine ward, with a 95% confidence interval between 36 and 39 beds. The results are compatible with the previous findings using the other two methods; however the survival bed occupancy table has some attractive advantages over the other two methods:

- Just one day of data is required to make predictions.
- Admission rates are allowed to vary per day, week or month.
- A range of useful statistics can be obtained rather than a crude average of bed requirements, e.g. maximum, minimum, standard deviation, confidence intervals, etc.
- The survival occupancy table can be split into two tables to account for the different types of patients and beds (e.g. short LoS or medium-long LoS).
- The survival occupancy table provides an insightful and graphical representation of patient flows.

Length of stay	Day number							
	0	1	2	3	4	5	6	7
y=0	**11**	11	16	6	8	9	3	8
y=1	**4**	**9.55**	9.55	13.89	5.21	6.95	7.81	2.60
y=2	6	**2.70**	**6.44**	6.44	9.37	3.51	4.68	5.27
y=3	2	4.22	**1.90**	**4.53**	4.53	6.59	2.47	3.30
y=4	1	1.49	3.14	**1.41**	**3.37**	3.37	4.91	1.84
y=5	1	0.77	1.14	2.42	**1.09**	**2.59**	2.59	3.77
y=6	1	0.78	0.60	0.90	1.89	**0.85**	**2.03**	2.03
y=7	0	0.79	0.62	0.48	0.71	1.50	**0.67**	**1.61**
y=8	2	0.00	0.63	0.49	0.38	0.57	1.19	**0.54**
y=9	0	1.61	0.00	0.51	0.40	0.31	0.46	0.96
y=10	0	0.00	1.30	0.00	0.41	0.32	0.25	0.37
y=11	1	0.00	0.00	1.06	0.00	0.34	0.26	0.20
y=12	0	0.82	0.00	0.00	0.87	0.00	0.28	0.22
y=13	0	0.00	0.68	0.00	0.00	0.72	0.00	0.23
y=14	0	0.00	0.00	0.56	0.00	0.00	0.60	0.00
y=15	0	0.00	0.00	0.00	0.47	0.00	0.00	0.50
Total	**33.73**	**42.01**	**38.69**	**36.70**	**37.00**	**31.54**	**31.71**	**33.73**
						Average		**36**

Table 8.10: Survival bed occupancy table for the MRC hospital

8.2. From a Macro to a Micro Perspective

In Chapter 7 a multilevel group-based model was developed to provide a better understanding of the environment in which the patient is treated. However it might be of interest for the user to extend the model to predict individual LoS.

The multilevel logistic regression model can be incorporated into a group-based approach, where patients are first assigned to a LoS category based on internal and external factors. Next, an expected value of LoS and a density curve are constructed according to the category to which

the patients belong, based on the two-component Lognormal mixture model discussed in Chapter 4.

Table 8.11 lists the characteristics of five from five different hospitals from the regional dataset. The results indicate that patients A and D belong to the short LoS category, whereas patients B, C and E clearly belong to the medium-long category. Moreover, an expected value of LoS (Table 8.11) and density curve (Figure 8.5) are appointed according to the category they belong (base on the two-component Lognormal mixture model).

	Patient A	**Patient B**	**Patient C**	**Patient D**	**Patient E**
Age (years)	7	85	58	32	73
Gender	Female	Male	Male	Female	Male
Previous admissions	1	1	10	1	1
Origin	O. clinic	A&E	A&E	O. clinic	A&E
Ward	G. Surgery	A. Medicine	A. Medicine	G. Surgery	G. Surgery
Category main diagnosis	3	1	3	3	1
Category surgical procedure	2	2	2	1	2
Hospital code	5095	4231	7673	4074	7673
Operating rooms at hospital of admission	2	4	4	3	4
Observed LoS	1	6	7	1	14
Observed LoS category	Short	Medium-long	Medium-long	Short	Medium-long
Predicted LoS_Category	Short	Medium-long	Medium-long	Short	Medium-long
Expected LoS[60]	1.8	6	6	1.8	6

Table 8.11: Characteristics of five selected patients

[60] The mean or expected LoS $E_s(y_i)$was calculated for each component

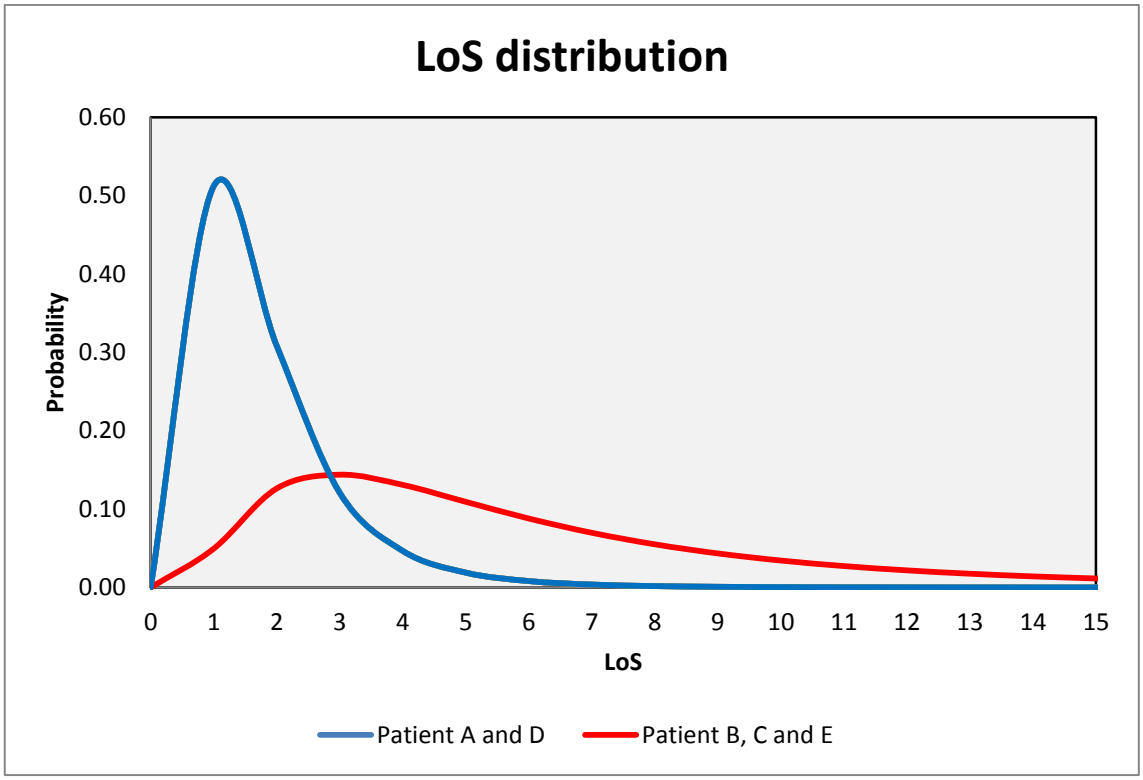

Figure 8.5: LoS density curves for five selected patients

Furthermore, the model can be extended to the equivalent of an individual-based approach to allow for "personalised" predictions, using the probability of the successful outcome π derived from the multilevel logistic regression model. Let us assume B is an arbitrary event (e.g. B is the event that a patient will have a LoS=m days), A_1 and A_2 are mutually exclusive events, where A_1 is the patient belonging to the short LoS category and A_2 is the patient belonging to the medium-long LoS category and $P(B|A_s)$ is the conditional probability of B assuming A_s. Thus, according to total probability theorem (Scheaffer and Young, 2010):

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) \tag{8.9}$$

In other words, the probability that a patient will have a LoS of exactly m days (i.e. $P(B)$) is equalto the probability of the patient belonging to short LoS category $P(A_1)$ multiplied bythe conditional probability that a patient will have a LoS=m days, given it belongs to short LoS category $P(B|A_1)$, plus the probability of the patient belonging to medium-long LoS category $P(A_2)$multiplied by the conditional probability that a patient will have a LoS=m days given it belongs to medium-long LoS category $P(B|A_2)$.

Notice that, because LoS is a continuous variable, the instance $P(B) = P(LoS = m)$ or the probability that a patient will have exactly a LoS=m days, is undefined in the continuous realm.

Therefore Equation (9.9) can be rewritten to account for LoS as a continuous variable Y with a probability density function f (Ross, 2010)

$$f_Y(y) = P(A_1)f_{Y|A}(y|A_1) + P(A_2)f_{Y|A}(y|A_2) \tag{8.10}$$

For example, if $f_Y(y)$ is the patient survival function which provides the probability of surviving beyond time y, Equation (9.10) can be rewritten as Equation (9.11),

$$R(y_i) = (1-\pi)R_1(y_i) + \pi R_2(y_i) \tag{8.11}$$

where $R_s(y_i)$ is the survival function of the component s of the finite mixture model (Section 4.1.3) and π is the (posterior) probability of the patient belonging to the medium-long LoS category extracted from the multilevel logistic regression model. The total probability theorem expressed in Equation (9.10) can be used to calculate also the density, cumulative and hazard functions.

Notice that Equation (9.11) looks very similar to Equation (9.2) which describes the density function $R(y_i)$ of the ith observation corresponding to the finite mixture model in the individual-based approach. However, they are conceptually different since π is not any longer the prior probability that an observation belongs to a certain component of the finite mixture model, which is often specified as constant for all individuals. Equation (9.11) is based on posterior probabilities π. The idea behind the group-based approach, which is the foundation of Equation (9.11), is that preceding any information, it is believed that the LoS of a patient X can be described by the density function of one of the components of a finite mixture model (i.e. either the patient belongs to component 1 (and its LoS is better described by the density function of such component) or the patient belongs to component 2). However, unlike the individual-based approach, there is no prior knowledge about which component is more likely to describe most of the patients[61]. Later, when information is available, the multilevel logistic regression gives the probability of success π, which can be understand as the posterior probability that patient X belongs to component s given patient (and hospital) information, the posterior probabilities are revised priori probabilities that take into account new available information.

Figure 8.6 shows how the posterior probabilities π derived from the multilevel logistic regression model shape the LoS density function. The dashed lines are the same curves as in

[61]This prior knowledge is represented, in the individual based approach, by the prior probabilities of the finite mixture model π_s i.e. if $\pi_1 = 0.60$ and $\pi_2 = 0.40$, it means that 60% of the patients would have their LoS described component 1 density function and 40% of them by component 2 density function.

Figure 8.5 which represents the first (short LoS) and second component (medium-long LoS) of the Lognormal finite mixture model. The continuous lines represent the density function for each patient after applying the total probability theorem. Notice the curves for patients A and D are very close to the theoretical curve for short LoS (dashed blue line), whereas the values of π had a significant impact on patient B, C and E, pushing their curves to the left of the graph, generating a more pronounced early peak compared to the theoretical curve medium-long LoS (dashed red line). However, from day 3 approximately, the patient curves behave very close to the theoretical curve. Furthermore, for every patient an associated expected value of LoS (Table 8.12) can be calculated as linear combination of the means of each component $E(y_i) = E_1(y_i)(1-\pi) + E_2(y_i)\pi$

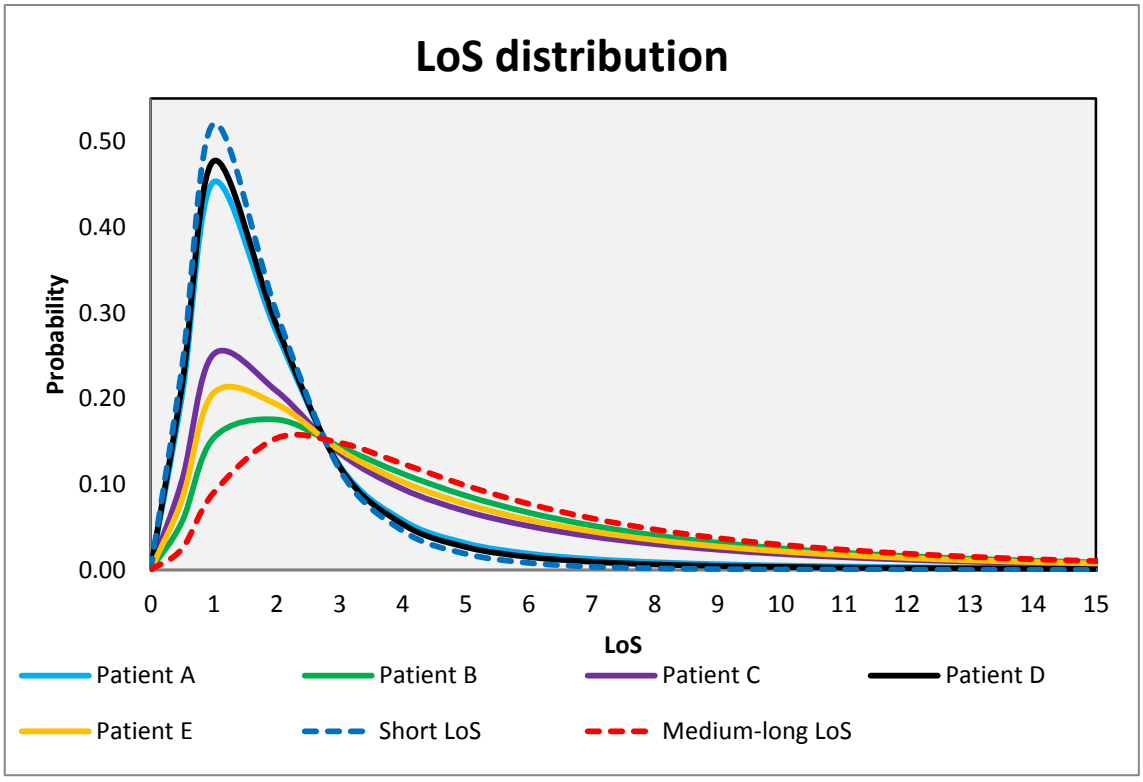

Figure 8.6: LoS density curves for five selected patients using the total probability law

	Patient A	Patient B	Patient C	Patient D	Patient E
Expected LoS	2.5	6.1	5.4	2.5	5.1

Table 8.12: Expected LoS (in days) curves for five selected patients using the total probability law

Finally, Table 8.13-Table 8.15list the cumulative density function, survival function and hazard function respectively for the five patients described in Table 8.11, using the total probability theorem where the probability functions $f_{Y|A}(y|A_s)$ are derived from the cumulative density

function, survival function or hazard function of the Lognormal finite mixture component *s*, and the probabilities $P(A_s)$ are derived from the multilevel logistic regression model.

LoS	Patient A	Patient B	Patient C	Patient D	Patient E
1 day	0.20	0.06	0.11	0.22	0.09
5 days	0.91	0.62	0.71	0.93	0.67
10 days	0.97	0.87	0.90	0.98	0.88

Table 8.13: The probabilities that selected patients will be discharged by day *y* or before.

LoS	Patient A	Patient B	Patient C	Patient D	Patient E
1 day	0.79	0.93	0.88	0.77	0.91
5 days	0.08	0.37	0.28	0.06	0.32
10 days	0.02	0.12	0.09	0.01	0.11

Table 8.14: The probabilities that selected patients will still be retained at hospital after *y* days

LoS	Patient A	Patient B	Patient C	Patient D	Patient E
1 day	0.59	0.18	0.31	0.62	0.25
5 days	0.69	0.30	0.43	0.72	0.37
10 days	0.51	0.24	0.33	0.53	0.29

Table 8.15: The probabilities that selected patients who has been in the hospital y days will be discharged in the next 24 hours.

8.3. Summary

The first sections of this chapter were devoted to exploring the applications of the individual and group approach for a better understanding of the patient flow, which is cornerstone for the day by day planning and allocation of beds. Later, the attention was focus on how ISSEMyM and MRC hospitals currently calculate their requirements of beds, and two new methods were proposed based on the estimate of expected LoS and the survival function, both derived from the finite mixture model developed earlier on this thesis.

The last section of this chapter, an application of the multilevel group-based model, which is meant to provide comparative estimations between hospitals, was firstly explored in a similar way to the group-based approach and later adapted to an equivalent of individual-based

approach to allow for "personalised" predictions. The result was a model to predict LoS based on patient characteristics and other external factors, which can be exploited by two different users: the decision maker at a large scale (i.e. government), interested in a strategic level of resource planning (i.e. national, regional or institutional level) and the decision maker (hospital manager or medical staff) at a local hospital, interested in the day to day planning of resources.

9 CONCLUSIONS

This final chapter concludes the thesis with a brief summary of main findings, the limitations of the study, suggestions for further work and the novel contributions to the field.

9.1. Summary of main findings

This thesis used model-based cluster analysis with finite mixture models to find a probabilistic model for LoS. The most common mixtures of distributions for lifetime data including Lognormal, Gamma, Gaussian and Poisson were fitted to the data. The results suggested that a two-component Lognormal mixture model was the most appropriate for describing LoS, yielding to the creation of a new variable named LoS category with two categories: Short (patients with LoS up to 2 days) and Medium/Long (patients with LoS more than 3 days). The same approach was carried out in the data by hospital: a two-component Lognormal mixture model was the most appropriate choice to describe LoS at the MRC hospital, and a two-component Gamma mixture model was the preferred option for the ISSEMyM hospital. These results yielded a redefinition of the categories in the variable LoS category: Short/Medium (patients with LoS up to 11 days) and Long (patients with LoS more than 12 days) for ISSEMyM hospital, and Short (patients with LoS up to 3 days) and Medium/Long (patients with LoS more than 4 days) for MRC hospital.

A variable selection process was carried out to select the significant variables for the LoS, using multiple stepwise regression with the backward method. The results for ISSEMyM indicated that 11 variables are significant in explaining 14% of the variance in LoS. Conversely, 7 variables in the MRC dataset were significant in explaining 21% of the LoS variance. These results were validated using non-parametric bootstrapping.

Next, the finite mixture models defined previously were extended to accommodate covariates (i.e. known now as finite mixture of generalised linear models). According to the AIC, a measure of goodness of fit, the Gamma mixture model for ISSEMyM was a better fit to the data after adding the covariates compared to the model with the intercept only. On the other hand, both AIC and BIC values agreed that the Lognormal mixture model with covariates was a better model to explain the LoS data at MRC. However, the accuracy rates when classifying patients into their correct LoS category indicated that both mixtures of generalised linear models have limited ability to predict accurately membership of the smallest component of the mixture. In consequence, for the estimation of patient LoS distribution, it is recommended to use the density function of the mixture, rather than the density function for each component.

In addition, Logit regression, decision trees (CART, QUEST, C4.5 and CHAID), Naive Bayes and hybrid methods (Naive Bayes trees and Logistic Trees) were evaluated to find the best method to predict LoS category based on patient characteristics. However, all the algorithms had a poor performance to predict long LoS category at ISSEMyM hospital. In consequence, other more sophisticated data mining techniques known as ensemble methods were analysed. However, the results did not indicate a significant improvement in the performance, which may suggest that the algorithms explored in this chapter reached the limit of information that can be predicted from the data. Nevertheless, Logit regression is suggested as the preferred option to predict LoS category at MRC, based on its good performance, simplicity and stability. On the other hand, the least bad option for the ISSEMyM hospital was the Naive Bayes. However in this case, an individual-based approach using mixtures of generalised linear models is suggested to model patient LoS in ISSEMyM.

Later, the logit regression model to predict LoS category was redesigned to account for the environment in which the patient is treated and how this affects LoS. In this context, a multilevel logistic regression model was suggested to address the new challenges. However, due to the sample size of the current dataset, the multilevel analysis was carried out on a different dataset containing patient records of 19 Mexican hospitals. A Kruskal-Wallis test supported the use of the existing LoS category parameterisation (a two-component Lognormal finite mixture model) on the new dataset. The results of the multilevel logistic model suggested that there are actually significant differences across hospitals in the way that the probabilities of LoS category are assigned by the logistic model and that, the effect of the ward, where the patient is treated, differs across hospitals. Another interesting finding is that just 3% of the total variance of the model is attributable to the hospital only. However, the analysis of the residuals suggested that the model does not meet all the assumptions on which it is based and therefore it is advisable to take the results of the multilevel model cautiously. However, valid estimates for 19 hospitals sample were generated.

This thesis closed with the applications of the individual, group and multilevel group based approaches by reviving the role of survival analysis in the field of finite mixture models. In addition, two new methods to calculate bed requirements were proposed based on the estimate of expected LoS and the survival function of the finite mixture models.

9.2. Research limitations

At this stage, a critical evaluation of the whole study is always required. Although this research satisfies the objectives defined in Chapter 1 and proposes a number of advantages from previous studies in the field, the present work has certain limitations that need to be acknowledged when evaluating the whole study and its contributions. This chapter will demonstrate how some of these limitations might be seen as areas of opportunity for future research under the same topic.

9.2.1. Data limitations

The sample of this research included two hospitals from the Mexican context that were selected following the applications of the convenient sampling technique that suits the research objectives. This was explained earlier in this thesis in (Section 1.4.1). Therefore, it cannot be claimed that the findings can be generalised over the population of the all public hospitals in Mexico. However the characteristics of the selected hospitals are a good and rich representation of the public hospital population.

In Section 4.2, the clustering of diagnoses and surgical procedures ICD codes was carried out using the particular case mix of two hospitals only. However, both hospitals do have a different case-mix, as a consequence of the level of care that each of them provides (i.e. MRC have the basic medical specialties and perform surgeries of low and medium levels of complexity, whereas ISSEMyM provides a wider range of medical subspecialties and can perform more complex surgeries in different specialties and subspecialties). Therefore, it would be natural to expect differences between other hospitals, healthcare providers or regions. Consequently the clustering arrangement suggested in Section 4.2may be inappropriate to fully capture a case-mix from other hospitals, regions or healthcare providers. One way to overcome this limitation can be researched if the implementation of DRG's becomes a reality for more hospitals (i.e. currently just SSA and IMSS have implemented them). DRG's will provide a national standard coding for diagnoses. This new classification scheme could be used to re-estimate the models developed in Chapters 5 and 6.

Another more feasible solution would be to expand the sample of ICD codes and associated LoS to include more hospitals and healthcare providers at a national scale. Then hierarchical

clustering analysis can be performed this time in a more representative sample of the population.

On the other hand, in Chapter 7, for purposes of the multilevel analysis the data was extended to account for 19 hospitals. However when an analysis of the residuals was performed, this suggested that the statistical assumptions were broken and therefore the model cannot be accurately applied to the whole population. It was advised then to interpret the results of the multilevel model cautiously.

Another limitation is that some information about the patient such as diagnosis and surgical procedure might not be available (or known) at the moment of patient admission (or when an estimation of patient LoS is needed). In the case of ISSEMyM, the impact should be lower since the models developed here are based on the first diagnosis variable which is usually entered at moment that the patient is admitted. However first diagnosis is the initial guess of the cause of admission and it might differ from the definite final diagnosis (which can be known days later). In the case of MRC general hospital, this piece of information is not mandatory at the moment of admission and moreover it is permitted to record the diagnosis variable as "cause under investigation" or "unknown" until the real patient condition is diagnosed (i.e. the patient diagnosis is a mandatory field just when the patient is about to be discharged). Thus, when data for individual predictions is missing, it is proposed to follow a group-based approach to estimate LoS using Naive Bayes to firstly identify what type of LoS the patient is more likely to have (i.e. LoS category) and then an expected value of LoS or density curve can be appointed accordingly. The reason why now Naïve Bayes is recommended instead of Logistic regression (which was the preferred model for MRC hospital) is because Naïve Bayes algorithm is a better option when dealing with missing values. If the value of a predictor is missing, such variable is simply omitted from the calculations. In other words, probability ratios would be based on the number of values that actually occur rather than on the total number of observations (Witten and Frank, 2005).

Finally, another limitation that was outside of this research's control is related to the quality of the data. Let us not forget that the success of these models and in general any other healthcare model hinges on the availability of clean healthcare data. In this context, staff at ISSEMyM was seriously concerned about the reliability of the conditions and interventions ICD coding. They have serious doubts about how standardised the coding for conditions and interventions is, since different doctors record this information in different ways at different points in time. According to Black and Payne (2003) particular emphasis should be placed on reliable recording of conditions and interventions, as it is important that any differences in patient diagnosis or treatment are due to real factors, and not because the variables were recorded in different ways

for different people. They suggest conducting a coding audit where different coders are given the same information to code and their coding is compared for concordance. Hopefully, the models developed here will show to hospitals how important good quality data is.

9.2.2. Failure to identify patients requiring long LoS

In view of the little success that the group-based models achieve on classifying ISSEMyM patients into long LoS category (Chapter 6), an individual-based approach using finite mixture regression was suggested as an alternative for the estimation.

In addition, other alternative solutions were proposed such as collecting more data based on the hypothesis that there may be other variables which are not being recorded at the moment that may have influence in the LoS category and that can help to boost the accuracy rates for the long LoS category.

Moreover, in Section 6.5 it was also suggested that the problem might be caused by the presence of ambiguous outliers. A couple of suggestions have been made on how to handle ambiguous observations: Trappenberg and Back (2000) suggested the idea of adding a new category IDK (I Do not Know) to the number of outcome categories and to classify ambiguous observations into the IDK class. Later, Hashemi and Trappenberg (2002) suggested separating typical data from atypical data (i.e. ambiguous outliers) and training the selected data mining algorithm just in the typical data. They demonstrated that with this method that although some information is lost, the accuracy rates increased significantly.

Another feasible explanation for the failure to predict long LoS is that these patients might be incurring in an inappropriate use of hospital days (more on this in Section 9.3.3).

9.2.3. Limitations to the bed management applications

The bed management methods presented in Chapter 9 were suggested for the special case of two Mexican hospitals that do not have a formal methodology (MRC hospital) or whose current procedures are inadequate (ISSEMyM) to calculate bed requirements. The aim was to provide a quick and easy-to-implement solution for the near future. However there are some pitfalls that need to be considered:

Firstly, the models developed in this thesis do not take in account that there is actually a time-gap between the patient's "formal discharge" (following a clinician authorisation) and the exact time when actually the patient leaves the bed. This is because, unlike the data from British hospitals, the data from both hospitals considers only the first type of discharge. Most patients leave a couple of hours after the doctor's authorisation; however both hospitals identified at least 20 different reasons that cause delays on discharges up to 8 hours (i.e. Some of the more

common reasons are that the patient file is not complete or it is missing, the patient is still awaiting for the catheter removal or the patient family has not arrive to pick up the patient, etc.). Omitting this time gap might be crucial for an efficient use of beds, where decisions are made day by day and minute by minute. On the other hand, knowing the time of the formal discharge could be more useful than knowing the time of actual discharge, as the latter represents the end of the patient's treatment.

In addition, the models developed in this thesis can help in identifying and monitoring patients that are likely to leave hospital in the next 24 hours and based on that information the right measures can be taken to ensure their discharge on time and under normal circumstances.

On the other hand, in Section 9.1.2 a method to calculate bed requirements was proposed, which assumes that hospital bed requirements are static over time. Albeit, bed occupancy patterns and emergency demand may be dependent on the time of the day, the day of the week and month of the year (Harper, 2002). To deal with the variation through different points of the day, it was suggested to take data from the peak times, in order to avoid underestimation of bed requirements. However this may lead as well to overestimated results.

Moreover, the model assumes that a patient occupies the exactly same bed from the day of admission until discharge; however this is often not the case: it is very common to exchange and move patients from one floor to another or from one ward to another. It is usually the case when beds are not available in the appropriate ward that patients are allocated in a temporary one. Therefore, as piece of future research, the models presented here should be incorporated in more complex simulation models or patient systems to successfully capture the complexity and uncertainty embedded in hospitals.

9.3. Extensions to this research

In addition to the areas of opportunity mentioned above, this research could be extended in a number of novel ways.

9.3.1. Dealing with LoS less than one day

Less than 5% of the admitted patients have a LoS of less than one day. They are usually patients that come to hospital for diagnosis or treatment that does not require an overnight stay. These can be emergency admissions or planned (i.e. haemodialysis patients). This research could be extended to identify this type of patients by incorporating a new component to the mixture of distributions to describe $LoS < 1\ day$. Garg et al. (2009) demonstrated that by modelling $LoS < 1\ day$ as a Gaussian component with mean 0 and standard deviation 0 (while fitting the remaining LoS with Gaussian finite mixture model) and treating this component as an

extra component of their Gaussian mixture model, it resulted in a significant improvement in the likelihoods.

A piece of future work could consist of exploring different probabilistic distributions to model LoS<1 day and re-estimating the finite mixture model for the rest of the data. Moreover, finite mixture of generalised linear models or logistic regression could be used to understand and identify the internal and external factors associated with LoS<1 day.

9.3.2. Extending the finite mixture model

In Chapter 4, four different mixtures of probabilistic distributions were fitted to the LoS data: Gaussian, Lognormal, Gamma and Poisson. The reasons why these distributions were selected are outlined in Section 4.1.1. However one can notice the absence of the Weibull distribution, which is widely recognised for its ability to capture the long right-sided tails in life-time data. Actually single Weibull distributions have proved to work well with LoS data as it was demonstrated by Marazzi et al. (1998) who carried out a study on 3279 hospital samples using single component Log-normal, Gamma and Weibull models to describe the distribution of LoS. Weibull distribution was found to be the second best fit for most of the samples. The reason why it was not included in this study is that at the time this research was being conducted, the code to fit mixtures of Weibull distributions was under development by the STATA code authors. Besides, the development of such code as part of this research would have been outside of the time frame of this research.

On the other hand, it was assumed that each component of the finite mixture model belongs to the same parametric family of distributions but with different parameters. However Atienza et al. (2008) explored a mixture of different families, through finite mixtures of Gamma, Weibull and Lognormal families to model LoS within several DRG's. Their proposed model with three components demonstrated improvements in the results obtained with mixtures of three distributions from the same family for most of the DGRs, although they indicated that the proposed model was far more complicated than any individual one.

In view of the above, fitting a mixture of Weibull distributions and fitting mixtures of different families of distributions should be in the future considered as an extension of this research.

Furthermore, in Chapter 7 of this thesis the group-based approach models, which cluster patients into LoS homogenous groups, were extended to account for external and environmental factors that might influence the patient classification (between hospitals) into those homogenous groups. This was done by using multilevel models, which assumes that hospital data is actually a hierarchical structure, where there are patients clustered in hospitals and hospitals clustered in healthcare providers and so on. Moreover, this type of analysis takes into account the variability

associated with each level of clustering, which if ignored, as happens with classical methods, may lead to wrong conclusions (Snijders and Bosker, 1999). In this context, it might be of interest to extend the individual-based approach models in the same way to account for such hierarchical structure. In other words, the gamma and lognormal mixtures of generalised linear models defined in Section 5.2can be extended by introducing random effects in the mixing coefficients and component densities, taking a form similar to Equation (9.1)

$$f(y_{ij}; x_{ij}, \varphi) = \pi_{1ij} f_{1ij}(y_{ij}; x_{ij}, \theta_1) + \pi_{2ij} f_{2ij}(y_{ij}; x_{ij,} \theta_2) \quad (9.1)$$

Notice that the density function and parameters have the suffix ij, indicating that the parameter value corresponds to patient *i* in hospital *j*.

Similar work was conducted by Yau et al. (2003) who fitted a Gaussian mixture regression model with random effects on neonatal LoS data to account for between hospital variations.

9.3.3. Understanding inappropriate hospital LoS

In the last 10 years the admission rates of elderly people at public hospitals in Mexico have increased considerably. With population ageing, the healthcare facilities for long term or rehabilitation patients are often scarce, putting public hospitals (designed originally for acute care only) under too much pressure and forcing them to frequently incur in an inappropriate use of hospital days. Inappropriate hospital stays occur when a patient has technically finished his/her treatment but he/she cannot be discharged due to the lack of long term care facilities or the appropriate quality of care at home and therefore it is forced to stay at hospital while consuming resources.

Let us recall that the models (and associated factors) presented in this thesis reflect the duration of a patient's treatment only. Identifying whether the length of stay is appropriate may improve performance of such models. Indeed, most of the public healthcare providers in Mexico have developed their own version of the US Appropriateness Evaluation Protocol (AEP) which consists in a number of explicit and objective criteria to identify inappropriate hospital admission and stays (Gertman and Restuccia, 1981).

A piece of future work would therefore be the use of such criteria to identify patients with inappropriate LoS and either exclude them from data when re-estimating the finite mixture models (and descendant models) or to adjust their LoS to account just for the time that corresponds to duration of the patient treatment.

9.3.4. Time-dependent predictor variables

So far in this research it has been assumed that the predictor variables are fixed over time. However some predictors actually vary over time such as: age, number of comorbidities, number of previous hospitalisations, occupation, educational level, inherited family history etc. Moreover, other important information exists that is usually difficult to include in standard statistical models, such as biomarkers, which may be highly dependent on time. A biomarker is substance whose detection or change of state may indicate a particular disease state (e.g. levels of glucose are frequently associated with diabetes).

One of the possible approaches to accommodate this type of variables in the models (developed in this thesis) is to adjust the hazard function $h_s(y_i)$ of the finite mixture model (see Section8.1) to include time-varying predictors. These types of models in survival analysis are known as time-dependent proportional hazard models (Fisher and Lin, 1999)

9.3.5. Handling mortality

There are four possible causes of discharge from hospital: recovery, transfer to another healthcare facility, voluntary discharge and death.

In this research, no discrimination was made between different types of discharge because only the variables that can be known on admission were included. However, patients who die in hospital might be different in many ways from the rest of the patients (age, severity of the illness, complications, comorbidities, etc.). In medical research, some authors have recognised the importance to make a separation of these patients from the rest, since their inclusion might lead to biased results or erroneous conclusions (Brock et al., 2011).

A piece of further work would be to study whether different conclusions can be drawn by separating or handling the deaths. Some suggestions to deal with this type of data are:

- Disregarding LoS data from individuals who die, although some valuable information may be lost.
- Analysing this data separately from the rest by following the same methodology outlined in this thesis. This was not an option for the current research since the sample size of deaths was very small (i.e. less than 1.7% of the total sample size), complicating any statistical analysis.
- Using survival analysis and labelling deaths as censored observations.
- Using more advanced models such as Brock et al. (2011) who proposed a method using multi-stage models where different states corresponds to the different discharge options represented by recovery and death. Then such options are represented as competing

events, indicating that a transition into either one of the two states precludes a transition into other. Similarly, the different discharge options can be modelled as multiple absorbing states in a mixed Coxian phase-type distribution (McClean et al., 2010)

9.4. Conclusions and novel contributions

This PhD thesis attempted to satisfy the research objectives defined in Chapter 1 while answering six research questions. In retrospect, after the completion of the study, it is possible now to evaluate how well the research objectives and questions were addressed.

This research was geared towards developing a statistical model to predict patient length of stay (LoS) in Mexican public hospitals that:

A. Captures the variability of the LoS distribution

The embedded variability of LoS was successfully captured by using finite mixture models. The advantage of finite mixture models is that they assume that the LoS distribution is intrinsically a linear combination of two or more sub-distributions of LoS, where each sub-distribution is a local model of some part of the true distribution. The LoS characteristics such as skewness and heavy tails were easily modelled by finding a sub-distribution that represents them.

Moreover, by addressing the heterogeneity problem (see below) it was possible as well to capture and minimize part of the variability.

B. Recognises and addresses the heterogeneity problem

One of the causes of LoS variability is the heterogeneity problem defined in the introductory chapter of this thesis. In this research heterogeneity was tackled by creating homogenous groups of patients and understanding how internal and external factors delineate such groups. In this context, heterogeneity was recognised to occur at three different contexts: between patients, within hospitals and between hospitals. Consequently three different approaches were designed to tackle this problem:

- Individual-based model: where patient attributes shape the LoS distribution of each individual. Here, the finite mixture model was extended to account for covariates using a generalised linear model principle.
- Group-based model: where each group of cohort of patients correspond to a component of the finite mixture model. Different data mining techniques were explored to understand the relationships between each group and patients attributes.

- Multilevel group-based model: The group-based approach was extended to accommodate variation between hospitals and other environmental and contextual variables.

C. Supplies LoS predictions for individual patients and cohort of patients (within and between hospitals)

   The three approaches (mentioned above) allowed estimations of expected LoS value and probabilities for individual patients (individual-based approach), groups of patients within hospitals (group-based approach) and group of patients between hospitals (multilevel group-based approach).

D. Demonstrates a solid application into the decision-making process

   In chapter 8 it was proposed how the individual and group approach can be exploited in order to provide insights into the patient flow. In addition, two new methods for calculations of bed requirements were proposed based on the estimate of expected LoS and the survival function. Moreover it was exemplified how the multilevel group-based approach can be a useful tool for two different users: the decision maker interested in a strategic level of resource planning and the decision maker at a local hospital, interested in the day to day planning of resources.

In addition, the following research questions were posed:

Can a statistical model approximate to the underlying LoS distribution?

The two-component mixture models developed in Section 4.1.1 and 4.1.3 successfully approximate to the LoS distribution. Visually one can see how the models mimicked accurately the main characteristics of the distribution such as the early peaks and long tails. In addition, Log likelihood, AIC and BIC values supported the model goodness of fit.

Conceptually Gamma and Lognormal are (for a couple of reasons) adequate models for LoS: Firstly, they are the usual choices when modelling positive, skewed, continuous variables such as LoS. Secondly, the shape of the hazard function for both distributions reflects the real nature of the patient flow: for the MRC hospital where the proportion of patients having a medium-long LoS is higher, the hazard function takes an inverted U shape, indicating that during the first days of admission the risk of being discharged is constantly increasing, to then becoming steady for a while and later starting to decrease over time. This gradual fall of the risk of being discharged means that the longer a patient stays at hospital the longer it will take to be discharged. This is a very common characteristics of geriatric patients (Harrison and Millard, 1991). On the other hand for the ISSEMyM hospital where the proportion of patients with short-medium LoS is by far the higher, the hazard function increases monotonically, indicating that the risk of being discharged tends to increase over time.

Also, by recognising and addressing the heterogeneity problem the models developed in this research become closer to reality, where the patient LoS distribution is governed by different internal and external factors.

Is it possible to use the same model for other hospitals or does each hospital needs its own customised model?

To satisfactorily answer this question, further research needs to be conducted. From Chapter 4, it was clear that each hospital needed a finite mixture model of very distinct nature. A two-component Lognormal model was more appropriate for MRC hospital whereas a two-component Gamma model was a better choice for ISSEMyM. However in Chapter 8, a Kruskal-Wallis test supported the use of the two-component lognormal model for a bigger sample including 19 hospitals. The most feasible explanation is that because the MRC hospital and the 19 hospitals from the regional sample belong to the same healthcare supplier and provide the same level of care, it is very likely that they share procedures and policies, and even treat similar type of population. Therefore it is expected that they have similar LoS behaviour.

However it is early to conclude if the two-component lognormal model can (or cannot) be applied for hospitals of other levels of care and healthcare providers. As far as this research could surmise, these models are suitable for second level hospitals belonging to the Secretariat of Health.

Which are the internal and external factors that affect LoS distribution and what it is the nature of this influence?

As it was stated in the literature review, an extensive number of studies have been more interested in understanding the variables that influence LoS rather than making predictions or estimations of patient LoS. One possible cause of such interest is that with the enormous pressure from outside authorities to reduce LoS, the medical staff and decision makers are interested in gaining some control over the LoS patterns. Therefore, there is great interest on identifying the factors involved on patient LoS, mainly those ones that can be controlled by human intervention.

The LoS predictors in the third-level ISSEMyM general hospital (according to the finite mixture of generalised linear models) were origin of the patient, surgical procedure, total number of current illnesses, ward where the patient is treated, patient age, previous blood transfusions, number of comorbidities, drinking and smoking behaviour, diagnosis, first diagnosis and number of previous hospital admissions. Conversely in the second-level MRC general hospital the predictors (according to the logit model) were patient age, number of previous hospital

admissions, origin of the patient, ward where the patient is treated, diagnosis and surgical procedure to be undergone.

Furthermore, those predictors helped to delineate the characteristics of each LoS-homogenous subpopulation of patients. For example at the ISSEMyM hospital, those who are older, have a diagnosis from category 2 (i.e. diabetes mellitus, stroke, hepatic failure, cirrhosis, gastrointestinal haemorrhage, etc.) or underwent a surgical procedure category 2 (i.e. appendectomy, bowel endoscopy, laparoscopic cholecystectomy, etc.) were more likely to be in component 2, with a long LoS (i.e. more than 12 days at hospital). On the other hand, at the MRC hospital, those patients who are older, male, have few previous admissions to hospital, enter to the hospital via A&E and are treated in the adult medicine ward were significantly more likely to have a medium-long LoS.

Moreover, different predictors had different effects for different groups of patients. Some variables were significant just for one of the LoS categories and non-significant for the other. For example: the effect of surgical procedure category 3 was very significant for MRC patients belonging to the short-medium LoS category (i.e. the LoS for a patient undergoing a surgical procedure of this category was almost 80% higher than that for a patient not undergoing surgical procedures). However this effect was not significant in patients belonging to long LoS category, indicating that surgical procedures under category 3 have influence just on patients with short-medium LoS rather than with long LoS.

On the other hand, according to the multilevel group-based model, the effect of the ward where the patient is treated differs across hospitals, indicating that different hospitals might have different ways of running their wards and this has an effect on the patient LoS. In addition, patients treated at hospitals with a number of operating rooms above the regional average have a higher probability of having medium-long LoS than of short LoS.

Can a statistical model be clinically and/or operationally meaningful?

The models developed here give insightful knowledge through the understanding of the patient population and its interaction with the hospital (and system). In this context, unlike some cases of phase-type distribution (see Section 2.4.1), the optimal number of components for each finite mixture model has actually an interpretation in the real world, where every component can be regarded as a subpopulation of patients with different LoS patterns. Each subpopulation LoS is delineated by a number of variables, including clinical variables, which can observable and quantified by the medical staff such as initial diagnosis, diagnosis, presence of comorbidities, surgical procedures, etc.

Moreover the density function of the models represents the duration of medical treatment and when the survival function is used instead, it reflects how the beds in the hospital are being used, moving from a clinical perspective to an operational perspective.

What type of information can be derived from the model that can be incorporated in a decision-making process?

Firstly the finite mixture models represent the duration of the patient treatment from where the probability or likelihood with which specific values of the variable LoS will occur can be ascertained.

In addition, the models allow identification of two LoS-homogenous groups of patients or subpopulations. The proportions of each group vary from one hospital to the other. Knowing the proportions of each type of subpopulations allows the hospital to identify different types of workload: for example, the proportions of patients with longer LoS are 18% at ISSEMyM and 45% at MRC hospital. Therefore bed-blocking might be a more frequent phenomenon at the MRC hospital than at ISSEMyM. This valuable information can be used to plan the number of beds needed according to the type of stay.

Not just the factors that influence LoS were identified but their size and direction. In the individual-based approach, it was possible to identify if a particular variable increases or decreases the LoS (and in which proportion this occurs) in comparison to a baseline scenario. In the group-based approach, it was possible to identify the factors associated with an increment (or decrement) in the likelihood of a patient belonging to a certain LoS category in comparison to a baseline scenario.

The choice of which approach to use is the decision of the final users based on their needs. Both approaches have different operational profiles: the individual-based approach allows tracking and tracing individual patients which may be useful in situations such as clinical studies or when there is a particular interest in a specific patient. However, tracking individual patients is costly, time-consuming, and constantly requires proactive attention from the medical staff. On the other hand, the group-based approach allows tracking homogenous groups of patients which may simplify planning tasks. However, this research suggested, based on the results of Chapter 5 and 6, that an individual based-approach is more appropriate for the ISSEMyM hospital in order to avoid misclassification patients.

From the multilevel group-based approach other interesting information can be derived: the model allows the calculation of probabilities (of belonging to a certain LoS category) for a patient on a baseline scenario in an “average” hospital and in this context, it is possible to identify which hospitals differ considerably from that average. Further, it is possible to make

probability estimations for different patient scenarios that could be expected in most of the sampled hospitals. Finally the model allows identification of how much of the patient LoS variability was attributable to the hospital and how much to the patient characteristics.

Finally a number of useful output measurements can be derived from the three different models (individual-based, group-based and multilevel group-based model):

- Expected length of stay for a particular patient or group of patient within and between hospitals.
- The probabilities that a particular patient will be discharged by a specific day or before.
- The probabilities that a particular patient will still be retained at hospital after certain time.
- The probabilities that a particular patient who has been in the hospital for a specific amount of time will be discharged in the next 24 hours.
- The probabilities that a patient will still remain in hospital for a specific number of days later, given that he or she has already been at hospital for a particular number of days.
- The probabilities that a patient will be discharged a specific number of days later, given that they have been at hospital for a particular number of days.
- The number of patients unlikely (and likely) to leave in the next 24 hours.
- The number of patients who will be discharged in the forthcoming days.
- Bed requirements calculations for different scenarios.

Furthermore, in the Appendix A (i.e. "Other classification methods"), alternative models for the group-based approach were presented. These models give a different interpretation to the effects of the predictors. For example the multinomial logit model (MNLM) compares the likelihood of being at the base LoS category with the likelihood of being at one of the other LoS categories and the stereotype ordered regression model (SORM) compares the likelihood of being at the lowest numbered category with the likelihood of being at the highest one. In the other hand, the generalised ordered logit model (GOLM) and the partial proportional odds model (PPOM) are very useful in understanding patient progression regarding to their LoS category, either moving upward to longer LoS or downward to shorter LoS.

Can this model derived from routinely collected data be accurate in predicting LoS?

The LoS data extracted directly from both hospitals data sets was suitable enough to build the finite mixture model which accurately captured the nature and variability of patient LoS.

However when patient attributes and other factors were included into the models some challenges were faced: in addition to the data limitations mentioned in Section 9.2.1, the

variable selection process (Section 5.1) suggested that the significant variables available in the data set might not be enough to explain LoS variability. This became more evident when in Chapter 6 most of the classification algorithms (used to predict LoS category for ISSEMyM patients) failed to predict long LoS. It might be the case that the information contained in the datasets is not enough to predict accurately LoS category (for the ISSEMyM dataset only) and the inclusion of other variables which are not being currently recorded might help to improve the performance of the models.

Nevertheless, the individual-based model for ISSEMyM and the group-based model for MRC hospital are efficient models to predict LoS using the data available currently routinely in both hospitals.

9.4.1. Contributions

Finally this chapter closes by enumerating the contributions of this thesis:

This thesis considers the hospital full case-mix. Most of the models and methods described in the literature review focused on a particular patient cohort or patient population: geriatric patients, stroke patients, patients with mental diseases, patients undergoing major surgery, etc. The models developed here included the entire case-mix of both hospitals during 2005-2009, including around 800 different diagnoses and 200 different surgical procedures, divided in two hospital wards.

The research proposes an alternative methodology to cluster categorical variables. A methodology to reduce the number of ICD codes of the variables “first diagnosis”, “diagnosis” and “surgical procedure” was proposed using hierarchical cluster methods based on the chi-square dissimilarity measure (see Section 4.2). Hierarchical cluster methods and clustering methods in general are often used to assign objects (e.g. patients, products, countries, companies, etc.) into clusters; however in this research they were used to assign categories of nominal variables into clusters. This methodology enabled the inclusion of the entire hospital case-mix in the models developed in this thesis, and provided a simple classification scheme for ICD codes in ISSEMYN and MRC hospitals.

The thesis provides a comparative study of different mixtures of distributions. Although finite mixture models have previously been used to model LoS, most of the research has focused on the Gaussian mixture model. Very few studies explored and compared other distributions: one of the few such studies is the work by Singh and Ladusingh (2010) that compared Poisson and negative binomial mixtures for LoS data in India. This current research contributes to the field with a comparative study of the most common (continuous and discrete) distributions for lifetime data that have previously been successfully used in modelling LoS as single models,

including Lognormal, Gamma, Gaussian and Poisson distributions. The results suggested that a two-component Lognormal and two-component Gamma mixture models were more appropriate to represent the nature of LoS, whereas the mixtures of Poisson were the less appropriate.

The thesis provides a comparative study of different data mining algorithms. This research contributes a comparative study of thirteen data mining algorithms to classify patients according to their LoS, including the most common algorithms such as logistic regression, classification trees and Naïve Bayes. In addition, new to LoS data, Logistic Regression trees and Naïve Bayes trees were included along with the novel ensemble methods Boosting and Bagging. One of the most interesting findings of the comparative study is that the accuracy rates of many algorithms were sufficiently similar that their differences were statistically insignificant. The differences are also probably insignificant in practical terms. It seems that the algorithms explored in this thesis reached the limit of information that can be predicted from the current data and there is no room for future improvement.

This thesis introduces a hierarchical group-based model. As was mentioned previously, the use of a group-based approach to model patient LoS is not new in the field. Similar models under the name of DCS models have being used in the last couple of years. However, the contribution of this research is an extension of the conditional component (logit model) to account for the effect of the environment (where the patient is treated) on the classification of patients into LoS-homogenous groups. The extended model reflects the real hierarchical structure of the LoS data, which consists of units grouped at different levels (at level-1, the units are patients and at level-2 they are hospitals). This type of analysis tries to address one of the issues mentioned in the introductory chapter by taking into account that the outcome variable has both an individual and a cluster effect to its variability. Furthermore in the previous chapter it was demonstrated how this model can be exploited by two different users: the decision maker at a large scale (i.e. government), interested in a strategic level of resource planning (i.e. national, regional or institutional level); and the decision maker (hospital manager or medical staff) at a local hospital, interested in the day to day planning of resources.

This thesis introduces the applications of survival analysis for finite mixture models. As far as this thesis is concerned, there is no previous research on the applications of survival analysis in finite mixture models. This research contributes to the field by reviving the use of the survival and hazard functions on finite mixture models which enhanced the interpretability and application of the models developed here. For example the use of survival analysis allowed identification of patients that are more likely to be discharged in the next 24 hours (hazard function) or the probability of being discharged after *n* days given that the patient has already been m days at hospital. The inclusion of survival analysis not only allows making LoS

predictions at the moment of the hospital admission but at any moment after admission. In addition, the inclusion of survival analysis allowed the development of the survival occupancy table (Section 8.1.2) for calculation of bed requirements, which can be an alternative for hospitals that do not have a formal methodology or whose current procedures are inadequate to calculate bed requirements.

The thesis provides a comparative study of different logistic regression models for ordinal data (See Appendix A). Most of the previous research that uses discretised versions of LoS as a categorical outcome variable (i.e. whereby the continuous variable is split in different intervals according to a certain criteria), thereby ignoring its embedded ordinal nature. This research contributes a comparative study of different logistic regression models to deal with ordinal outcome variables and to interpret such ordinality according to different research questions.

Overall, this thesis meets the research objectives of developing a statistical model to predict patient LoS in public hospitals in Mexico which approximates to the distribution of the LoS, recognises and addresses the heterogeneity population problem, supplies LoS predictions for individual patients and cohorts of patients within and between hospitals and demonstrates a solid application into the decision-making process. Finally, therefore, locating it within the existing literature knowledge, this thesis extended previous models and employed new methods within the area of predicting patient LoS. This thesis focused on specific contributions to the understanding of patient LoS and the methods that have been previously and traditionally exploited within this area of research. In addition, in terms of the geographical context, the applications of the models developed are easy to implement and understand and provide useful information in predicting patients LoS at Mexican hospitals at all levels of healthcare planning.

Appendix A

Other classification methods

During the initial encounters with the MRC hospital staff, two general concerns related to patient LoS came out. Firstly, they were interested in understanding how different factors (patient characteristics and other internal and external factors) influence patient LoS at hospital and, secondly whether it is possible to predict beforehand (based on those factors) how long a patient will stay at their hospital.

However, it is just recently that the MRC hospital is incorporating planning strategies to its practices. One of these strategies was to classify patients according to their LoS into three categories: “short LoS” (patients with LoS up to 3 days), “Medium LoS” (patients with LoS from 4 to 11 days) and “long LoS” (patients with LoS from 12 days). This classification was made merely based on empirical observation and personal judgment.

Accordingly, a very early stage of this research was devoted exclusively to explore different alternatives to bring answers to the previous questions, working under the existing patient LoS classification scheme described above. The approach was equivalent to the group-based, where patient attributes or variables were used to predict the LoS category to which the patient belongs. Therefore, this chapter describes the first research journey (when just data from MRC hospital was available) and the valuable lessons that were learnt from it.

Hierarchical clustering was used as Section 4.2 outlined in order to cluster both diagnosis and surgical procedure into categories with similar length of stay, taking into account there are now three LoS categories. In addition, a variable selection process was performed according to the methodology followed in Section5.1.

Initially, it was decided to start with a more classical approach and select a well-established method for classification and prediction as logistic regression methods. Therefore, different models from the family of the logistic regression models were analysed as possible classification/prediction tools of patient LoS category: Ordinal regression model, generalised regression model, partial proportional odds model multinomial Logit model, and stereotype

ordered regression. For each model a brief introduction of the theoretical foundation behind each model and the results of the estimation of the parameters are presented, giving a brief interpretation of the results applied to the context of LoS.

A.1. Ordinal Regression Model

Ordinal regression model (ORM), commonly known as the proportional odds model or cumulative odds model, is the preferable option for predicting LoS category among the family of regression models for categorical variables, because it assumes ordinality of the outcome, as it happens with the variable LoS category where the order of the coding actually has a ranked meaning (0 for short, 1 for medium and 2 for long LoS).

The ORM can be defined as a probability model in Equation (A.1):

$$\ln \frac{\Pr(y \le j \mid \tilde{x})}{\Pr(y > j \mid \tilde{x})} = \tau_j - \tilde{x}\beta_j \quad \text{For } j = 1 \text{ to } J-1 \tag{A.1}$$

Where $\tilde{x}$ is the vector of independent variables, βs are the slope coefficients, τ_j are the cut points and J is the number of categories of the ordinal dependent variable. The meaning of the cut points will become clearer shortly when the latent variable model is explained.

The predicted probabilities are calculated as:

$$\Pr(y = 1 \mid \tilde{x}) = \frac{\exp(\tau_1 - \tilde{x}\beta_1)}{1 + \exp(\tau_1 - \tilde{x}\beta_1)} \tag{A.2}$$

$$\Pr(y = j \mid \tilde{x}) = \frac{\exp(\tau_j - \tilde{x}\beta_j)}{1 + \exp(\tau_j - \tilde{x}\beta_j)} - \frac{\exp(\tau_{j-1} - \tilde{x}\beta_{j-1})}{1 + \exp(\tau_{j-1} - \tilde{x}\beta_{j-1})} \quad \text{For } j = 2 \text{ to } J-1 \tag{A.3}$$

$$\Pr(y = J \mid \tilde{x}) = 1 - \frac{\exp(\tau_{J-1} - \tilde{x}\beta_{J-1})}{1 + \exp(\tau_{J-1} - \tilde{x}\beta_{J-1})} \tag{A.4}$$

When $J = 2$ the model is equivalent to the Logit regression model and when $J > 2$ the model becomes equivalent to a series of binary logistic regressions where categories of the dependent variable are combined. For example in the MRC dataset $J = 3$ which represents three categories: short, medium and long LoS. Therefore, when $j = 1$, category short LoS is contrasted with categories medium and long LoS and when $j = 2$, category short and medium LoS is contrasted with category long LoS.

The ORM is often formulated as a latent variable model, defined as:

$$y_i' = \tilde{x}\beta + \varepsilon_i \tag{A.5}$$

$$y_i = j \text{ if } \tau_{j-1} \leq y_i' < \tau_j \quad \text{for } j\text{=1 to } J \tag{A.6}$$

where y_i' is the latent variable ranging from ∞ to -∞, and ε_i is the random error.

The continuous latent variable y_i' can be thought of as the propensity of a patient to belongs to a certain category (Long, 1997).The LoS category of a patient now relies on the latent variable:

$$y_i = Short \text{ if } \tau_0 \leq y_i' < \tau_1$$

$$y_i = Medium \text{ if } \tau_1 \leq y_i' < \tau_2$$

$$y_i = Long \text{ if } \tau_2 \leq y_i' < \tau_3 = \infty$$

Thus when the latent variable crosses a cut point τ_j the patient category changes.

A special type of ordinal model is the Continuation ratio model in which the categories represent levels, where the lowest level must occur before the second, the second before the third, and so forth until the highest level (Hilbe, 2009). It can be thought as stages in some process through which an individual can advance. A key characteristic of the process is that an individual must pass through each stage (Long, 1997). This special characteristic suits with the nature of the patient journey through the hospital where the patient can evolve from a short to a medium LoS and so on.

The continuation ratio model is defined as:

$$\ln \frac{\Pr(y = m \mid \tilde{x})}{\Pr(y > m \mid \tilde{x})} = \tau_m - \tilde{x}\beta \quad \text{For } m\text{=1 to } J\text{-1} \tag{A.7}$$

Where *m* is the stage and *J* is the number of categories of the outcome variable.

The predicted probabilities are calculated by:

$$\Pr(y = m \mid \tilde{x}) = \frac{\exp(\tau_m - \tilde{x}\beta)}{\prod_{j=1}^{m}[1 + \exp(\tau_j - \tilde{x}\beta)]} \quad \text{For } m\text{= 1 to } J\text{-1} \tag{A.8}$$

$$\Pr(y = J \mid \tilde{x}) = 1 - \sum_{j=1}^{J-1} \Pr(y = j \mid \tilde{x}) \quad \text{(A.9)}$$

ORMs and continuation ratio models are based on the parallel regression assumption or proportional odds assumption. This assumption sustain that the ordinal model is equivalent to $J-1$ binary regressions, where the slope coefficients are identical across each regression. Assuming the equality of slopes among categories allows interpreting the model in the same way for all categories, making it more parsimonious. However there is a general consensus that the parallel regression assumption is quite stringent and the chance of all the dependent variables in the model having identical slope coefficients is likely to be quite rare (Lall et al., 2002). This assumption was tested in the MRC dataset using the `omodel` command on STATA. The command computes a likelihood ratio test and compares the log-likelihood from ORM to that obtained from pooling J-1 binary logistic models. Table A.1depicts the results.

Leaving aside the interpretation of the parameters, notice that the *p-value* of the chi-squared (bottom of Table) indicates that the null hypothesis that the model parameters are equal across categories can be rejected at the .0001 level. When statistical assumptions are broken as they are in the current dataset, the model cannot be accurately applied to the whole population (the parameters of the model are said to be biased). In other words it is not possible to draw conclusions about the population, although valid estimates of an ORM for the sample were generated.

Due to the restrictions of the parallel assumption, other models have been presented in statistical literature as alternatives to the ORMs, including: generalised ordered logit model (GOLM), partial proportional odds model (PPOM), multinomial Logit model (MNLM) and stereotype regression model (SORM) are the most common alternatives.

Log-likelihood	--5729.0052
LR chi^2(13)	2559.6
Prob > chi^2	0.0000
Pseudo R^2	0.1826

	β	**Std. Err.**	***z***	***P>z***	**[95% Conf.**	**Interval]**
Female	0.1626	0.0501	3.2400	0.0010	0.0643	0.2609
Previous admissions	-0.0095	0.0034	-2.8000	0.0050	-0.0161	-0.0028
Age	0.0103	0.0014	7.4300	0.0000	0.0076	0.0130
General surgery ward	-0.3427	0.0739	-4.6400	0.0000	-0.4875	-0.1979
Outpatient clinic	-0.5912	0.1208	-4.9000	0.0000	-0.8279	-0.3545
Number of s. procedures	0.5629	0.0703	8.0100	0.0000	0.4252	0.7007
Diagnosis_category2	1.2559	0.1832	6.8500	0.0000	0.8967	1.6150
Diagnosis_category3	1.6321	0.1847	8.8400	0.0000	1.2702	1.9940
Diagnosis_category4	3.2461	0.2111	15.3700	0.0000	2.8323	3.6599
Diagnosis_category5	2.3750	0.1862	12.7500	0.0000	2.0100	2.7401
Sp_category2	0.7638	0.1854	4.1200	0.0000	0.4003	1.1272
Sp_category3	-0.8799	0.1133	-7.7700	0.0000	-1.1019	-0.6579
Sp_category4	-1.4895	0.1091	-13.6500	0.0000	-1.7034	-1.2757
_cut1	1.8655	0.1881				
_cut2	4.61099	0.1943				

Approximate likelihood-ratio test of proportionality of odds across response categories:	chi2(13)	104.61
	Prob > chi2	0.0000

Table A.1:Ordinal logistic regression model and test for parallel assumption

A.2. Generalised Ordered Logit Model

The generalised ordered logit model (GOLM) allows the slope coefficients to differ for each *J-1* binary regressions as represented in Equation (A.10)

$$\frac{\Pr(y \le j \mid \tilde{x})}{\Pr(y > j \mid \tilde{x})} = \tau_j - \tilde{x}\beta_j \text{ For } j = 1 \text{ to } J-1 \qquad \text{(A.10)}$$

The predicted probabilities are calculated as:

$$\Pr(y=1 \mid \tilde{x}) = \frac{\exp(\tau_1 - \tilde{x}\beta_1)}{1+\exp(\tau_1 - \tilde{x}\beta_1)} \tag{A.11}$$

$$\Pr(y=j \mid \tilde{x}) = \frac{\exp(\tau_j - \tilde{x}\beta_j)}{1+\exp(\tau_j - \tilde{x}\beta_j)} - \frac{\exp(\tau_{j-1} - \tilde{x}\beta_{j-1})}{1+\exp(\tau_{j-1} - \tilde{x}\beta_{j-1})} \text{ For } j = 2 \text{ to } J\text{-}1 \tag{A.12}$$

$$\Pr(y=J \mid \tilde{x}) = 1 - \frac{\exp(\tau_{J-1} - \tilde{x}\beta_{J-1})}{1+\exp(\tau_{J-1} - \tilde{x}\beta_{J-1})} \tag{A.13}$$

Notice that the equations for the GOLM are similar to the ORM. GOLM retains the nature of the ORM, which considers simultaneously the effects of a set of independent variables across successive dichotomizations of the outcome (O'Connell, 2006). Yet the slope coefficients β_j are not delineated by the parallel regression assumption.

The GOLM was fitted on STATA using the `gologit` function. The next figure shows the output

Log-likelihood	-5678.135
LR chi^2(26)	2661.34
Prob > chi^2	0.0000
Pseudo R^2	0.1899

	β	Std. Err.	z	$P>z$	[95% Conf.	Interval]
Short LoS						
Female	0.1331	0.0524	2.5400	0.0110	0.0303	0.2359
Previous admissions	-0.0090	0.0034	-2.6700	0.0080	-0.0157	-0.0024
Age	0.0096	0.0015	6.5900	0.0000	0.0067	0.0124
General surgery ward	-0.3581	0.0781	-4.5800	0.0000	-0.5113	-0.2050
Outpatient clinic	-0.5738	0.1225	-4.6800	0.0000	-0.8140	-0.3337
Number of s. procedures	0.5509	0.0737	7.4700	0.0000	0.4064	0.6954
Diagnosis_category2	1.2234	0.1847	6.6300	0.0000	0.8615	1.5853
Diagnosis_category3	1.6425	0.1857	8.8400	0.0000	1.2785	2.0065
Diagnosis_category4	3.3890	0.2322	14.6000	0.0000	2.9340	3.8441
Diagnosis_category5	2.4775	0.1882	13.1600	0.0000	2.1085	2.8464
Sp_category2	1.2091	0.2768	4.3700	0.0000	0.6667	1.7515
Sp_category3	-0.8343	0.1167	-7.1500	0.0000	-1.0630	-0.6055
Sp_category4	-1.4464	0.1105	-13.0900	0.0000	-1.6630	-1.2299
cons	-1.8786	0.1903	-9.8700	0.0000	-2.2517	-1.5056
Medium LoS						
Female	0.3832	0.1011	3.7900	0.0000	0.1851	0.5812
Previous admissions	-0.0138	0.0083	-1.6600	0.0960	-0.0301	0.0025
Age	0.0151	0.0027	5.5600	0.0000	0.0098	0.0205
General surgery ward	-0.2648	0.1429	-1.8500	0.0640	-0.5450	0.0154
Outpatient clinic	-0.9858	0.3650	-2.7000	0.0070	-1.7012	-0.2704
Number of s. procedures	0.5998	0.1267	4.7300	0.0000	0.3515	0.8481
Diagnosis_category2	1.1148	0.4274	2.6100	0.0090	0.2770	1.9525
Diagnosis_category3	0.9610	0.4393	2.1900	0.0290	0.1001	1.8220

Diagnosis_category4	2.2966	0.4455	5.1600	0.0000	1.4234	3.1698
Diagnosis_category5	1.2790	0.4317	2.9600	0.0030	0.4329	2.1252
Sp_category2	0.4890	0.2656	1.8400	0.0660	-0.0316	1.0095
Sp_category3	-1.3966	0.2832	-4.9300	0.0000	-1.9517	-0.8415
Sp_category4	-1.3614	0.2252	-6.0400	0.0000	-1.8029	-0.9200
cons	-3.9248	0.4386	-8.9500	0.0000	-4.7845	-3.0652

Table A.2: Generalised Ordered Logit STATA output

The output of Table A.2is divided into two panels: the first panel contrasts category 0 (short LoS) with category 1 and 2 (medium and long LoS), and the second panel contrast categories 0 and 1 (short and medium LoS) with category 2 (long LoS). In terms of interpretation, positive coefficients of the parameters indicate that higher values in the independent variables make it more likely that the patient belong to an upper category of LoS category than the current one, while negative coefficients indicate that higher values on the independent variable increase the likelihood of belonging to the current or to a lower category of LoS category (Williams, 2006)

The results of Table A.2can be interpreted in terms of odds ratios as it was described in Section 6.1 (i.e. e^{β})

Diagnosis category 4: The positive coefficient of diagnosis category 4 (HIV, respiratory tuberculosis, pleural effusion, etc.) is higher in the first panel, indicating that a patient with a diagnosis category 4 is more likely (29.63 times more) to have a medium or long LoS rather than a short one[62].

Outpatient clinic: The negative value of the coefficient of ward is higher in the first panel, indicating that a patient who enters to the hospital via outpatient clinic is more likely (1.43 times more) to have a short LoS rather than a medium or long one[62,63].

Diagnosis category 2: The positive coefficient of diagnosis category 2 (chronic renal failure or ventral hernia) is higher in the first panel, indicating that a patient with a diagnosis category 2 is more likely (3.39 times more) to have a medium/long LoS rather than a short one[62].

Surgical procedure category 3: The negative value of the coefficient of surgical procedure category 3 is higher in the second panel indicating that a patient undergoing one of those

[62]This interpretation is true only if the effects of the other variables are held constant.
[63]To ease the interpretation of the parameters for all the models, when the odds ratio was lower than one (negative effect), the order of presenting odds was reversed and the inverse of the odds ratio was used as the new factor change.

surgical procedures (e.g. Cholecystectomy, appendectomy, tonsillectomy, etc.) is more likely (4.04 times more) to have a short or medium LoS than a long one[62].

A.3. Partial Proportional Odds Model

One of the drawbacks of the GOLM is that it includes many more parameters than the ORM. As a result of setting free all variables from parallel line constraints.

Although it is very common to find that the parallel assumption has been violated, usually not all the slope coefficients of the model transgress the assumption. The partial proportional odds model (PPOM) imposes constraints for parallel lines just where they are needed. In other words, some slope coefficients can be the same for all values of *J*, while others can differ, and hence, avoids including unnecessary extra parameters in the model. Equation (A.10) is extended to accommodate the unconstrained parameters which violated the assumption (see equation A.14):

$$\frac{\Pr(y < j \mid \tilde{x})}{\Pr(y > j \mid \tilde{x})} = \tau_j - \left(\tilde{x}\beta_j + \left[x_q\beta_q + T_q\gamma_{jq}\right]\right) \text{ For } j = 1 \text{ to } J-1 \qquad \text{(A.14)}$$

Here $\tilde{x}$ is the vector of independent variables where *q* of them are known to violate the parallel assumption. T_q exists only for the *q* variables that violate the parallel assumption. Thus γ_{jq} are non-zero coefficients for the *q* variables and zero otherwise, and they are the components of the log odds that vary over the different categories (Lall et al., 2002).

The `gologit2` command on STATA with the `autofit` option identifies which variables violated the parallel lines assumption and imposes constraints on those where the assumption is not violated and then the model is re-estimated. STATA imposed seven constraints in the final model for "number of previous visits", "age", "ward", "origin", "number of surgical procedure to undergo", "diagnosis category 2" and "surgical procedure category 4". Table A.3shows the final output of STATA. Note that the parameter estimates for those constrained variables are the same in both panels and their interpretation can be the same as in ordinal regression:

The odds of having a longer LoS are 3.38 times larger for patients with a disease classified in the diagnosis category 2 like chronic renal failure or ventral hernia.

The odds of having a shorter LoS are 1.40 times larger for patients who enter to the hospital via outpatient clinic.

				Log-likelihood		-5681.3463
				LR chi^2(19)		1894.59
				Prob > chi^2		0.0000
				Pseudo R^2		0.1894
	β	**Std. Err.**	***z***	***P>z***	**[95% Conf.**	**Interval]**
Short LoS						
Female	0.1342	0.0524	2.5600	0.0100	0.0314	0.2370
Previous admissions*	-0.0092	0.0034	-2.7100	0.0070	-0.0158	-0.0025
Age*	0.0105	0.0014	7.4800	0.0000	0.0077	0.0132
General surgery ward*	-0.3395	0.0748	-4.5400	0.0000	-0.4861	-0.1929
Outpatient clinic*	-0.5963	0.1215	-4.9100	0.0000	-0.8345	-0.3581
Number of s. procedures*	0.5603	0.0704	7.9600	0.0000	0.4223	0.6984
Diagnosis_category2*	1.2194	0.1837	6.6400	0.0000	0.8594	1.5795
Diagnosis_category3	1.6375	0.1853	8.8400	0.0000	1.2743	2.0008
Diagnosis_category4	3.3878	0.2317	14.6200	0.0000	2.9337	3.8420
Diagnosis_category5	2.4718	0.1873	13.2000	0.0000	2.1048	2.8389
Sp_category2	1.1886	0.2755	4.3100	0.0000	0.6486	1.7286
Sp_category3	-0.8404	0.1145	-7.3400	0.0000	-1.0648	-0.6161
Sp_category4*	-1.4474	0.1087	-13.3100	0.0000	-1.6605	-1.2343
cons	-1.8824	0.1889	-9.9600	0.0000	-2.2527	-1.5121
Medium LoS						
Female	0.3601	0.0999	3.6100	0.0000	0.1644	0.5559
Previous admissions*	-0.0092	0.0034	-2.7100	0.0070	-0.0158	-0.0025
Age*	0.0105	0.0014	7.4800	0.0000	0.0077	0.0132
General surgery ward*	-0.3395	0.0748	-4.5400	0.0000	-0.4861	-0.1929
Outpatient clinic*	-0.5963	0.1215	-4.9100	0.0000	-0.8345	-0.3581
Number of s. procedures*	0.5603	0.0704	7.9600	0.0000	0.4223	0.6984
Diagnosis_category2*	1.2194	0.1837	6.6400	0.0000	0.8594	1.5795

Diagnosis_category3	1.0477	0.2399	4.3700	0.0000	0.5776	1.5178
Diagnosis_category4	2.3826	0.2468	9.6500	0.0000	1.8990	2.8663
Diagnosis_category5	1.3840	0.2133	6.4900	0.0000	0.9659	1.8021
Sp_category2	0.5753	0.2338	2.4600	0.0140	0.1171	1.0335
Sp_category3	-1.3978	0.2611	-5.3500	0.0000	-1.9096	-0.8860
Sp_category4*	-1.4474	0.1087	-13.3100	0.0000	-1.6605	-1.2343
cons	-3.95839	0.211355	-18.73	0	-4.37264	-3.54414

Table A.3: PPOM STATA output. * the indicates constrained variables

For the variables which were set free of constraints, the interpretation is similar to the GOLM:

The higher coefficient of diagnosis category 4 (HIV, respiratory tuberculosis, pleural effusion, etc.) in the first panel indicates that patient with a condition classified under that diagnosis category is more likely (29.6 times more) to have a medium or long LoS rather than a short one.

The higher negative value of the coefficient of surgical procedure category 3 in the second panel indicates that a patient undergoing one of those surgical procedures (e.g. Cholecystectomy, appendectomy, tonsillectomy, etc.) is more likely (4.04 times more) to have a short or medium LoS than a long one.

A.4. Multinomial Logit Model

Multinomial Logit model (MNLM) is the preferred option when the ordinal logistic model fails to meet the parallel assumption (Lunt, 2005). It can be thought as an extension of generalised linear models Section 6.1) where the categories of the dependent variable are no longer considered as ordered (Hilbe, 2009) and the effects of the independent variables are allowed to differ for each outcome. The multinomial Logit model can be expressed as:

$$\ln \frac{\Pr(y = m \mid \tilde{x})}{\Pr(y = b \mid \tilde{x})} = \tilde{x}\beta_{m|b} \text{ For } m=1 \text{ to } J \tag{A.15}$$

where *b* is the base category or the comparison group.

The predicted probabilities are calculated by:

$$\Pr(y = m \mid \tilde{x}) = \frac{\exp(\tilde{x}\beta_{m|b})}{\sum_{j=1}^{J} \exp(\tilde{x}\beta_{j|b})} \tag{A.16}$$

				Log-likelihood		-5675.28
				LR chi^2(26)		2667.04
				Prob > chi^2		0.0000
				Pseudo R^2		0.1903
	β	**Std. Err.**	***z***	***P>z***	**[95% Conf.**	**Interval]**
Medium LoS						
Female	0.0812	0.0548	1.4800	0.1380	-0.0262	0.1886
Previous admissions	-0.0082	0.0036	-2.2600	0.0240	-0.0152	-0.0011
Age	0.0081	0.0015	5.3600	0.0000	0.0051	0.0111
General surgery ward	-0.3598	0.0809	-4.4500	0.0000	-0.5183	-0.2013
Outpatient clinic	-0.5018	0.1266	-3.9600	0.0000	-0.7499	-0.2537
Number of s. procedures	0.4862	0.0778	6.2500	0.0000	0.3337	0.6388
Diagnosis_category2	1.1989	0.1990	6.0300	0.0000	0.8090	1.5889
Diagnosis_category3	1.7050	0.1997	8.5400	0.0000	1.3136	2.0965
Diagnosis_category4	3.3655	0.2453	13.7200	0.0000	2.8846	3.8463
Diagnosis_category5	2.5555	0.2023	12.6300	0.0000	2.1590	2.9519
Sp_category2	1.1744	0.2832	4.1500	0.0000	0.6193	1.7295
Sp_category3	-0.7110	0.1215	-5.8500	0.0000	-0.9491	-0.4729
Sp_category4	-1.3645	0.1178	-11.5900	0.0000	-1.5953	-1.1337
cons	-2.0565	0.2043	-10.0600	0.0000	-2.4570	-1.6561
Long LoS						
Female	0.4178	0.1048	3.9900	0.0000	0.2123	0.6233
Previous admissions	-0.0155	0.0086	-1.8000	0.0730	-0.0324	0.0014
Age	0.0191	0.0028	6.7600	0.0000	0.0136	0.0246
General surgery ward	-0.4703	0.1488	-3.1600	0.0020	-0.7620	-0.1786
Outpatient clinic	-1.1947	0.3697	-3.2300	0.0010	-1.9193	-0.4700
Number of s. procedures	0.8759	0.1320	6.6400	0.0000	0.6173	1.1346
Diagnosis_category2	1.1030	0.4308	2.5600	0.0100	0.2586	1.9473
Diagnosis_category3	1.0776	0.4376	2.4600	0.0140	0.2199	1.9352

Diagnosis_category4	3.4445	0.4596	7.4900	0.0000	2.5437	4.3453
Diagnosis_category5	1.8347	0.4332	4.2400	0.0000	0.9856	2.6838
Sp_category2	1.4214	0.3513	4.0500	0.0000	0.7330	2.1099
Sp_category3	-1.7150	0.2920	-5.8700	0.0000	-2.2873	-1.1428
Sp_category4*	-1.8603	0.2324	-8.0000	0.0000	-2.3158	-1.4048
cons	-3.5007	0.4386	-7.9800	0.0000	-4.3602	-2.6411

Table A.4:MNLM STATA output. Short LoS is the base category

The MNLM was fitted on STATA using the `mlogit` function. Since the MNLM can be thought as *J-1* simultaneous binary logistic regressions, the interpretation of the parameters should consider all possible scenarios among categories. For the variable LoS category with three outcomes, there are three possible scenarios to consider: medium vs. short, long vs. short and medium vs. long. Table A.4 shows the STATA output of the multinomial model where short LoS is the base category; usually STATA picks the category with the highest frequency to play the role of the baseline category. From the first panel medium vs. short, the results can be interpreted as follows in terms of odds ratios:

Patients are a lot more likely (29 times more) to have a medium LoS than a short LoS if they have a disease classified in the diagnosis category 4 e.g. HIV, respiratory tuberculosis, pleural effusion, etc.

A patient who enters to hospital via the outpatient clinic is 1.43 times more likely to have a short LoS than a medium.

From the second panel long vs. short, the results can be interpreted:

A patient is more likely (3.01 times more) to have a long LoS than a short LoS if he/she has a disease classified in the diagnosis category 2 e.g. chronic renal failure or ventral hernia.

A patient undergoing one of the surgical procedures of category 3, e.g. Cholecystectomy, appendectomy or tonsillectomy, is more likely (5.55 times more) to have a short LoS than a long one.

On the other hand, the MNLM relies on the assumption of independence of irrelevant alternatives (IIA), where the odds do not depend on other alternatives that are available. In other words, adding or deleting outcome categories does not affect the odds among other outcomes.

The most common tests of IIA are the Hausman-McFaden test and the Small-Hsiao test. Testing the IIA involve comparing the estimated coefficients from the full model to those from a

restricted model that excludes at least one of the alternatives. If the test statistic is significant, the assumption of IIA is rejected indicating that the MNLM is inappropriate.

Table A.5andTable A.6 show the STATA output of both tests. In Table A.5 the results of the Hausman-McFaden test differ considerably depending on the category considered. If the category 1 (medium LoS) is omitted there is evidence that the IIA is violated; however the other two test statistics are negative, which might indicate evidence that IIA has not been violated (Hausman and McFadden, 1984). In each individual test of the Small-Hsiao test indicates that the assumption has not been violated. Long and Freese (2006) comment that both tests often give inconsistent results and provide little guidance to violations of the IIA assumption.

Ho: Odds (Outcome J vs. Outcome K) are independent of other alternatives				
Omitted	**chi2**	**df**	**P>chi2**	**evidence**
Medium	58.144	14	0	Against Ho
Long	-6.125	14	---	---
Short	-9.388	14	---	---

Table A.5:Hausman test of IIA

Ho: Odds (Outcome J vs. Outcome K) are independent of other alternatives						
Omitted	**lnL(full)**	**lnL(omit)**	**chi2**	**df**	**P>chi2**	**evidence**
Medium	-624.131	-617.956	12.35	14	0.574	For Ho
Long	-2025.922	-2022.501	6.841	14	0.941	

Table A.6: Small and Hsiao test of IIA

Nevertheless the categories of LoS can plausibly be assumed to be distinct and weighted independently, e.g. the patient odds of having a short LoS does not change if the other two categories are omitted. Consequently if all the evidence is set together, it is fair to assume that the MNLM model for MRC dataset does not violate the IIA assumption.

A.5. Stereotype Ordered Regression

The stereotype ordered regression model (SORM) can be thought of as imposing ordering constraints on a multinomial model (Lunt, 2005). It was proposed by (Anderson, 1984) in response to the restrictive parallel assumption in the ordered regression model.

The name stands for on how the model was introduced by Anderson. He described the outcome categories as "assessed". In addition, each respondent is considered to have "stereotypes" that characterise the outcomes categories. The respondent assesses each category and then picks the one whose "stereotype" most closely matches the respondent's view. Albeit how SORM was originally defined, nowadays it is used in many other applications aside from assessed choices.

The SORM is defined as:

$$\ln \frac{\Pr(y = q \mid \tilde{x})}{\Pr(y = r \mid \tilde{x})} = (\theta_q - \theta_r)\beta_0 - (\phi_q - \phi_r)(\tilde{x}\beta) \tag{A.17}$$

Where β_0 is the intercept, $\theta's$ and $\phi's$ are scale factors associated with the outcome categories. The model allows the coefficients associated with each independent variable to differ by a scale factor that depends on the pair of outcomes on the left hand side of the equation. Similarly, the $\theta's$ allow different intercepts for each pair of outcomes. If the relationship between the independent variables and dependent variable is ordinal then $\phi_1 > \phi_2 > ... > \phi_{J-1} > \phi_J$

Constraints need to be added to the model to make it identified[64], Long (1997) suggests $\phi_1 = 1$, $\phi_J = 0$ $\theta_1 = 1$ and $\theta_J = 0$ .The predicted probabilities are calculated by:

$$\Pr(y = m \mid \tilde{x}) = \frac{\exp(\theta_m \beta_0 - \phi_m \tilde{x}\beta)}{\sum_{j=1}^{J} \exp(\theta_j \beta_0 - \phi_j \tilde{x}\beta)} \tag{A.18}$$

The SORM was fitted on STATA using the `slogit` command. The next figure shows the output:

[64] The model is identifiable if it is theoretically possible to learn the true value of the model's underlying parameter after obtaining an infinite number of observations from it.

				Log-likelihood	-5714.74	
				Wald chi^2(13)	604.18	
				Prob > chi^2	0.0000	
	β	**Std. Err.**	***z***	***P>z***	**[95% Conf.**	**Interval]**
Female	0.1628	0.0615	2.6500	0.0080	0.0423	0.2834
Previous admissions	-0.0109	0.0040	-2.7500	0.0060	-0.0187	-0.0031
Age	0.0116	0.0018	6.6100	0.0000	0.0082	0.0150
General surgery ward	-0.4328	0.0909	-4.7600	0.0000	-0.6110	-0.2546
Outpatient clinic	-0.6848	0.1458	-4.7000	0.0000	-0.9704	-0.3991
Number of s. procedures	0.6474	0.0899	7.2000	0.0000	0.4713	0.8236
Diagnosis_category2	1.3846	0.2202	6.2900	0.0000	0.9529	1.8162
Diagnosis_category3	1.8545	0.2239	8.2800	0.0000	1.4157	2.2932
Diagnosis_category4	3.9601	0.3020	13.1100	0.0000	3.3681	4.5521
Diagnosis_category5	2.7912	0.2348	11.8900	0.0000	2.3310	3.2515
Sp_category2	1.4796	0.3212	4.6100	0.0000	0.8501	2.1091
Sp_category3	-0.9800	0.1421	-6.9000	0.0000	-1.2585	-0.7016
Sp_category4	-1.6893	0.1437	-11.7500	0.0000	-1.9710	-1.4076
cons	0.1628	0.0615	2.6500	0.0080	0.0423	0.2834
phi1_1	1					
phi1_2	0.1552	0.0351	4.41	0	0.0862	0.2241
phi1_3	0	(base category outcome)				
theta1	4.0633	0.2397	16.95	0	3.5938	4.533
theta2	2.1155	0.115581	18.3	0	1.8889	2.3420
theta3	0	(base category outcome)				
phi1_1	1					
phi1_2	0.1552	0.0351	4.41	0	0.0862	0.2241
phi1_3	0	(base category outcome)				

Table A.7:SORM STATA output

The parameters of the model can be interpreted in terms of the odds of the base category (STATA selected the last category as the base category) versus the first category:

The odds of having a long versus a short LoS are 4 times larger for patients with a disease classified in the diagnosis category 2, like chronic renal failure or ventral hernia.

The odds of having a short versus a long LoS are 1.54 times larger for patients who enter to the hospital via outpatient clinic.

The odds of having a long versus a short LoS are 52.46 times larger for patients with a disease classified in the diagnosis category 4 like HIV, respiratory tuberculosis, pleural effusion, etc.

The odds of having a short versus a long LoS 2.66 times larger for patients undergoing one of the surgical procedures of category 3 like Cholecystectomy, appendectomy and tonsillectomy.

It is important to clarify that the `slogit` command uses an order of the categories that does not necessarily correspond to how the dependent variables has been numbered. Fortunately looking at $\phi's$ on Table A.7: $\phi_1 > \phi_2 > \phi_3$. This means that SORM ordered LoS category as short, medium and long, which is what it was expected. However, Long (2006) presented a case when the original order of the categories was not respected and it was required to examine the effects of the independent variables on all pairs of outcomes. This makes the interpretation of parameters slightly more complicated.

A.6. Validation and Performance

Notice that the validation the previous models followed the same strategy than in previous chapters, where the estimation was performed using 2/3 of the MRC dataset for the training set and the remaining was used for testing.

It is important to notice that although the interpretation of the odds ratios varied from one model to another, the direction of the effects was similar in all models. For a more formal evaluation of the models, the log-likelihood, AIC and BIC of each model were calculated to give an idea about the adequacy of the fit. Accuracy rates for each category, overall performance and prediction profit were calculated using the testing set in order to evaluate the capacity of the algorithms to predict on new data.

The left hand side of the Table A.8 shows the results of the log-likelihood, AIC and BIC for the training set. The model which represents the best fit according to the log-likelihood and AIC is the MNLM. However, in terms of BIC, the best model is the PPOM. The lower value of BIC

could indicate a better fit or the presence of fewer parameters; because BIC penalizes free parameters more strongly than AIC.

Model	Log-likelihood	AIC	BIC	Accuracy rate			
				Short	Medium	Long	Overall
GOLM	-5678.14	1.295	-68446.5	85.99%	53.73%	0.4%	71.65%
MNLM	-5675.28	1.294	-68452.2	86.10%	53.73%	0%	71.69%
SORM	-5714.74	1.301	-68473.2	86.24%	53.27%	0%	71.65%
PPOM	-5681.35	1.296	-68440.1	85.89%	54.11%	0%	71.67%

Table A.8: Comparative chart Logistic regression models

The second part of the table shows the accuracy rates on the testing set, giving an idea of how well the models are in predicting new patients. The four models perform exceptionally well in predicting patients with short LoS. However, they failed drastically to predict patients with a long LoS. Although patients with long LoS account for just 5% of the total patients, it is a concern that around 40% of these patients are not just incorrectly classified, but they are being classified in the category "short". In other words, 40% of the patients who ended having a long LoS of more than 12 days were classified initially as patients who would stay no more than 3 days at hospital.

In terms of overall accuracy rate, the four models present minimal differences, being the MNLM slightly superior to other three, and SORM inferior. However, these differences might be insignificant in practical and statistical terms. Consequently the process of splitting randomly the dataset in 2/3 for creating the models and 1/3 for testing purposes was repeated 10 times in order to check if the models follow the same behaviour. Table A.9shows the results of the 10 trials where the overall accuracy rate presents a stable and reliable behaviour. Table A.10shows the 10 trails average accuracy rate of each category, whose results are very similar to the original values.

Run	GOLM	MNLM	SORM	PPOM
1	71.65%	71.69%	71.65%	71.67%
2	71.24%	71.29%	71.15%	71.36%
3	72.29%	72.31%	72.29%	72.29%
4	71.79%	71.79%	71.63%	71.79%
5	72.57%	72.57%	72.48%	72.41%
6	72.30%	72.30%	72.23%	72.23%
7	71.01%	71.03%	71.03%	71.07%
8	72.82%	72.82%	72.91%	72.96%
9	73.40%	73.50%	73.36%	73.45%
10	72.56%	72.61%	72.50%	72.63%
Average	72.16%	72.19%	72.12%	72.19%

Table A.9:Overall accuracy rates in 10 trials for logistic regression models

	GOLM	MNLM	SORM	PPOM
Short	86.07%	86.08%	86.16%	85.93%
Medium	54.78%	54.88%	54.51%	55.14%
Long	0.66%	0.16%	0.00%	0.46%

Table A.10:Average accuracy rates per category

The values of the average overall accuracy for the four models remain very close to each other. To determine whether those small differences are statistically significant, ANOVA was performed. The results confirmed that there is NOT a statistical significant difference across the performance measurement of the four models, $F(3,36)=.018$, $p=.997$.

For the objectives of this chapter, the interested reader may ask, "among the logistic regression models presented on this chapter, which is the best option for predicting LoS category?"

In general their accuracy rates to predict patients with short and medium LoS (95% of the patients) were quite good; although the four models failed to predict patients with long LoS. Furthermore in the last section, it was demonstrated on a 10 trials experiment and supported by

ANOVA analysis that the four models have on average equal performance. Therefore other factors described below should be taken into account when selecting the most appropriate model for predicting patient LoS category.

In spite having less parameters, which simplifies the interpretation, the SORM model has the highest values for the log-likelihood, AIC and BIC. This indicates it is a less adequate fit of the MRC data compared with the other models, except from the case mentioned on A.5

MNLM is the most popular model and there is a wide selection of software in the market available for its implementation, which is naturally an advantage. However the biggest drawback of the MNLM is that it ignores the ordinality of the data; some authors criticise treating the independent variable as nominal firmly, when there is clearly an ordinal nature, because it hinders the ability to assess directionality and progression (Cliff, 1996 and O'Connell, 2006).

On the other hand, GOLM preserves ordinality at the cost of the same number of parameters than MNLM and a more complex interpretation of the parameters.

There are three key advantages of the PPOM: It preserves the ordinality of the outcome variable, it reduces the number of parameters in relation with GOLM and MNLM and it has the capacity of identifying the variables which change across the different outcome categories (when testing the parallel assumption). The latest gives a better understanding of the underlying nature of the LoS.

Finally, in terms of interpretation of the results, SORM and MNLM are the easiest models to analyse. MNLM compares the likelihood of being at the base category with the likelihood of being at one of the other categories and SORM compares the likelihood of being at the lowest numbered category with the likelihood of being at the highest one. On the other hand, GOLM compares the likelihood of being at or below a given category with the likelihood of being at any of the remaining categories. In the case of the PPOM, the interpretation remains the same as in GOLM for the unconstrained variables; conversely, for the constrained variables, the PPOM reveals the likelihood of moving in certain direction, either the patient move towards longer LoS or shorter LoS. Notice that it might be a cause for confusion for non-experts the fact that odds (for PPOM) between unconstrained variables and constrained variables are interpreted differently. Nevertheless, GOLM and PPOM are very useful in identifying trends in the odds; and in practical terms, understanding patient progression regarding to their LoS category, either upward or downward, could bring more valuable information than comparing against a fixed category.

Considering all the previous points, PPOM appears to be the most adequate option from the family of logistic regression models to predict the patient LoS category. PPOM has parsimony; it gives a better insight on the patterns of the independent variables; and it respects the ordinal nature of dependent variable, which enhances the understanding of the progression of the patients in terms of LoS.

A.7. Discussion

From this initial analysis, some important lessons were learnt which shaped the direction of this research.

Based on the fact that all the methods explored in this section failed drastically to predict long LoS category, the researcher came with two solutions to overcome the problem:

1. Exploring other classification methods: It was decided to consider a wider selection of data mining techniques, from standard methods such as classification trees to more advanced models like logistic regression trees, with the aim of finding a method or model which would increase the accuracy rate for the category long LoS. These models were fully explored in Chapter 6.
2. Redesigning LoS category: In this initial stage, LoS category was defined for the planning team at MRC hospital based on personal judgment and empirical observation only. However it was believed that by redesigning carefully the characteristics of each category to make them statistically and clinically meaningful, it would increase the chance of any model to be more accurate when classifying patients into those categories. This is how originally model-based cluster analysis using finite mixture models came to be part of this research.

However, two things happened while this chapter was being written. Firstly, a literature review on the field and a better understanding and broader perspective of the research problem, derived from further informal interviews with both hospitals, made clear that the original objectives, defined at the beginning of this chapter, were insufficient and limited. Therefore a broader general objective was defined as the developing a statistical model to predict patient length of stay (LoS) in public hospitals in Mexico which approximates to the distribution of the LoS, recognises and addresses the heterogeneity population problem; supplies LoS predictions for individual patients and cohort of patients (within and between hospitals) and demonstrates a solid application into the decision-making process.

Secondly, the logistic regression models developed on this section were originally part of the collection of models and algorithms analysed in Chapter 6 for the group-based approach. However, when ISSEMyM data becomes available and different mixtures of distributions were explored for the model-based clustering analysis (for both hospitals) it became clear that just two categories of LoS were enough to describe the data (see Section 4.1.2). This had a big impact on this research because all the models explored in this section became inappropriate, considering that they are defined for polytomous outcome variables (i.e. variables with more than 2 categories). Instead, other model from the family logistic regression models was found more suitable for the binary nature of the newly defined LoS category variable, the Logit model (see Section 6.1).

A.8. Summary

In this chapter, ordinal regression models (ORM) and specifically the continuation ratio model (CRM) were presented as the first option to predict the patient LoS category since the nature of the dependent variables is clearly ordinal. However, it was demonstrate, for the MRC dataset, that those models resulted to be unsuitable since the parallel regression assumption was violated. Therefore it is not possible to draw conclusions about the population. Other models from the family of logistic regression models were presented as viable alternatives, including: generalised ordered logit model (GOLM), partial proportional odds model (PPOM), multinomial Logit model (MNLM) and stereotype regression model (SORM).

Those models were fitted on STATA, using 2/3 of the MRC dataset, whereas the remaining was used for testing purposes. Initial performance measurements indicated that the four models had similar performance on predicting new data. Their accuracy rates to predict patients with short and medium LoS were quite good, although they failed to predict patients with long LoS category. Further repeated experiments confirmed that indeed the four models have in average the same accuracy to predict patient LoS category.

The latest evidence suggested that it is important to take into account other factors when choosing the most appropriate model. The number and interpretability of the model parameters, proper management of the outcome ordinal nature and software availability were some of the factors that influence the selection of the partial proportional odds model as the most suitable model from the family of the logistic regression models to predict LoS category.

Finally, important lessons were learnt from this research journey: the PPOM model meet the basic requirements of the MRC medical staff answering the two concerns described at the beginning of this chapter (based on the existing patient LoS classification scheme). However, in

the future it may be desirable to predict patient LoS in more accurately form rather than just a general classification into three categories. This latter idea was the foundation for the group-based approach developed earlier in this thesis, where each category of patients (short, medium, long LoS) has an associated LoS probabilistic curve. Moreover, this motivates as well for exploring a different patient LoS classification scheme, one where each LoS category is statistically (and clinically) meaningful, which was the aim of model-based cluster analysis described in Chapter 4.

Therefore the logistic regression models developed on this section were planned to be part of the collection of models and algorithms analysed in Chapter 6 for the group-based approach. However the results in Section4.1.2 indicated that the definition of the variable LoS category containing just two groups was more appropriate to describe the LoS, leading the simple binary logistic regression to be the only option, among the logistic regressions family, to predict LoS category.

Appendix B

Clustering diagnoses and surgical procedures

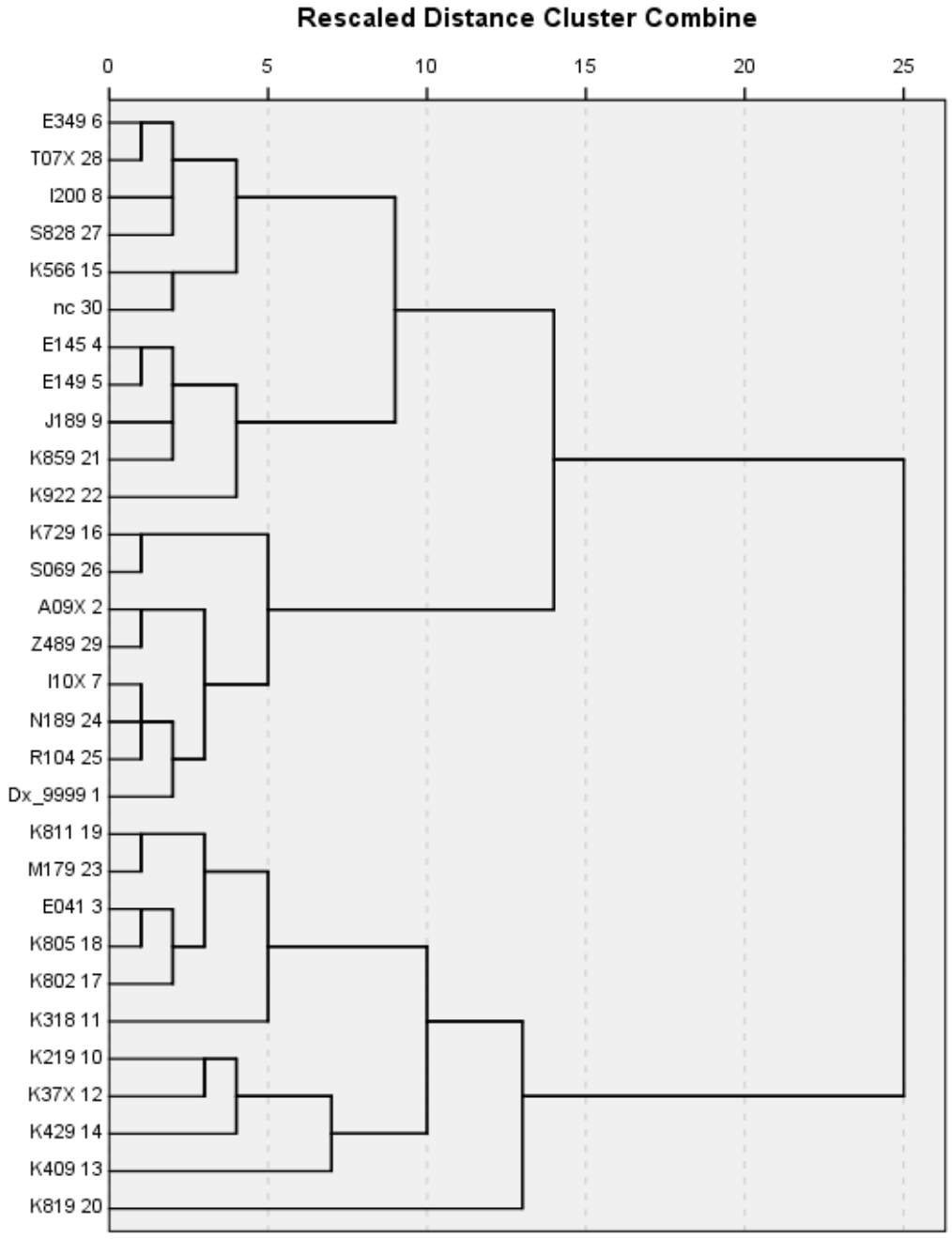

Figure B.1Dendrogram using Average linkage (within groups) for the variable “first diagnosis”

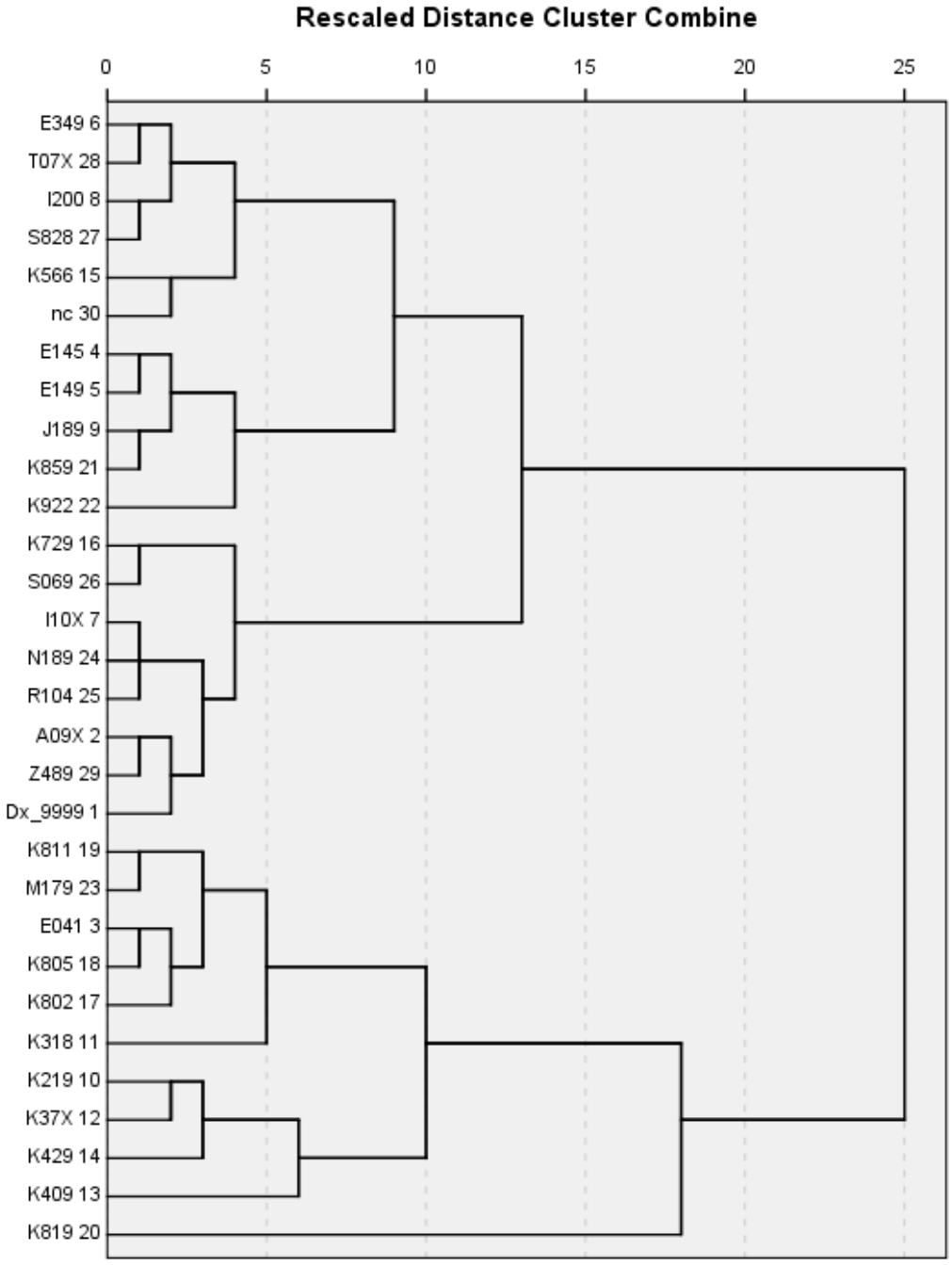

Figure B.2: Dendrogram using Average linkage (between groups) for the variable "first diagnosis"

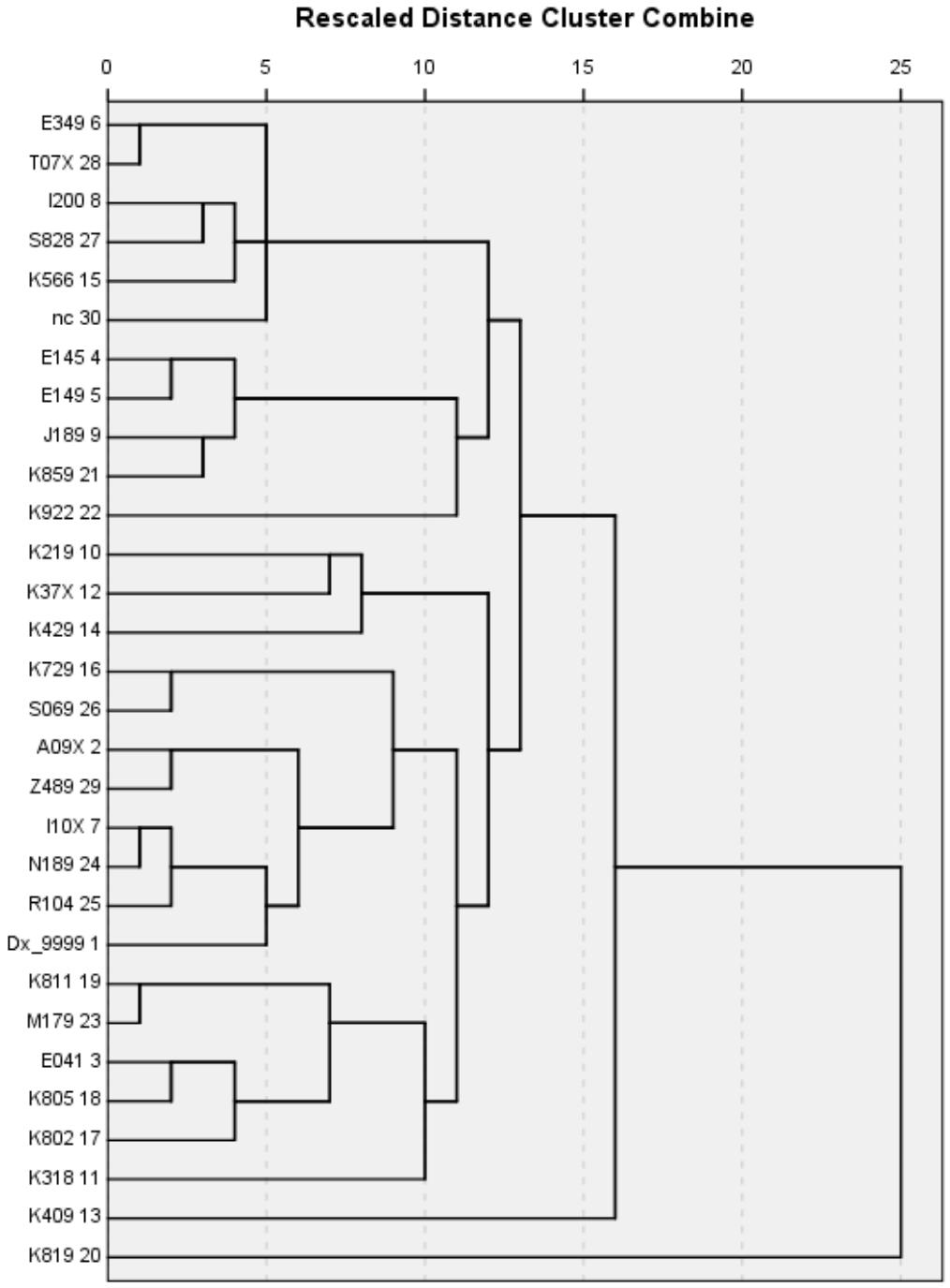

Figure B.3:Dendrogram using Single linkage for the variable “first diagnosis”

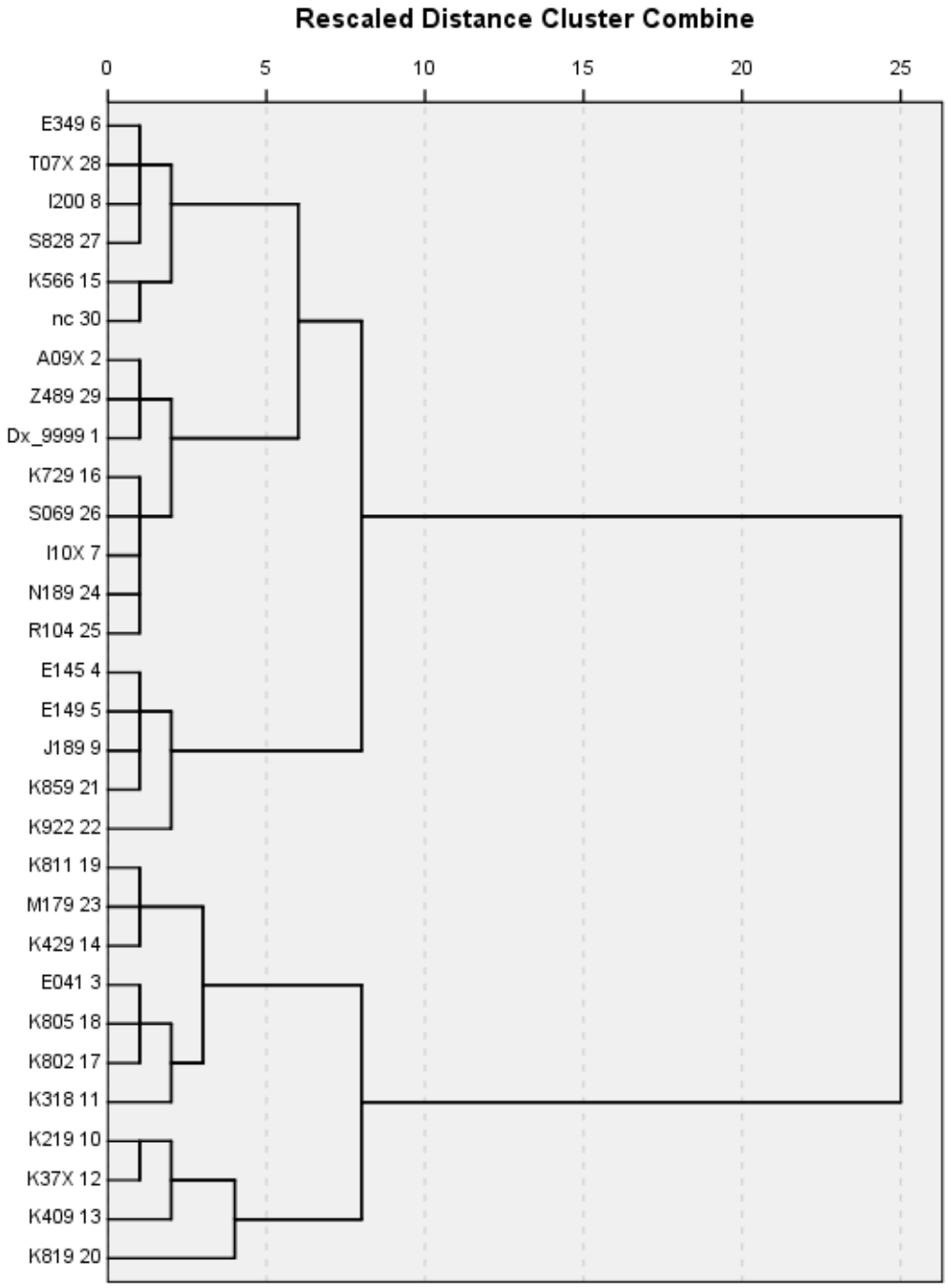

Figure B.4:Dendrogram using Complete linkage for the variable “first diagnosis”

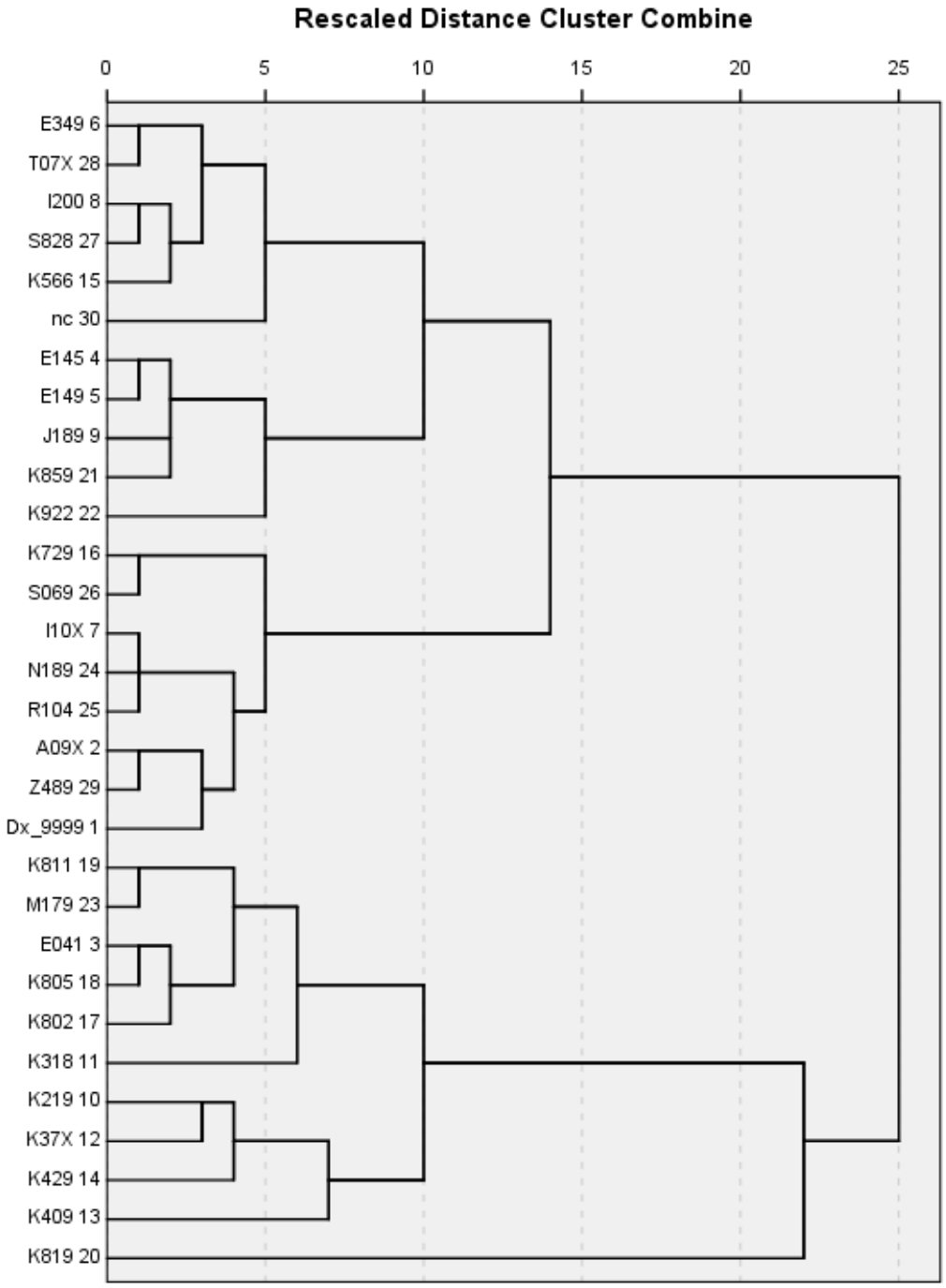

Figure B.5:Dendrogram using Centroid linkage for the variable “first diagnosis”

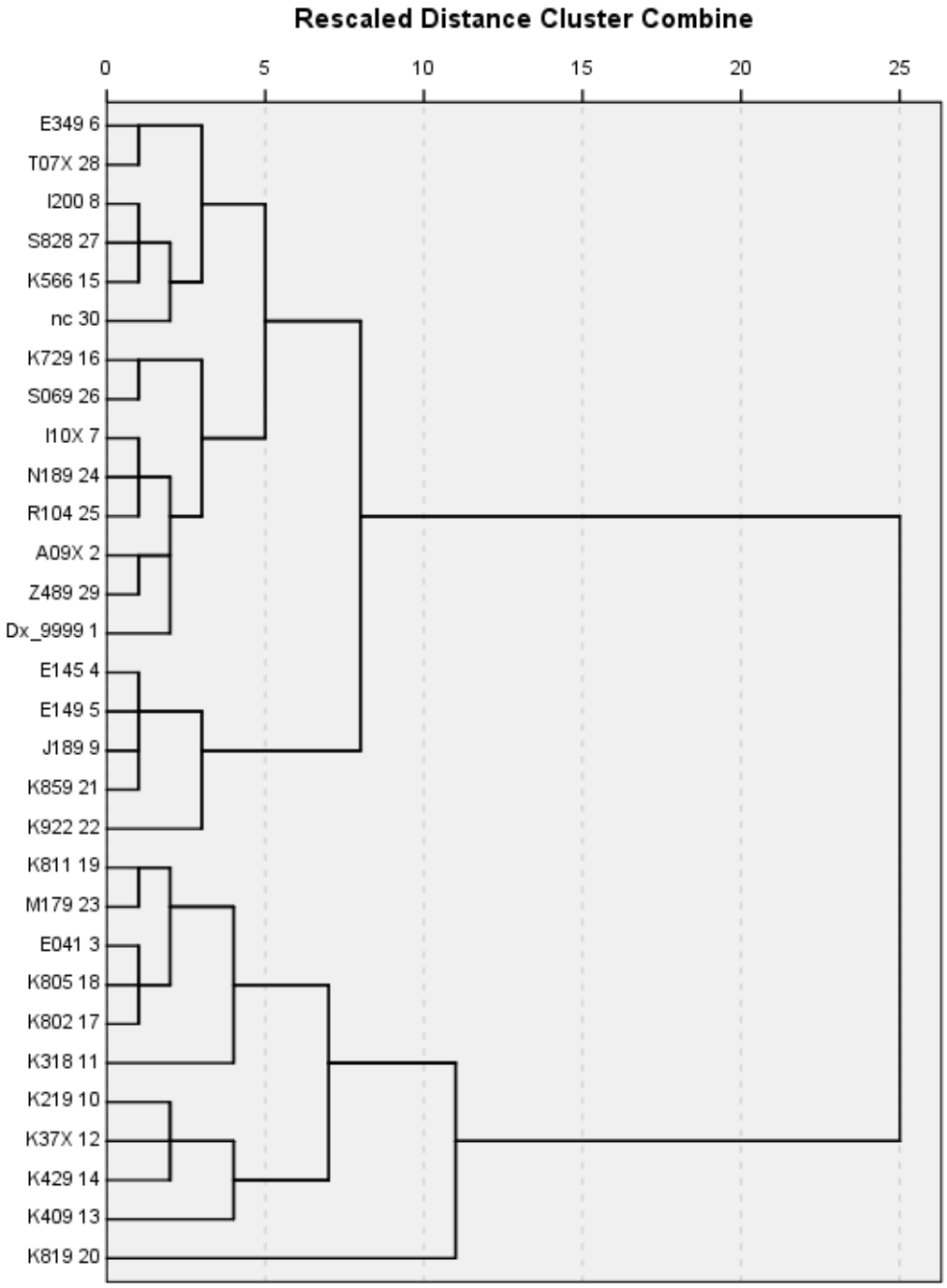

Figure B.6:Dendrogram using Median linkage for the variable “first diagnosis”

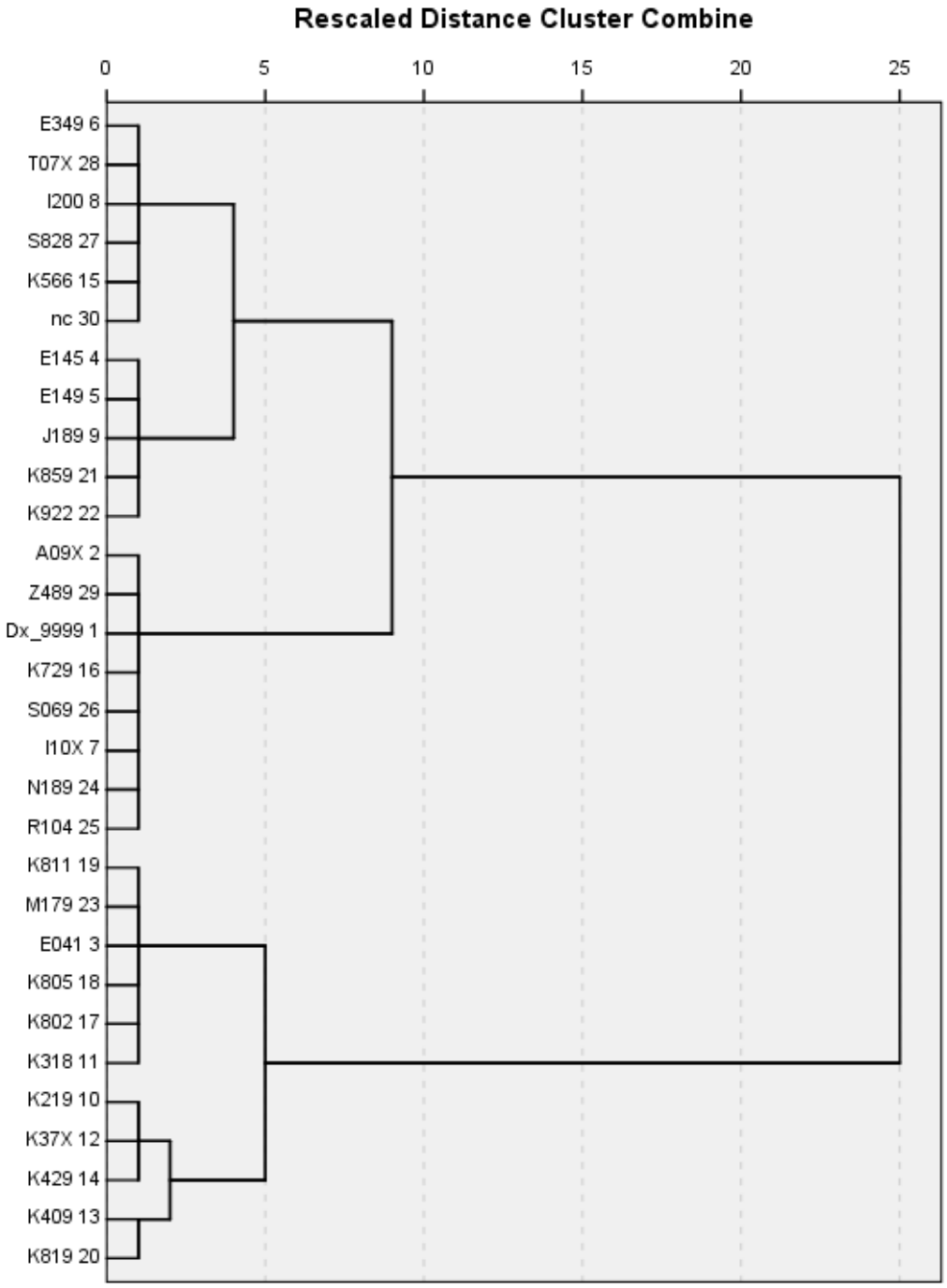

Figure A.1:Dendrogram using Ward linkage for the variable “first diagnosis”

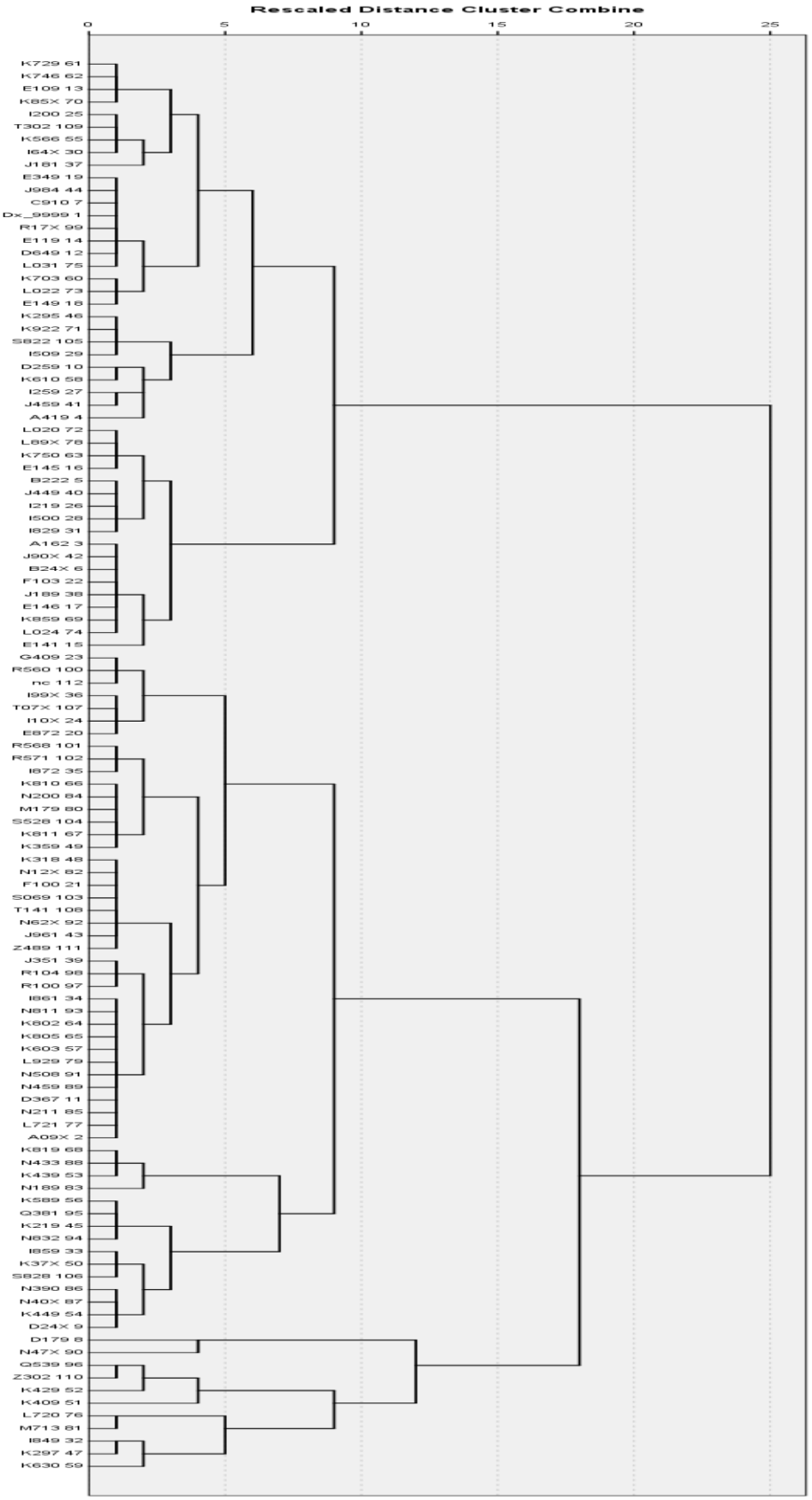

Figure B.7:Dendrogram using Average linkage (within groups) for the variable “diagnosis”

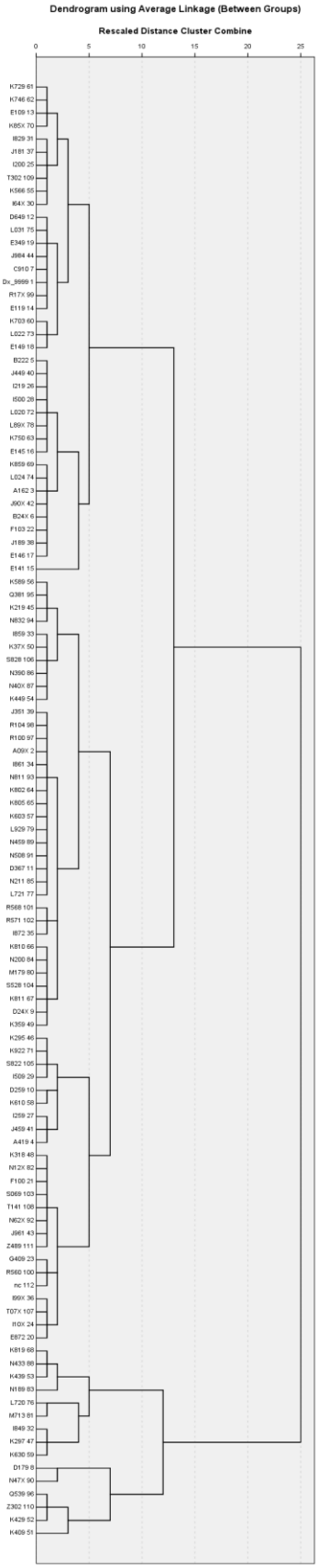

Figure B. 8: Dendrogram using Average linkage (between groups) for the variable “diagnosis”

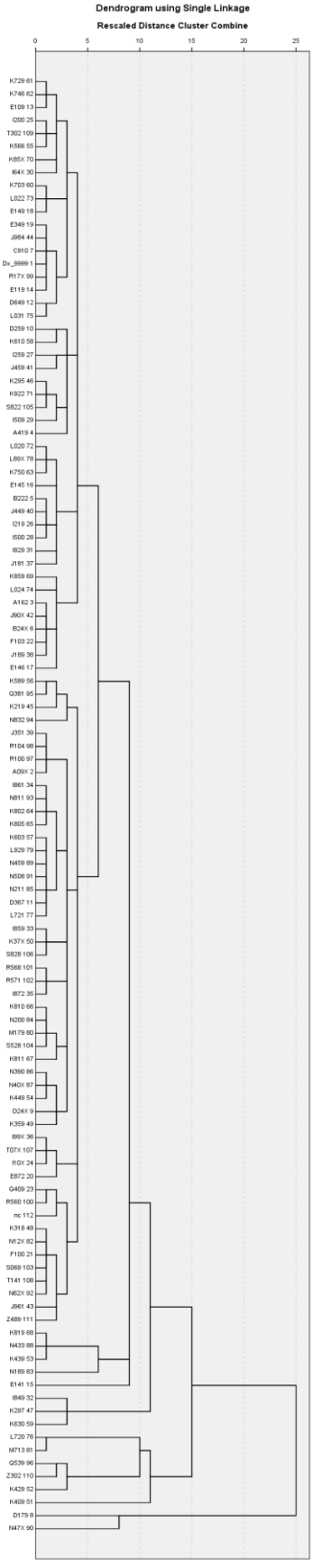

Figure B.9:Dendrogram using Single linkage for the variable ““diagnosis””

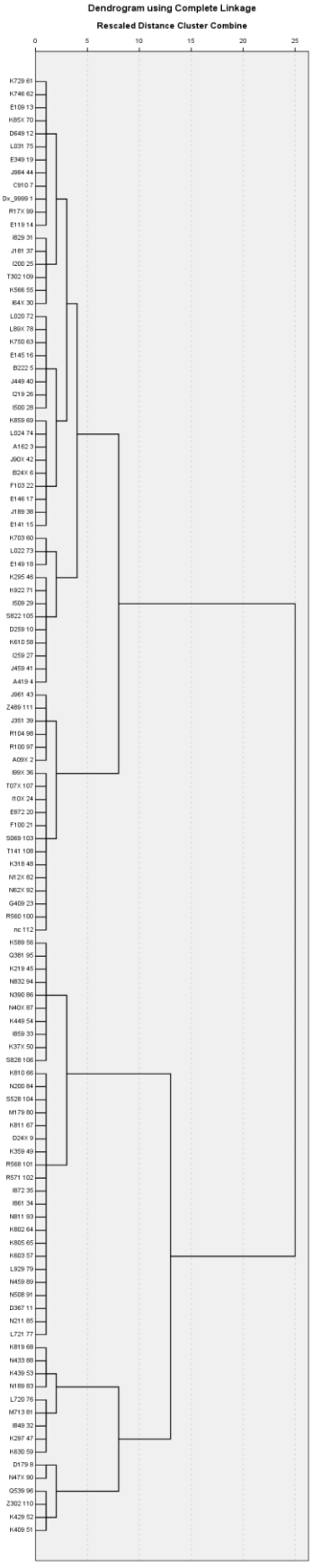

Figure B.10:Dendrogram using Complete linkage for the variable ““diagnosis””

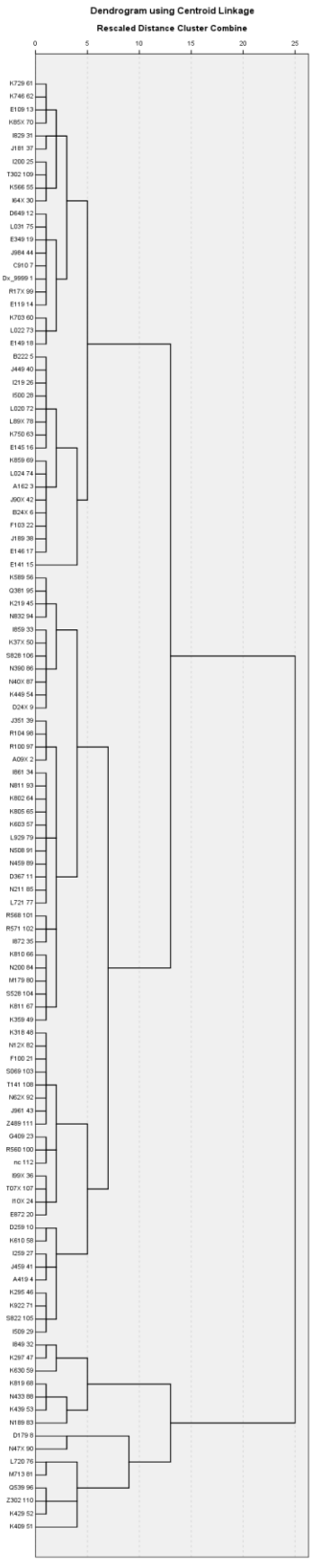

Figure B.11:Dendrogram using Centroid linkage for the variable ““diagnosis””

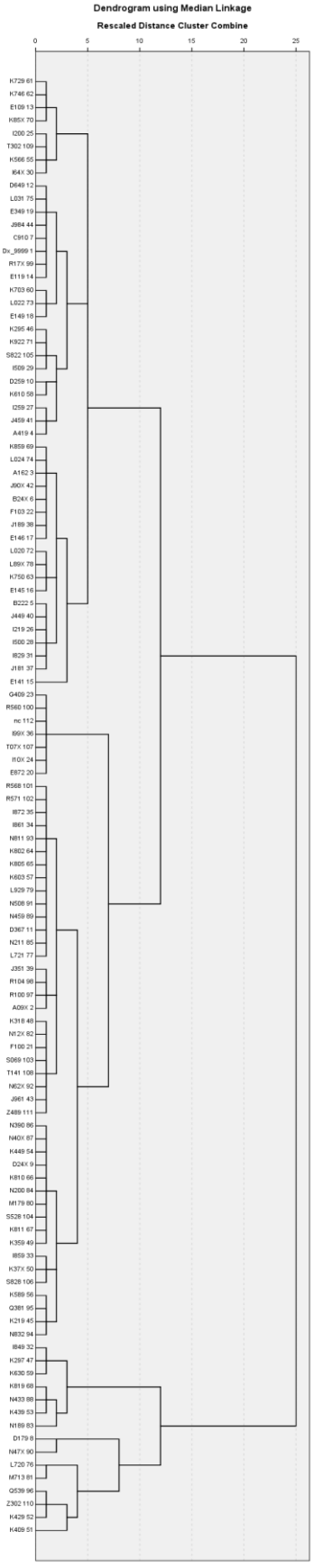

Figure B.12: Dendrogram using Median linkage for the variable ““diagnosis””

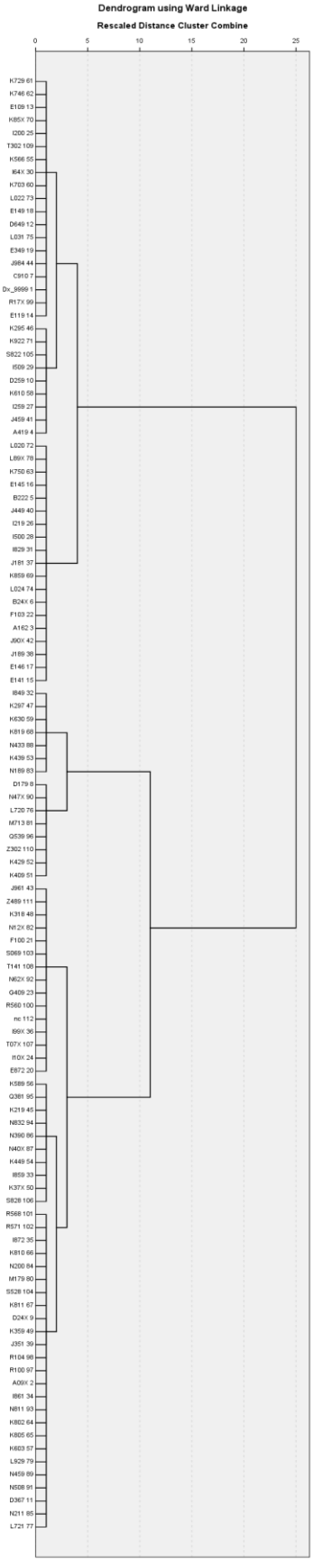

Figure B.13:Dendrogram using Ward linkage for the variable ““diagnosis””

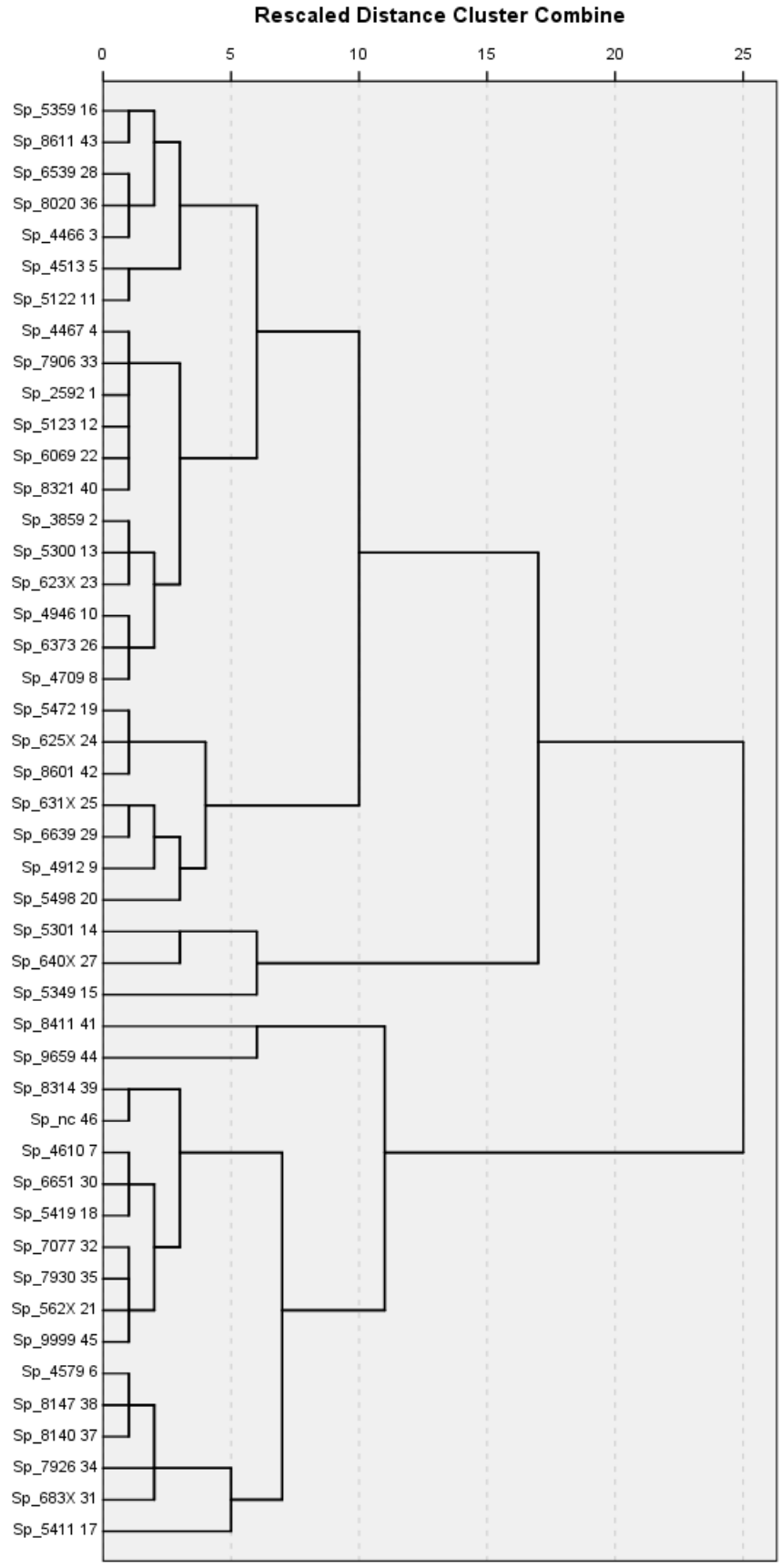

Figure B.14:Dendrogram using Average linkage (within groups) for the variable “surgical procedure”

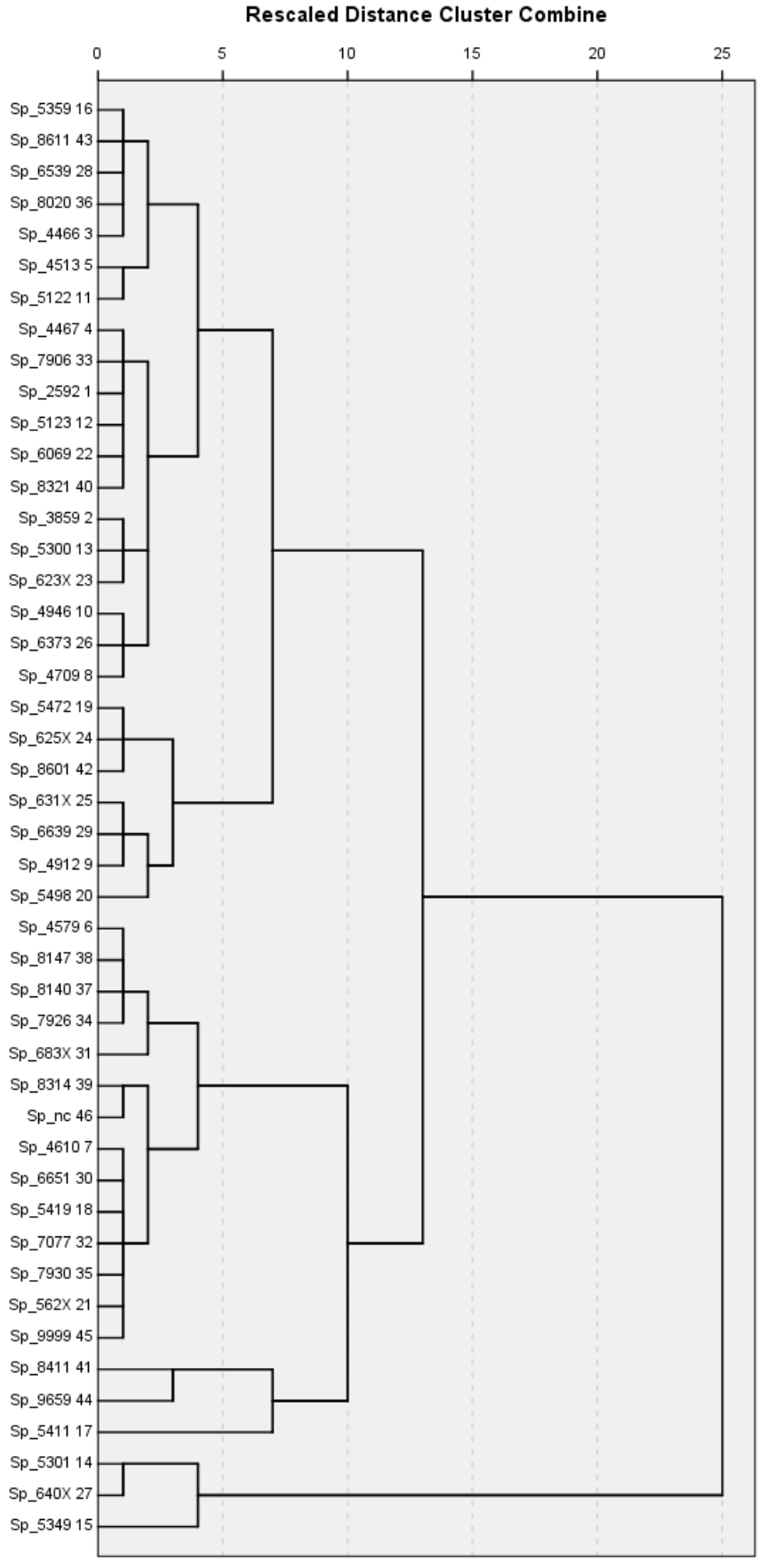

Figure B.15:Dendrogram using Average linkage (between groups) for the variable “surgical procedure”

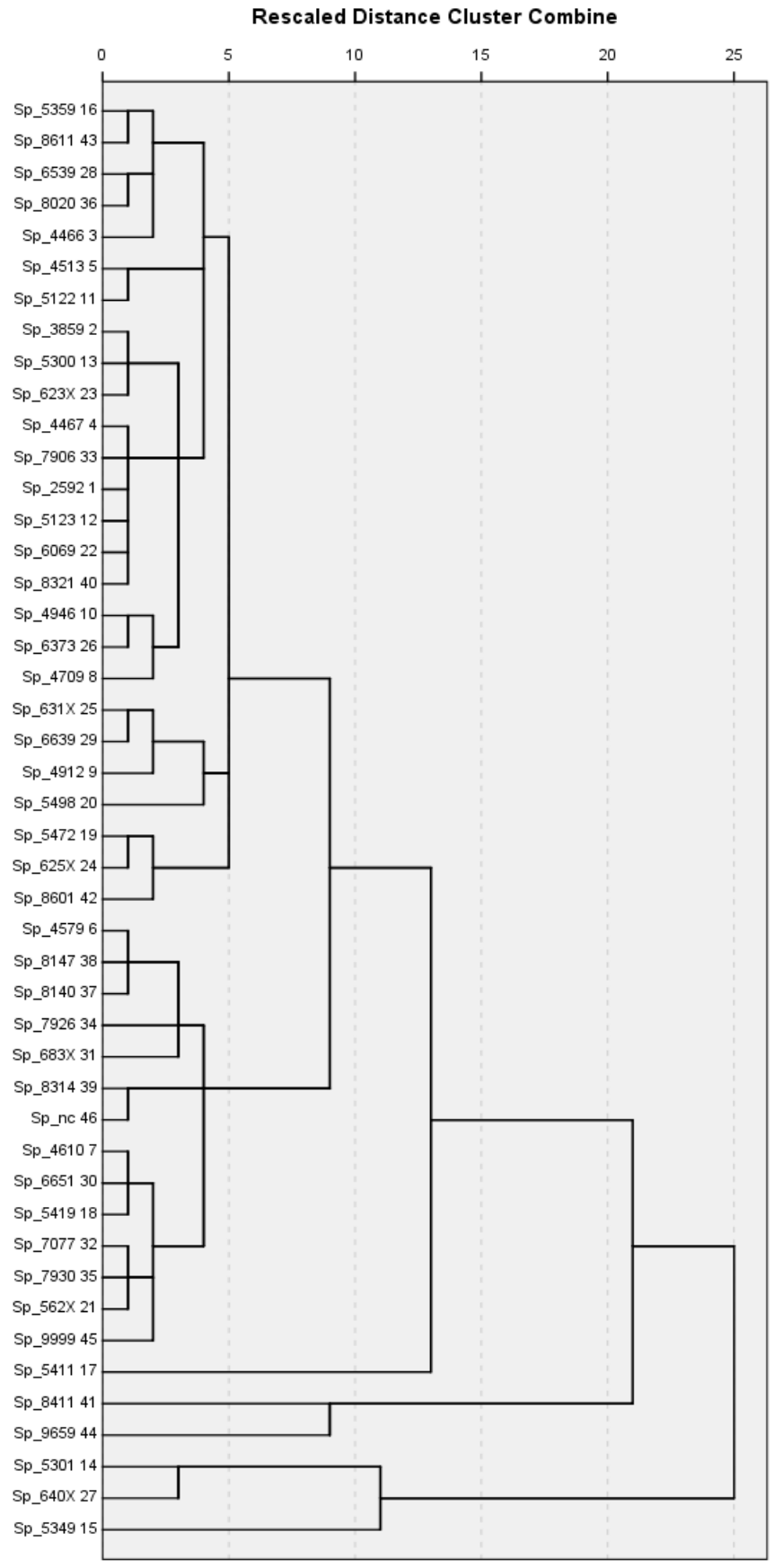

Figure B.16:Dendrogram using Single linkage for the variable “surgical procedure”

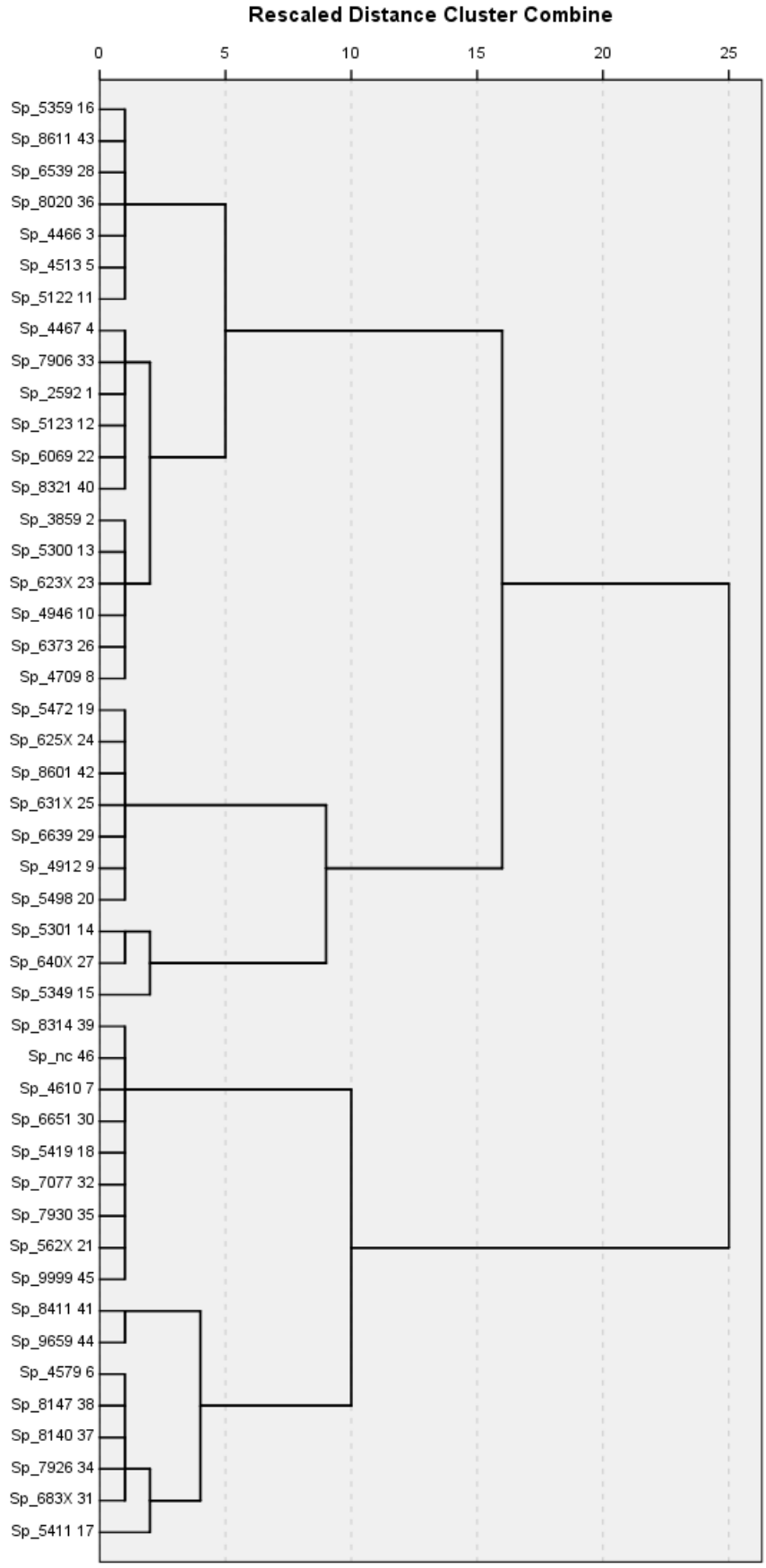

Figure B.17:Dendrogram using Complete linkage for the variable “surgical procedure”

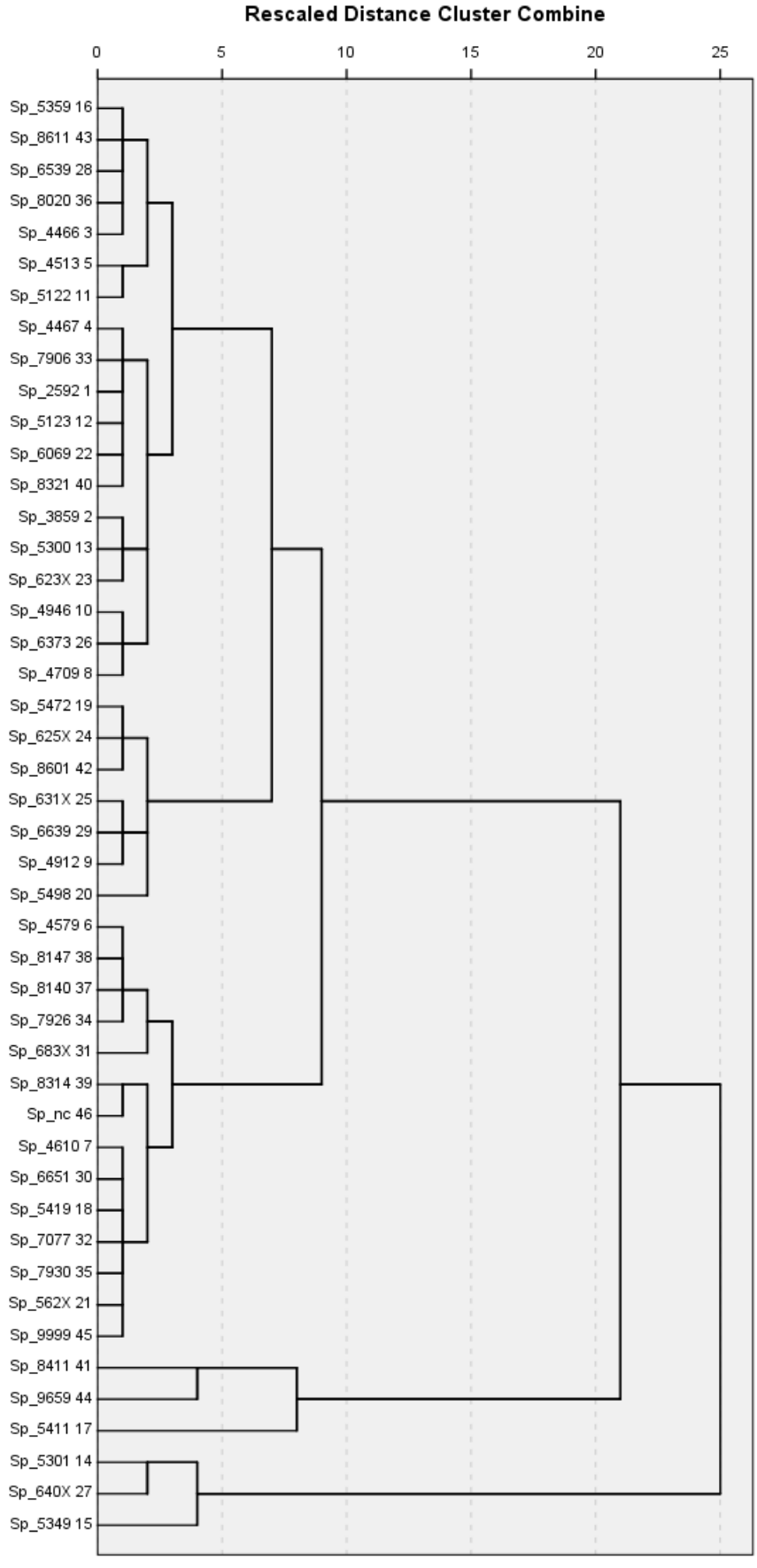

Figure B.18:Dendrogram using Centroid linkage for the variable “surgical procedure”

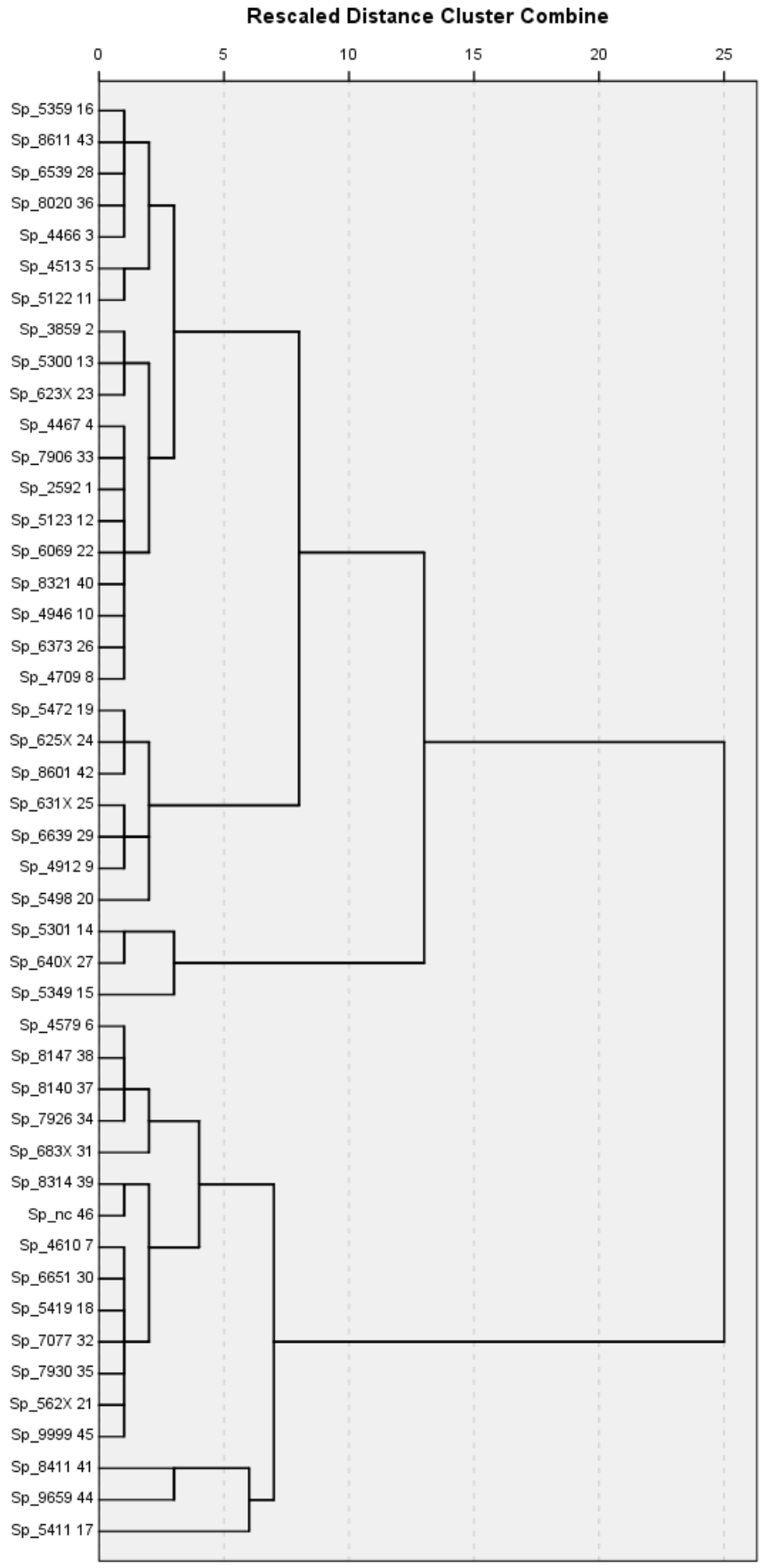

Figure B.19:Dendrogram using Median linkage for the variable “surgical procedure”

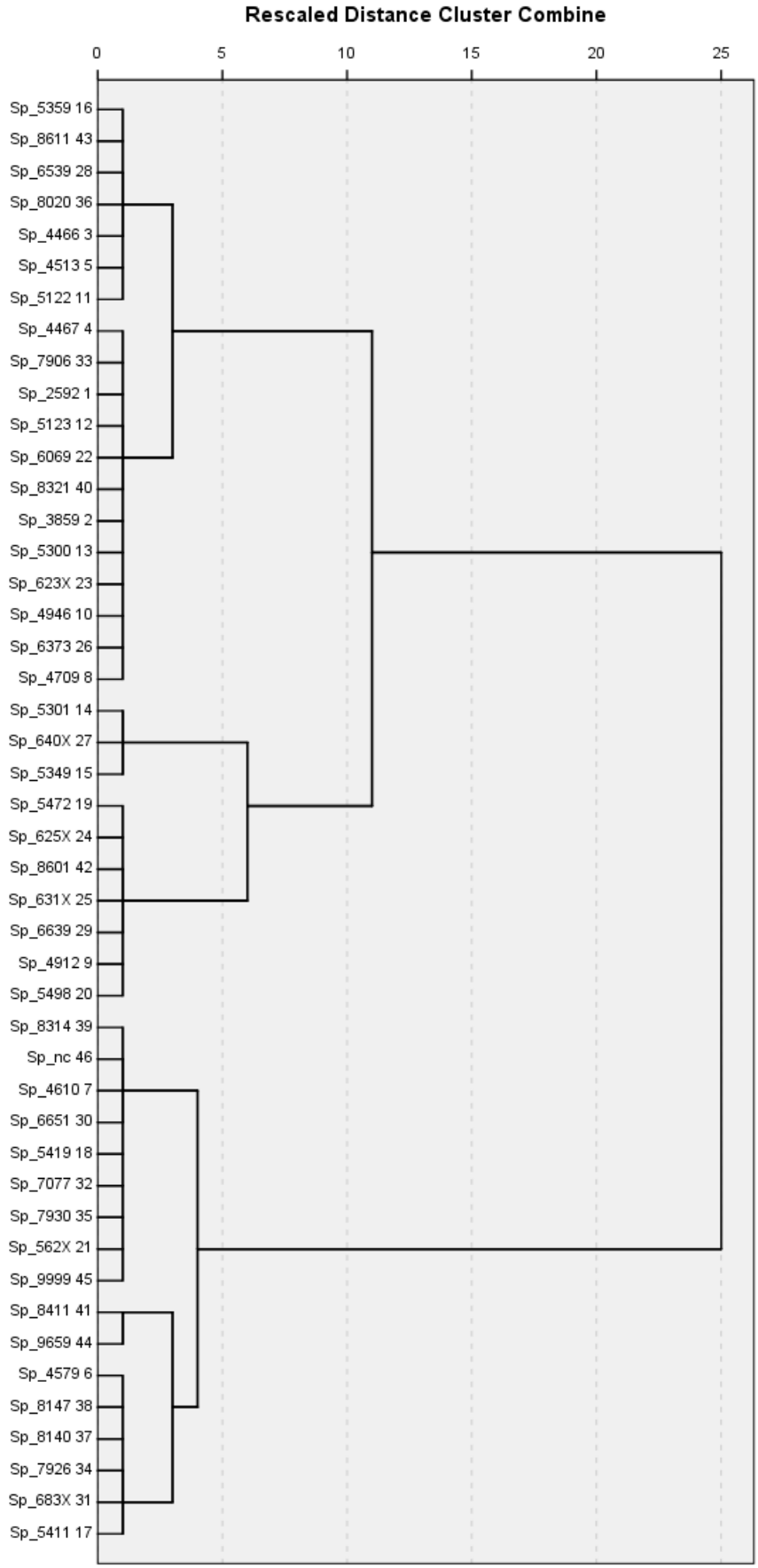

Figure B.20:Dendrogram using Ward linkage for the variable “surgical procedure”

Appendix C

ICD codes per category

ICD code	Description	Diagnosis category	First diagnosis category
C349	Malignant neoplasm of bronchus or lung, unspecified	Category 1	Category 1
D110	Benign neoplasm of parotid gland	Category 1	Category 1
D136	Benign neoplasm of pancreas	Category 1	Category 1
D180	Hemangioma, any site	Category 1	Category 1
D239	Benign neoplasm of skin, unspecified	Category 1	Category 1
D410	Neoplasm of uncertain behavior kidney	Category 1	Category 1
D649	Anemia, unspecified	Category 1	Category 1
E669	Obesity, unspecified	Category 1	Category 1
E871	Hypo-osmolality and hyponatremia	Category 1	Category 1
H050	Acute inflammation of orbit	Category 1	Category 1
I472	Ventricular tachycardia	Category 1	Category 1
I499	Cardiac arrhythmia, unspecified	Category 1	Category 1
I729	Aneurysm of unspecified site	Category 1	Category 1
J039	Acute tonsillitis, unspecified	Category 1	Category 1
J068	Other acute upper respiratory infections of multiple sites	Category 1	Category 1
J969	Respiratory failure, unspecified	Category 1	Category 1
K318	Other specified diseases of stomach and duodenum	Category 1	Category 1
K449	Diaphragmatic hernia without obstruction or gangrene	Category 1	Category 1
K610	Anal abscess	Category 1	Category 1
K623	Rectal prolapse	Category 1	Category 1
K632	Fistula of intestine	Category 1	Category 1
K802	Calculus of gallbladder without cholecystitis	Category 1	Category 1
L984	Chronic ulcer of skin, not elsewhere classified	Category 1	Category 1
M172	Posttraumatic gonarthrosis, bilateral	Category 1	Category 1
M502	Other cervical disc displacement	Category 1	Category 1
M841	Nonunion of fracture [pseudarthrosis]	Category 1	Category 1
M960	Pseudarthrosis after fusion or arthrodesis	Category 1	Category 1
N210	Calculus in bladder	Category 1	Category 1
N832	Other and unspecified ovarian cysts	Category 1	Category 1
Q521	Doubling of vagina	Category 1	Category 1
R072	Precordial pain	Category 1	Category 1
S223	Fracture of rib	Category 1	Category 1
S528	Fracture of other parts of forearm	Category 1	Category 1
S623	Fracture of other metacarpal bone	Category 1	Category 1
T814	Infection following a procedure, not elsewhere classified	Category 1	Category 1
C349	Malignant neoplasm of bronchus or lung, unspecified	Category 1	Category 1
D110	Benign neoplasm of parotid gland	Category 1	Category 1
D136	Benign neoplasm of pancreas	Category 1	Category 1
D180	Hemangioma, any site	Category 1	Category 1
D239	Benign neoplasm of skin, unspecified	Category 1	Category 1
D410	Neoplasm of uncertain behavior kidney	Category 1	Category 1
D649	Anemia, unspecified	Category 1	Category 1
E669	Obesity, unspecified	Category 1	Category 1

E871	Hypo-osmolality and hyponatremia	Category 1	Category 1
H050	Acute inflammation of orbit	Category 1	Category 1
I472	Ventricular tachycardia	Category 1	Category 1
I499	Cardiac arrhythmia, unspecified	Category 1	Category 1
I729	Aneurysm of unspecified site	Category 1	Category 1
J039	Acute tonsillitis, unspecified	Category 1	Category 1
J068	Other acute upper respiratory infections of multiple sites	Category 1	Category 1
J969	Respiratory failure, unspecified	Category 1	Category 1
K318	Other specified diseases of stomach and duodenum	Category 1	Category 1
K449	Diaphragmatic hernia without obstruction or gangrene	Category 1	Category 1
K610	Anal abscess	Category 1	Category 1
K623	Rectal prolapse	Category 1	Category 1
K632	Fistula of intestine	Category 1	Category 1
K802	Calculus of gallbladder without cholecystitis	Category 1	Category 1
L984	Chronic ulcer of skin, not elsewhere classified	Category 1	Category 1
M172	Posttraumatic gonarthrosis, bilateral	Category 1	Category 1
M502	Other cervical disc displacement	Category 1	Category 1
M841	Nonunion of fracture [pseudarthrosis]	Category 1	Category 1
M960	Pseudarthrosis after fusion or arthrodesis	Category 1	Category 1
N210	Calculus in bladder	Category 1	Category 1
N832	Other and unspecified ovarian cysts	Category 1	Category 1
Q521	Doubling of vagina	Category 1	Category 1
R072	Precordial pain	Category 1	Category 1
S223	Fracture of rib	Category 1	Category 1
S528	Fracture of other parts of forearm	Category 1	Category 1
S623	Fracture of other metacarpal bone	Category 1	Category 1
T814	Infection following a procedure, not elsewhere classified	Category 1	Category 1
C349	Malignant neoplasm of bronchus or lung, unspecified	Category 1	Category 1
D110	Benign neoplasm of parotid gland	Category 1	Category 1
D136	Benign neoplasm of pancreas	Category 1	Category 1
D180	Hemangioma, any site	Category 1	Category 1
D239	Benign neoplasm of skin, unspecified	Category 1	Category 1
D410	Neoplasm of uncertain behavior kidney	Category 1	Category 1
D649	Anemia, unspecified	Category 1	Category 1
E669	Obesity, unspecified	Category 1	Category 1
E871	Hypo-osmolality and hyponatremia	Category 1	Category 1
H050	Acute inflammation of orbit	Category 1	Category 1
I472	Ventricular tachycardia	Category 1	Category 1
I499	Cardiac arrhythmia, unspecified	Category 1	Category 1
I729	Aneurysm of unspecified site	Category 1	Category 1
J039	Acute tonsillitis, unspecified	Category 1	Category 1
J068	Other acute upper respiratory infections of multiple sites	Category 1	Category 1
J969	Respiratory failure, unspecified	Category 1	Category 1
K318	Other specified diseases of stomach and duodenum	Category 1	Category 1
K449	Diaphragmatic hernia without obstruction or gangrene	Category 1	Category 1
K610	Anal abscess	Category 1	Category 1
K623	Rectal prolapse	Category 1	Category 1
K632	Fistula of intestine	Category 1	Category 1

K802	Calculus of gallbladder without cholecystitis	Category 1	Category 1
L984	Chronic ulcer of skin, not elsewhere classified	Category 1	Category 1
M172	Posttraumatic gonarthrosis, bilateral	Category 1	Category 1
M502	Other cervical disc displacement	Category 1	Category 1
M841	Nonunion of fracture [pseudarthrosis]	Category 1	Category 1
M960	Pseudarthrosis after fusion or arthrodesis	Category 1	Category 1
N210	Calculus in bladder	Category 1	Category 1
N832	Other and unspecified ovarian cysts	Category 1	Category 1
Q521	Doubling of vagina	Category 1	Category 1
S223	Fracture of rib	Category 1	Category 1
S528	Fracture of other parts of forearm	Category 1	Category 1
S623	Fracture of other metacarpal bone	Category 1	Category 1
T814	Infection following a procedure, not elsewhere classified	Category 1	Category 1
A162	Tuberculosis of lung, without mention of bacteriological or	Category 2	Category 1
C910	Acute lymphoblastic leukemia	Category 2	Category 1
E109	Insulin-dependent diabetes mellitus without complications	Category 2	Category 1
E119	Non-insulin-dependent diabetes mellitus without complications	Category 2	Category 1
E141	Unspecified diabetes mellitus with ketoacidosis	Category 2	Category 1
E145	Unspecified diabetes mellitus with peripheral circulatory	Category 2	Category 3
E149	Unspecified diabetes mellitus without complications	Category 2	Category 3
E349	Endocrine disorder, unspecified	Category 2	Category 1
F100	Mental and behavioral disorders due to use of alcohol, acute	Category 2	Category 1
I200	Unstable angina	Category 2	Category 3
I219	Acute myocardial infarction, unspecified	Category 2	Category 1
I259	Chronic ischemic heart disease, unspecified	Category 2	Category 1
I829	Embolism and thrombosis of unspecified vein	Category 2	Category 1
J181	Lobar pneumonia, unspecified	Category 2	Category 1
J189	Pneumonia, unspecified	Category 2	Category 1
J449	Chronic obstructive pulmonary disease, unspecified	Category 2	Category 1
J984	Other disorders of lung	Category 2	Category 1
K219	Gastroesophageal reflux disease without esophagitis	Category 2	Category 2
K566	Other and unspecified intestinal obstruction	Category 2	Category 3
K703	Alcoholic cirrhosis of liver	Category 2	Category 1
K729	Hepatic failure, unspecified	Category 2	Category 3
K746	Other and unspecified cirrhosis of liver	Category 2	Category 1
K750	Abscess of liver	Category 2	Category 1
K810	Acute cholecystitis	Category 2	Category 1
K859	Acute pancreatitis, unspecified	Category 2	Category 1
K922	Gastrointestinal hemorrhage, unspecified	Category 2	Category 3
L031	Cellulitis of other parts of limb	Category 2	Category 1
S822	Fracture of the shaft of tibia	Category 2	Category 1
T302	Burn of second degree, body region unspecified	Category 2	Category 1
D179	Benign lipomatous neoplasm, unspecified	Category 3	Category 1
I859	Esophageal varices without bleeding	Category 3	Category 1
K297	Gastritis, unspecified	Category 3	Category 1
K409	Unilateral or unspecified inguinal hernia, without obstruction	Category 3	Category 2
K429	Umbilical hernia without obstruction or gangrene	Category 3	Category 2
K439	Ventral hernia without obstruction or gangrene	Category 3	Category 1

K589	Irritable bowel syndrome without diarrhea	Category 3	Category 1
K630	Abscess of intestine	Category 3	Category 1
K819	Cholecystitis, unspecified	Category 3	Category 2
M179	Gonarthrosis, unspecified	Category 3	Category 1
N189	Chronic renal failure, unspecified	Category 3	Category 1
R571	Hypovolemic shock	Category 3	Category 1

Table C.1: Selected ICD codes version 10 for the variables “diagnosis” and “first diagnosis” and their category assigned by hierarchical clustering.

ICD code	Description
Category 1	
25.90	Other operations on tongue
49.10	Incision or excision of anal fistula
49.40	Procedures on hemorrhoids
53.00	Repair of hernia
53.40	Open rep umbilical hernia
54.70	Abdomen wall repair
62.50	Orchiopexy
63.10	Spermatic varicocele
63.70	Vasectomy and ligation of vas deferens
64.00	Operations on penis
64.00	Circumcision
66.30	Other bilateral destruction or occlusion of fallopian tubes
83.20	Diagnostic procedures on muscle, tendon, fascia, and bursa, including that of hand
86.00	Operations on skin and subcutaneous tissue
86.00	Incision of skin and subcutaneous tissue
25.90	Other operations on tongue
Category 2	
6.20	Unilateral thyroid lobectomy
6.30	Other partial thyroidectomy
6.40	Complete thyroidectomy
11.60	Corneal transplant
12.5	Intraocul circ facilitat
13.9	Other operations on lens
15.00	Operations on extraocular muscles
19.10	Stapedectomy
19.50	Other tympanoplasty
20.20	Incision of mastoid and middle ear
21.20	Diagnostic procedures on nose
21.80	Repair and plastic operations on the nose
27.20	Diagnostic procedures on oral cavity
27.40	Excision of other parts of mouth
28.20	Tonsillectomy without adenoidectomy
28.30	Tonsillectomy with adenoidectomy
31.10	Temporary tracheostomy
35.30	Operations on structures adjacent to heart valves
36.00	Operations on vessels of heart
36.00	Removal of coronary artery obstruction and insertion of stent(s)
36.10	Bypass anastomosis for heart revascularization
37.20	Diagnostic procedures on heart and pericardium
39.70	Endovascular repair of vessel
43.30	Pyloromyotomy
45.70	Partial excision of large intestine
46.00	Other operations on intestine
46.50	Closure of intestinal stoma

47.00	Operations on appendix
47.00	Appendectomy
48.00	Operations on rectum, rectosigmoid and perirectal tissue
49.00	Operations on anus
49.00	Incision or excision of perianal tissue
49.30	Local excision or destruction of other lesion or tissue of anus
50.10	Diagnostic procedures on liver
50.20	Local excision or destruction of liver tissue or lesion
51.10	Diagnostic procedures on biliary tract
51.10	Endoscopic retrograde cholangiopancreatography (ERCP)
51.40	Incision of bile duct for relief of obstruction
51.50	Other incision of bile duct
51.70	Repair of bile ducts
51.80	Other operations on biliary ducts and sphincter of Oddi
52.30	Marsupialization of pancreatic cyst
53.7	abd repair-diaphr hernia
56.30	Diagnostic procedures on ureter
57.10	Cystotomy and cystostomy
57.30	Diagnostic procedures on bladder
57.80	Other repair of urinary bladder
60.20	Transurethral prostatectomy
61.00	Operations on scrotum and tunica vaginalis
65.40	Unilateral salpingo-oophorectomy
65.60	Bilateral salpingo-oophorectomy
66.50	Total bilateral salpingectomy
66.70	Repair of fallopian tube
68.40	Total abdominal hysterectomy
68.50	Vaginal hysterectomy
68.90	Other and unspecified hysterectomy
69.00	Other operations on uterus and supporting structures
69.00	Dilation and curettage of uterus
69.40	Uterine repair
69.50	Aspiration curettage of uterus
71.20	Operations on Bartholin's gland
73.50	Manually assisted delivery
76.60	Other facial bone repair and orthognathic surgery
76.70	Reduction of facial fracture
77.00	Incision, excision, and division of other bones
77.00	Sequestrectomy
77.30	Other division of bone
83.60	Suture of muscle, tendon, and fascia
83.80	Other plastic operations on muscle, tendon, and fascia
84.00	Other procedures on musculoskeletal system
84.00	Amputation of upper limb
85.20	Excision or destruction of breast tissue
85.40	Mastectomy
85.70	Total reconstruction of breast
85.80	Other repair and plastic operations on breast

86.10	Diagnostic procedures on skin and subcutaneous tissue
86.20	Excision or destruction of lesion or tissue of skin and subcutaneous tissue
86.50	Suture or other closure of skin and subcutaneous tissue
86.80	Other repair and reconstruction of skin and subcutaneous tissue
88.40	Arteriography using contrast material
99.90	Other miscellaneous procedures
Category 3	
46.10	Colostomy
51.20	Cholecystectomy
54.10	Laparotomy
60.60	Other prostatectomy
65.30	Unilateral oophorectomy
68.30	Subtotal abdominal hysterectomy
70.70	Other repair of vagina
79.30	Open reduction of fracture with internal fixation
84.10	Amputation of lower limb
84.10	Lower limb amputation, not otherwise specified

Table C. 2:Selected ICD codes version 9 for the variable “surgical procedure” and their category assigned by hierarchical cluster.

Appendix D

Categorical and dummy variables

Dummies variables	Type of variable	Description	Data mining acronym
No_1st Diagnosis	Binary (base category)	Indicates a patient without first diagnosis	1stDx_cluster=0
1st Diagnosis_category1	Binary	Indicates whether the patient has a first diagnosis classified under category 1	1stDx_cluster=1
1st Diagnosis_category2	Binary	Indicates whether the patient has a first diagnosis classified under category 2	1stDx_cluster=2
1st Diagnosis_category3	Binary	Indicates whether the patient has a first diagnosis classified under category 3	1stDx_cluster=3
No_Diagnosis	Binary (base category)	Indicates a patient without diagnosis	2ndDx_cluster=0
Diagnosis_category1	Binary	Indicates whether the patient has a medical condition classified under category 1	2ndDx_cluster=1
Diagnosis_category2	Binary	Indicates whether the patient has a medical condition classified under category 2	2ndDx_cluster=2
Diagnosis_category3	Binary	Indicates whether the patient has a medical condition classified under category 3	2ndDx_cluster=3
No addictions	Binary	Indicates that the patient does not smoke, drink or consume drugs	addiction=none
Smoking/drinking	Binary	Indicates whether the patient smokes or drinks alcohol	addiction=smoker/drinker
Smoking&drinking	Binary	Indicates whether the patients smokes and drinks alcohol	addiction=both
Age	Continuous	Patient age	Age
Number of previous admissions	Continuous	Number of previous hospitalisations	Previous_adm
A&E	Binary	Indicates whether the A&E was the first hospital area from where the patient was referred for hospitalisation	Origin=A&E
Outpatient clinic	Binary	Indicates whether the outpatient clinic was the first hospital area from where the patient was referred for hospitalisation	Origin=outpatient
Transfer(other)	Binary	Indicates whether patient was transferred from other healthcare facility.	Origin=transfer

No surgical procedure	Binary (base category)	Indicates whether a patient without surgical intervention	Sp_category=0
Surgical procedure_category1	Binary	Indicates whether the patient has a surgical procedure classified under category 1	Sp_category=1
Surgical procedure_category2	Binary	Indicates whether the patient has a surgical procedure classified under category 2	Sp_category=2
Surgical procedure_category3	Binary	Indicates whether the patient has a surgical procedure classified under category 3	Sp_category=3
Number of diagnoses	Continuous	Total number of diagnosed medical conditions	Total_2ndDx
Number of comorbidities	Continuous	Total number of previously diagnosed medical conditions in addition to the primary disease or disorder.	Totalcomorbidities
Transfusions	Binary	Indicates whether the patient had a blood transfusion in the past	Transfusions
Adult Medicine	Binary	Indicates whether the patient is treated at the adult medicine ward	Ward=adult medicine
General Surgery	Binary	Indicates whether the patient is treated at the general surgery ward	Ward=g.surgery
Trauma	Binary	Indicates whether the patient is treated at the trauma ward	Ward=trauma
Gender	Binary	Indicates whether the patient is female or male	Gender

Table D.1Dummies variables derived from the categorical variables.

Appendix E

Classification tree

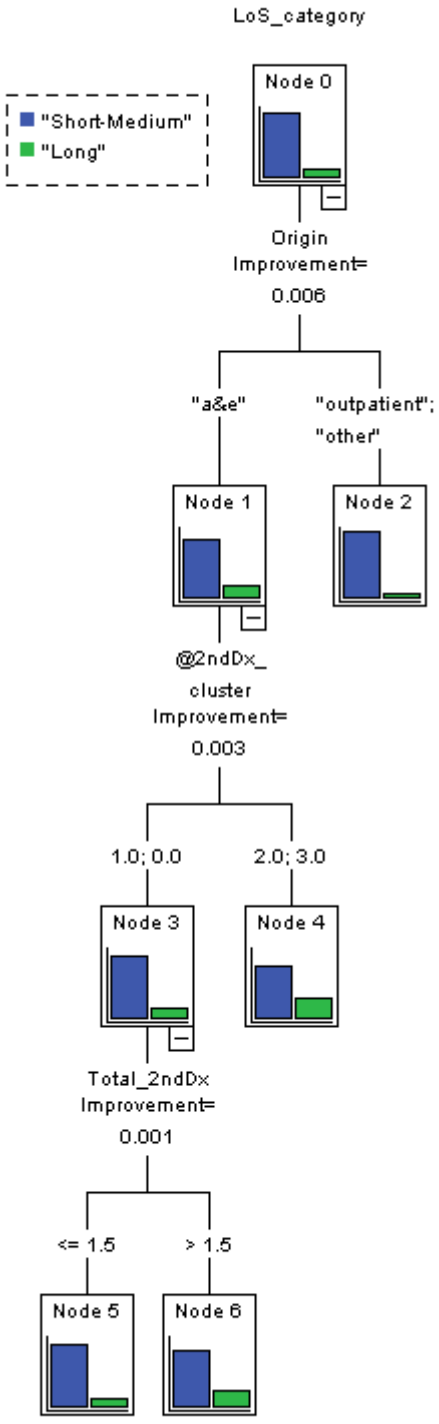

Figure E.1: CART for MRC hospital

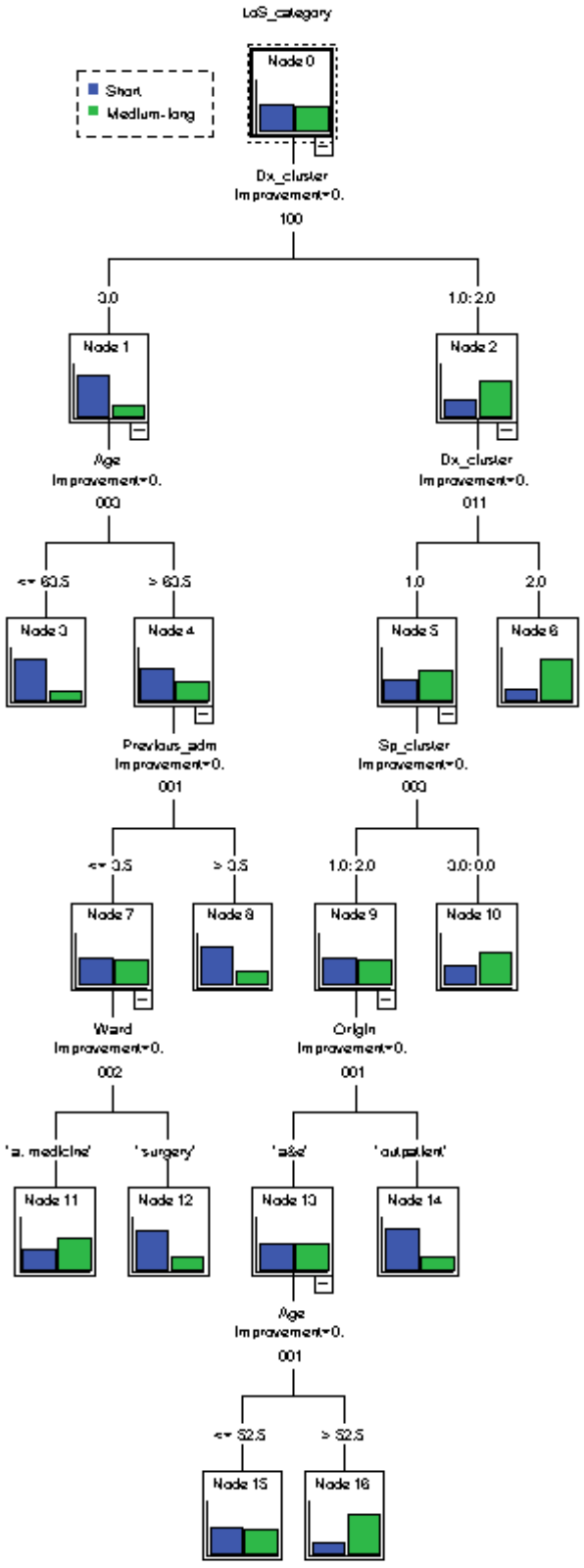

Figure E.2:CART for ISSEMyM hospital

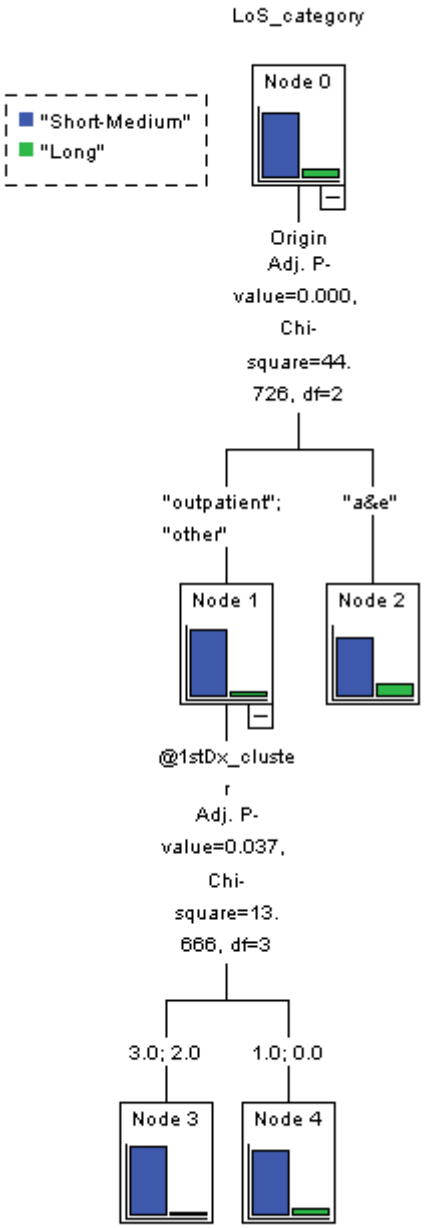

Figure E.3:QUEST for MRC hospital

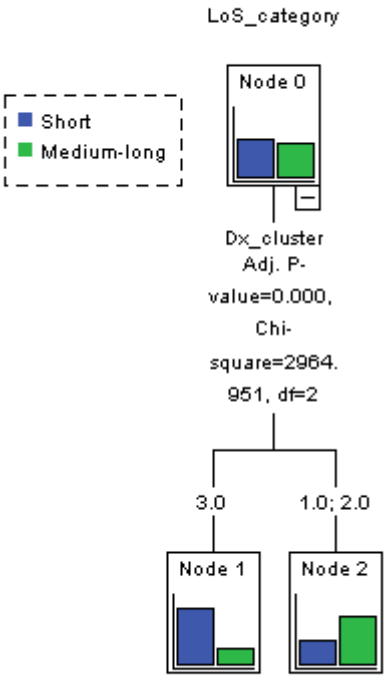

Figure E.4:QUEST for ISSEMyM hospital

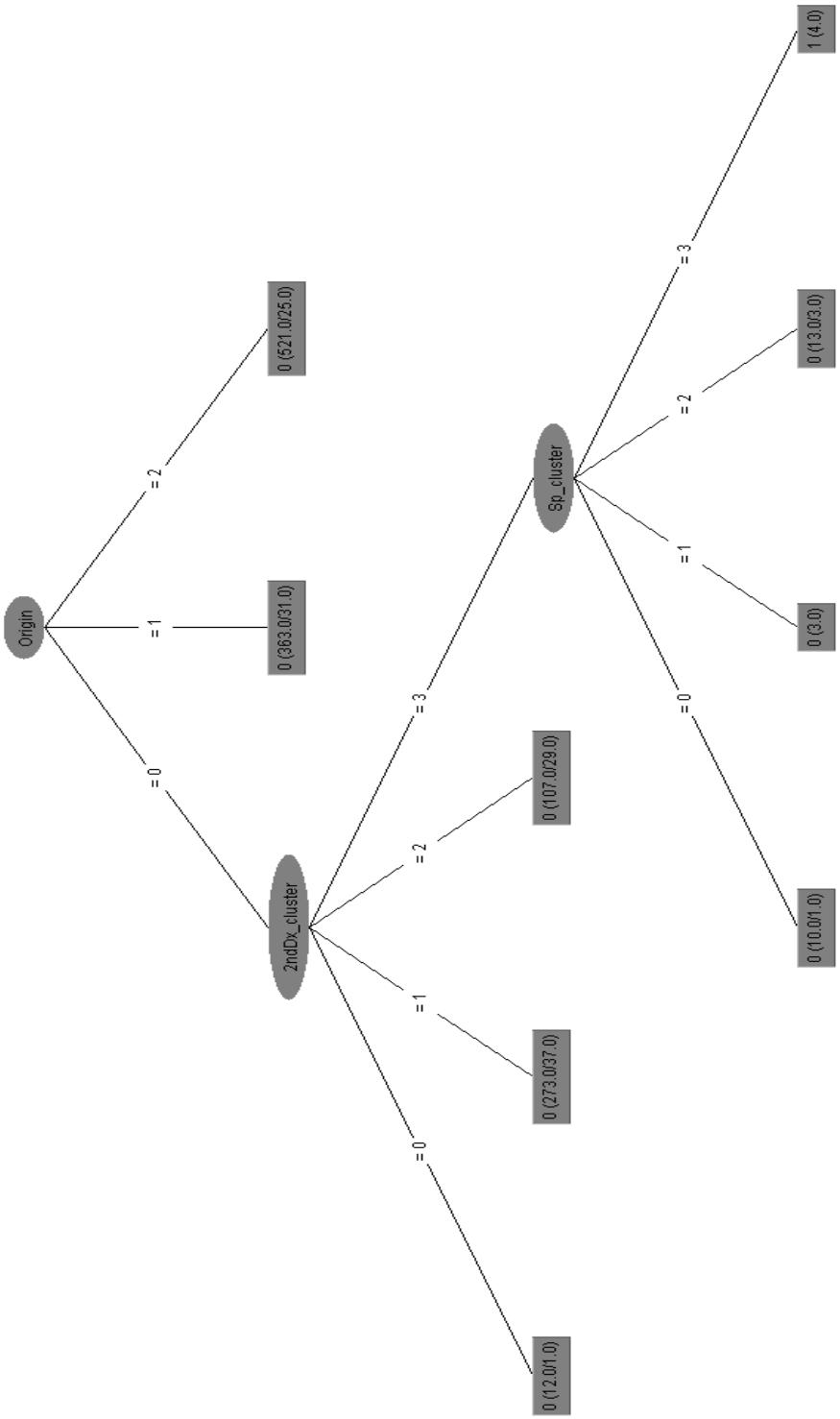

Figure E.5:C4.5 tree for MRC hospital

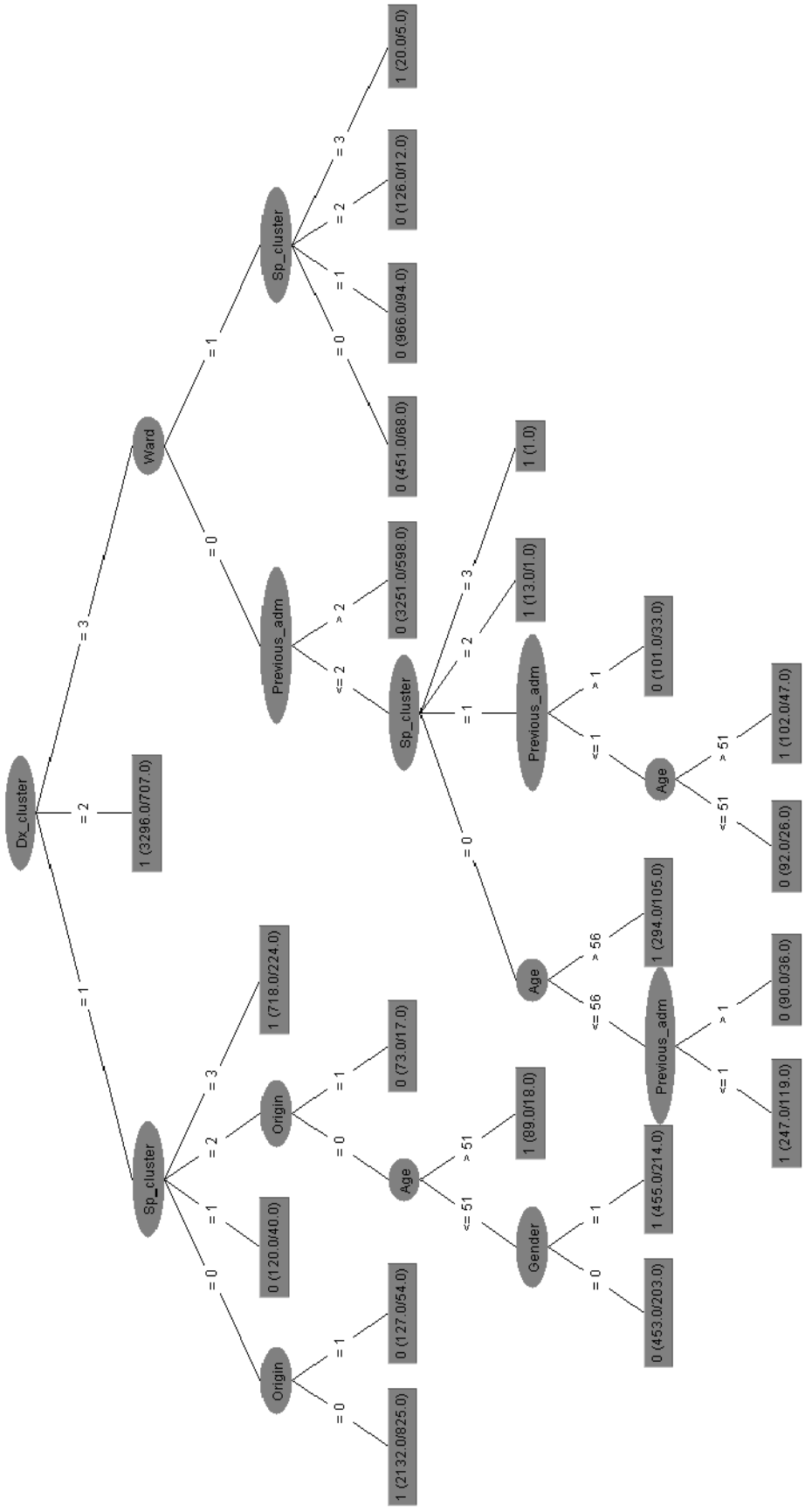

Figure E.6:C4.5 tree for ISSEMyM hospital

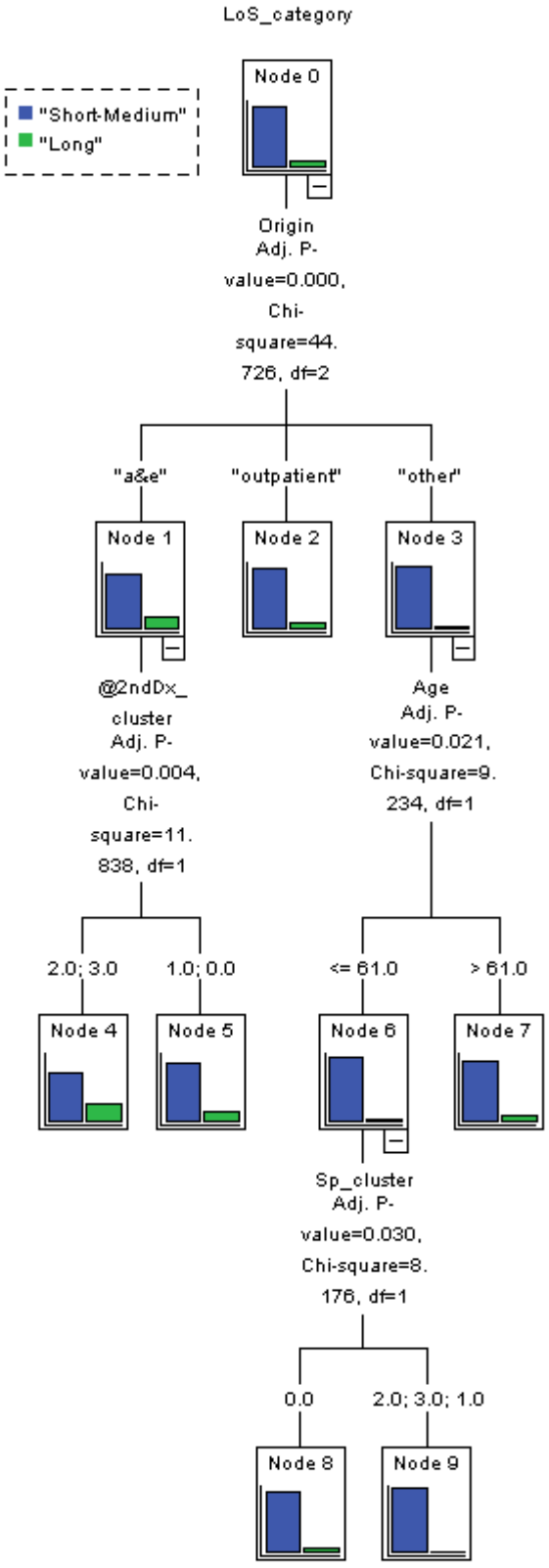

Figure E.7:CHAID tree for ISSEMyM hospital

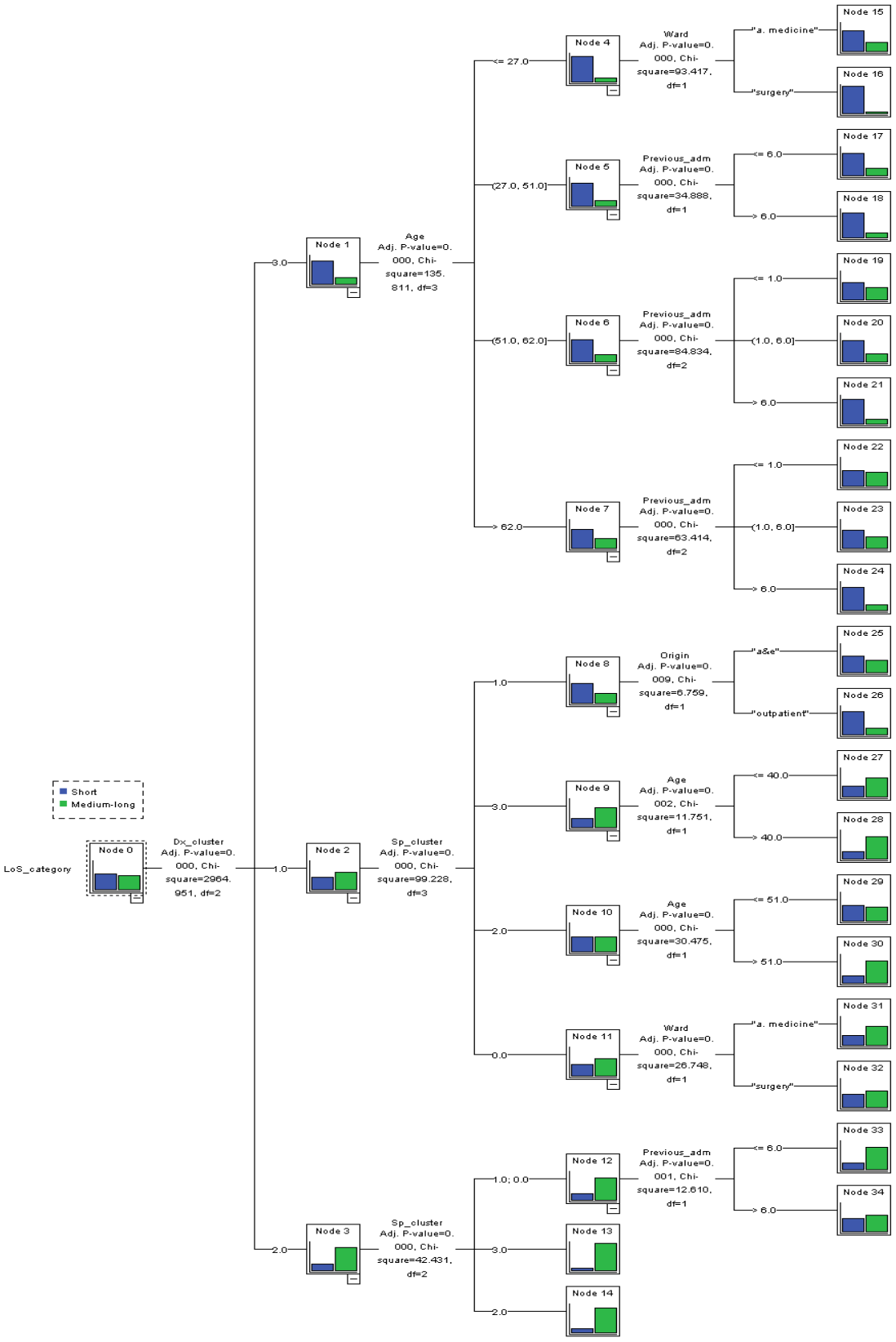

Figure E.8:CHAID tree for MRC hospital

Appendix F

Survival occupancy table

LoS\admission	Midnight Census	Monday 11:59pm	Tuesday 11:59pm	Wednesday 11:59pm	Thusrday 11:59pm	Friday 11:59pm	Saturday 11:59pm	Sunday 11:59pm	Monday 11:59pm	Tuesday 11:59pm	Wednesday 11:59pm	Thusrday 11:59pm	Friday 11:59pm	Saturday 11:59pm	Sunday 11:59pm
1	0	11	16	6	8	9	3	8	11	16	6	8	9	3	8
2	11	9.550795143	9.550795143	13.89206566	5.209524623	6.946032831	7.814286935	2.604762312	6.946032831	9.550795143	13.89206566	5.209524623	6.946032831	7.814286935	2.604762312
3	4	9.550795143	6.440074352	6.440074352	9.367380876	3.512767829	5.269151743	4.683690438	5.269151743	1.756383914	4.683690438	6.440074352	9.367380876	3.512767829	4.683690438	5.269151743
4	6	2.697188771	6.440074352	6.440074352	9.367380876	3.512767829	4.683690438	5.269151743	1.756383914	4.683690438	6.440074352	9.367380876	3.512767829	4.683690438	5.269151743
5	2	4.221419917	1.897661066	4.531043023	4.531043023	6.590608034	2.471478013	3.295304017	3.707217019	1.235739006	3.295304017	4.531043023	6.590608034	2.471478013	3.295304017
6	1	1.489117442	3.143095014	1.412920096	3.373627598	3.373627598	4.907094688	1.840160508	2.453547344	2.760240762	0.920080254	2.453547344	3.373627598	4.907094688	1.840160508
7	1	0.768561846	1.144478851	2.415662907	1.085916478	2.592841456	2.592841456	3.771405754	1.414277158	1.885702877	2.121415736	0.707138579	1.885702877	2.592841456	3.771405754
8	0	0.782309351	0.601253118	0.895336505	1.889795678	0.849522614	2.028404113	2.028404113	2.950405983	1.106402243	1.475202991	1.659603365	0.553201122	1.475202991	2.028404113
9	0	0.791432831	0.619145304	0.475851458	0.708598705	1.495646344	0.672340087	1.60534561	1.60534561	2.33504816	0.87564306	1.16752408	1.31346459	0.43782153	1.16752408
10	2	0	0.631989869	0.494411584	0.379985588	0.565843601	1.194331724	0.536889686	1.28193085	1.28193085	1.864626692	0.699235009	0.932313346	1.048852514	0.349617505
11	9	1.609394604	0	0.508560543	0.397851667	0.305773612	0.455332819	0.961075516	0.432033681	1.031566296	1.031566296	1.500460068	0.562672525	0.750230034	0.844008788
12	0	0	1.304149029	0	0.412104487	0.322393193	0.247779108	0.368972191	0.778793279	0.350092081	0.835914437	0.835914437	1.215875545	0.455953329	0.607937772
13	1	0	0	1.063693961	0	0.336121904	0.262951308	0.202094343	0.300942211	0.635201721	0.285543158	0.681791052	0.681791052	0.991696076	0.371886028
14	0	0.820630782	0.677340093	0	0.872900007	0	0.275831981	0.215785937	0.165844839	0.246962442	0.521266085	0.234325505	0.559498724	0.559498724	0.813816326
15	0	0	0	0.562135502	0	0.720482567	0	0.227668842	0.178107463	0.136886601	0.203840225	0.430247593	0.193409829	0.461804415	0.461804415
16	0	0	0	0	0.468946112	0	0.597940138	0	0.188946055	0.147814265	0.113604405	0.169170301	0.357069438	0.160513946	0.383258955
17	0	0	0	0	0	0	0	0.498815147	0	0.157623061	0.123309993	0.094771356	0.141125681	0.297875377	0.133904354
18	0	0	0	0	0	0.393126598	0	0	0.418166388	0	0.132138461	0.103373153	0.079448661	0.118308387	0.249714691
19	0	0	0	0	0	0	0.331097921	0	0	0.352186858	0	0.111289263	0.087062631	0.066913017	0.09964134
20	0	0	0	0	0	0	0	0.280084399	0	0	0.297924083	0	0.094142501	0.073648559	0.056603472
21	0	0	0	0	0	0	0	0	0.237919731	0	0	0.253073781	0	0.079970033	0.062561304
22	0	0	0	0	0	0	0	0	0	0.202902866	0	0	0.215826553	0	0.068200098
23	0	0	0	0	0	0	0	0	0	0	0.173690216	0.149214511	0	0.184753234	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0.128623348	0	0.158718573
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0.111232374	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.096489388
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	**29**	**33.73**	**42.01**	**38.69**	**36.70**	**37.00**	**31.54**	**31.71**	**35.82**	**44.10**	**40.60**	**38.36**	**38.42**	**32.74**	**32.73**

Table F.1:Survival occupancy table (Part 1)

Monday 11:59pm	Tuesday 11:59pm	Wednesday 11:59pm	Thusrday 11:59pm	Friday 11:59pm	Saturday 11:59pm	Sunday 11:59pm	Monday 11:59pm	Tuesday 11:59pm	Wednesday 11:59pm	Thusrday 11:59pm	Friday 11:59pm	Saturday 11:59pm	Sunday 11:59pm
15	16	17	18	19	20	21	22	23	24	25	26	27	28
11	16	6	8	9	3	8	11	16	6	8	9	3	8
6.946032831	9.550795143	13.89206566	5.209524623	6.946032831	7.814286935	2.604763212	6.946032831	9.550795143	13.89206566	5.209524623	6.946032831	7.814286935	2.604762312
1.756383914	4.683690438	6.440074352	9.367380876	3.512767829	4.683690438	5.269151743	1.756383914	4.683690438	6.440074352	9.367380876	3.512767829	4.683690438	5.269151743
3.707217019	1.235739006	3.295304017	4.531043023	6.590608034	2.471478013	3.295304017	3.707217019	1.235739006	3.295304017	4.531043023	6.590608034	2.471478013	3.295304017
2.453547344	2.760240762	0.920080254	2.453547344	3.373627598	4.907094688	1.840160508	2.453547344	2.760240762	0.920080254	2.453547344	3.373627598	4.907094688	1.840160508
1.414277158	1.885702877	2.121415736	0.707138579	1.885702877	2.592841456	3.771405754	1.414277158	1.885702877	2.121415736	0.707138579	1.885702877	2.592841456	3.771405754
2.950405983	1.106402243	1.475202991	1.659603365	0.553201122	1.475202991	2.028404113	2.950405983	1.106402243	1.475202991	1.659603365	0.553201122	1.475202991	2.028404113
1.605345561	2.33504816	0.87564306	1.16752408	1.31346459	0.43782153	1.16752408	1.605345561	2.33504816	0.87564306	1.16752408	1.31346459	0.43782153	1.16752408
0.932313346	1.28193085	1.864626692	0.699235009	0.932313346	1.048852514	0.349617505	0.932313346	1.28193085	1.864626692	0.699235009	0.932313346	1.048852514	0.349617505
0.281336263	0.750230034	1.031566296	1.500460068	0.562672525	0.750230034	0.844008788	0.281336263	0.750230034	1.031566296	1.500460068	0.562672525	0.750230034	0.844008788
0.683929994	0.227976665	0.607937772	0.835914437	1.215875545	0.455953329	0.607937772	0.683929994	0.227976665	0.607937772	0.835914437	1.215875545	0.455953329	0.607937772
0.495848038	0.557829043	0.185943014	0.495848038	0.681791052	0.991696076	0.371886028	0.495848038	0.557829043	0.185943014	0.495848038	0.681791052	0.991696076	0.371886028
0.305181122	0.406908163	0.457771683	0.152590561	0.406908163	0.559498724	0.813816326	0.305181122	0.406908163	0.457771683	0.152590561	0.406908163	0.559498724	0.813816326
0.671715512	0.251893317	0.335857756	0.377839976	0.125946659	0.335857756	0.461804415	0.671715512	0.251893317	0.335857756	0.377839976	0.125946659	0.335857756	0.461804415
0.383258955	0.557467571	0.209050339	0.278733785	0.313575509	0.10452517	0.278733785	0.383258955	0.557467571	0.209050339	0.278733785	0.313575509	0.10452517	0.278733785
0.319723263	0.319723263	0.465052019	0.174394507	0.232526009	0.261591761	0.087197254	0.232526009	0.319723263	0.465052019	0.174394507	0.232526009	0.261591761	0.087197254
0.11225461	0.268030197	0.268030197	0.389862104	0.146198289	0.194931052	0.219297434	0.073099145	0.194931052	0.268030197	0.389862104	0.146198289	0.194931052	0.219297434
0.210313969	0.094542746	0.225739599	0.225739599	0.328348508	0.123130691	0.164174254	0.184696036	0.061565345	0.164174254	0.225739599	0.328348508	0.123130691	0.164174254
0.084289218	0.177910092	0.079976184	0.190959037	0.190959037	0.277758599	0.104159475	0.138879299	0.156239212	0.052079737	0.138879299	0.190959037	0.277758599	0.104159475
0.048082232	0.07160009	0.151127022	0.067936351	0.162211543	0.162211543	0.235944063	0.088479024	0.117972031	0.132718535	0.044239512	0.117972031	0.162211543	0.235944063
0.053353574	0.041005521	0.061062037	0.086906288	0.057937525	0.138337359	0.138337359	0.201217976	0.075456741	0.100608988	0.113185112	0.037728371	0.100608988	0.138337359
0.058381086	0.045672069	0.03510181	0.024798026	0.074394079	0.049596053	0.118420435	0.118420435	0.172247906	0.064592965	0.086123953	0.096889447	0.032296482	0.086123953
0	0.050154265	0.039236151	0.010651796	0.021303591	0.063910774	0.042607182	0.101733118	0.101733118	0.147975444	0.055490791	0.073987722	0.083236187	0.027745396
0.136815878	0	0.043233124	0.009181879	0.009181879	0.018363758	0.055091275	0.036727517	0.087694247	0.087694247	0.127555269	0.047833226	0.063777635	0.071749839
0	0.118317204	0	0.007940411	0.007940411	0.007940411	0.015880822	0.047642465	0.031761643	0.075837237	0.075837237	0.110308709	0.041365766	0.055154354
0	0	0.102635179	0	0.006887971	0.006887971	0.006887971	0.013775942	0.041327827	0.027551885	0.065785601	0.065785601	0.095688147	0.035883055
0.083946677	0	0	0.011985 2	0	0.005992 6	0.005992 6	0.005992 6	0.011985 2	0.035955599	0.023970399	0.057234093	0.057234093	0.083249559
0	0.073239568	0	0	0.010456529	0	0.005228264	0.005228264	0.005228264	0.010456529	0.031369586	0.020913057	0.049934082	0.049934082
0	0	0.064069709	0	0	0.009147334	0	0.004573667	0.004573667	0.004573667	0.009147334	0.027442001	0.018294667	0.043682154
0	0	0	0.004011308	0	0	0.008022617	0	0.004011308	0.004011308	0.004011308	0.008022617	0.024067851	0.016045234
0	0	0	0	0.003526754	0	0	0.007053508	0	0	0.003526754	0.003526754	0.007053508	0.021160524
0	0	0	0	0	0.003108048	0	0	0.006216097	0	0.003108048	0.003108048	0.003108048	0.006216097
0	0	0	0	0	0	0.002745254	0	0	0.005490507	0	0.002745254	0.002745254	0.002745254
0	0	0	0	0	0	0	0.002430079	0	0	0.004860158	0	0.002430079	0.002430079
0	0	0	0	0	0	0	0	0.002155584	0	0	0.004311168	0	0.002155584
0	0	0	0	0	0	0	0	0	0.001915939	0	0	0.003831879	0
0	0	0	0	0	0	0	0	0	0	0.001706232	0	0	0.003412465
0	0	0	0	0	0	0	0	0	0	0	0.001522312	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0.001360658	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0.001218278
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
36.69	44.85	41.25	38.64	38.67	32.95	32.91	36.85	44.99	41.36	39.02	38.99	33.24	33.16

Table F.2:Survival occupancy table (Part 2)

References

Abbi, R., El-Darzi, E., Vasilakis, C. and Millard, P. A Gaussian mixture model approach to grouping patients according to their hospital length of stay. *Symposium on Computer-Based Medical Systems*, 2008 Jyvaskyla, Finland. Elsevier B.V., pp 524-529.

Adeyemi, S. and Chaussalet, T. 2009. Models for extracting information on patient pathways. *Intelligent Patient Management,* 189**,** 171-182.

Anderson, J. 1984. Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B (Methodological),* 46**,** 1-30.

Atienza, N., García-Heras, J., Muñoz-Pichardo, J. and Villa, R. 2008. An application of mixture distributions in modelization of length of hospital stay. *Statistics in Medicine,* 27**,** 1403-1420.

Basu, A., Manning, W. G. and Mullahy, J. 2004. Comparing alternative models: log vs Cox proportional hazard? *Health Economics,* 13**,** 749-765.

Black, N. and Payne, M. 2003. Directory of clinical databases: improving and promoting their use. *Quality and Safety in Health Care,* 12**,** 348-352.

Boaden, R., Proudlove, N. and Wilson, M. 1999. An exploratory study of bed management. *Journal of Management in Medicine,* 13**,** 234-250.

Böhning, D., Dietz, E. and Schlattmann, P. 1998. Recent developments in computer-sssisted analysis of mixtures. *Biometrics,* 54**,** 525-536.

Breiman, L. 1996a. Bagging predictors. *Machine Learning,* 24**,** 123-140.

Breiman, L. 1996b. Heuristics of instability and stabilization in model selection. *Annals of Statistics,* 24**,** 2350-2383.

Brock, G., Barnes, C., Ramirez, J. and Myers, J. 2011. How to handle mortality when investigating length of hospital stay and time to clinical stability. *BMC Medical Research Methodology,* 11**,** 144.

Bull, S. 1993. Sample size and power determination for a binary outcome and an ordinal exposure when logistic regression analysis is planned. *American journal of epidemiology,* 137**,** 676-684.

Cairns, K. J. and Marshall, A. H. Assessment of the benefits of Discrete Conditional Survival Models in modelling ambulance response times. *International Conference on Operational Research Applied to Healthcare (ORAHS)*, July 12-17 2009 Leuven, Belgium.

Carey, K. 2002. Hospital length of stay and cost: a multilevel modeling analysis. *Health Services and Outcomes Research Methodology,* 3**,** 41-56.

Channon, A. 2010. Modelling Multilevel Data. *STAT6080*. University of Southampton, unpublished.

Chen, J. S. 2003. Market segmentation by tourists' sentiments. *Annals of Tourism Research,* 30**,** 178-193.

Chen, Y. Y., Chou, Y. C. and Chou, P. 2005. Impact of nosocomial infection on cost of illness and length of stay in intensive care units. *Infection Control and Hospital Epidemiology,* 26**,** 281-287.

Chernick, M. 1999. *Bootstrap methods: a practitioner's guide*, Wiley New York.

Chertow, G. M., Burdick, E., Honour, M., Bonventre, J. V. and Bates, D. W. 2005. Acute kidney injury, mortality, length of stay, and costs in hospitalized patients. *Journal of the American Society of Nephrology,* 16**,** 3365-3370.

Clark, D. R. and Thayer, C. A. 2004. A primer on the exponential family of distributions. *Call paper program on Generalized Linear Models* [Online]. Available: https://www.casact.org/pubs/dpp/dpp04/04dpp117.pdf [Accessed 18 Jun 2011].

Classen, D. C., Pestotnik, S. L., Evans, R. S., Lloyd, J. F. and Burke, J. P. 1997. Adverse drug events in hospitalized patients. *JAMA: the Journal of the American Medical Association,* 277**,** 301.

Cleves, M. A., Gould, W. and Gutierrez, R. 2008. *An introduction to survival analysis using Stata,* Texas, Stata Corp.

Cliff, N. 1996. *Ordinal methods for behavioral data analysis,* Mahwah, Erlbaum.

Consejo Nacional de Poblacion. Available: www.conapo.gob.mx [Accessed December 12 2009].

Cosgrove, S. E., Qi, Y., Kaye, K. S., Harbarth, S., Karchmer, A. W. and Carmeli, Y. 2005. The impact of methicillin resistance in Staphylococcus aureus bacteremia on patient outcomes: mortality, length of stay, and hospital charges. *Infection Control and Hospital Epidemiology,* 26**,** 166-174.

Cots, F., Elvira, D., Castells, X. and Sáez, M. 2003. Relevance of outlier cases in case mix systems and evaluation of trimming methods. *Health care management science,* 6**,** 27-35.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society***,** 187-220.

Cramer, D. 2004. *Advanced quantitative data analysis,* Maidenhead, Open University Press.

De'ath, G. and Fabricius, K. E. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology,* 81**,** 3178-3192.

Deb, P., Gallo, W. T., Ayyagari, P., Fletcher, J. M. and Sindelar, J. L. 2011. The effect of job loss on overweight and drinking. *Journal of Health Economics,* 30**,** 317-327.

Deb, P. and Holmes, A. M. 2000. Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. *Health Economics,* 9**,** 475-489.

Deb, P. and Trivedi, P. K. 1997. Demand for medical care by the elderly: a Finite Mixture approach. *Journal of Applied Econometrics,* 12**,** 313-336.

Degroot, M. H. 1986. Probability and Statistics, California, Addison-Wesley.

Dempster, A. P., Laird, N. M. and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society,* 39**,** 1-38.

Dettling, M. and Bühlmann, P. 2003. Boosting for tumor classification with gene expression data. *Bioinformatics,* 19**,** 1061.

Dias, J. G. 2004. *Finite mixture models : review, applications, and computer-intensive methods,* Ridderkerk, Labyrint Publications.

Dietterich, T. G. 2000a. Ensemble methods in machine learning. *Multiple Classifier Systems***,** 1-15.

Dietterich, T. G. 2000b. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting, and Randomization. *Machine Learning,* 40**,** 139-157.

Dobbin, K. and Simon, R. 2011. Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics,* 4**,** 1-8.

Dodd, S., Bassi, A., Bodger, K. and Williamson, P. 2006. A comparison of multivariable regression models to analyse cost data. *Journal of Evaluation in Clinical Practice,* 12**,** 76-86.

Dougherty, J., Kohavi, R. and Sahami, M. Supervised and unsupervised discretization of continuous features. *Machine Learning International Workshop Conference* 1995. 194-202.

Dunteman, G. H. and Ho, M. R. 2005. *An introduction to generalized linear models,* London, Sage Publications.

Efron, B. 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics,* 7**,** 1-26.

Efron, B. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association,* 82**,** 171-185.

Efron, B. and Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician,* 37**,** 36-48.

El-Darzi, E., Abbi, R., Vasilakis, C., Gorunescu, F., Gorunescu, M. and Millard, P. 2009. Length of stay-based clustering methods for patient grouping. *Intelligent Patient Management,* 189**,** 39-56.

Esposito, F., Malerba, D., Semeraro, G. and Kay, J. 2002. A comparative analysis of methods for pruning decision trees. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 19**,** 476-491.

Everitt, B., Landau, S. and Leese, M. 2001. *Cluster analysis,* London, Arnold.

Fackrell, M. 2009. Modelling healthcare systems with phase-type distributions. *Health Care Management Science,* 12**,** 11-26.

Faddy, M. and McClean, S. 1999. Analysing data on lengths of stay of hospital patients using phase-type distributions. *Applied Stochastic Models in Business and Industry,* 15**,** 311-317.

Faddy, M. J. 1994. Examples of fitting structured phase–type distributions. *Applied Stochastic Models and Data Analysis,* 10**,** 247-255.

Faddy, M. J., Graves, N. and Pettitt, A. 2009. Modeling length of stay in hospital and other right skewed data: comparison of Phase-type, Gamma and Log-normal distributions. *Value in Health,* 12**,** 309-314.

Faraway, J. J. 2006. *Extending the linear model with R : generalized linear, mixed effects and nonparametric regression models,* Boca Raton, Chapman & Hall/CRC.

Field, A. P. 2009. *Discovering statistics using SPSS : (and sex and drugs and rock 'n' roll),* Los Angeles, Sage Publications.

Fisher, L. and Lin, D. 1999. Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health,* 20**,** 145-157.

Fleischmann, K. E., Goldman, L., Young, B. and Lee, T. H. 2003. Association between cardiac and noncardiac complications in patients undergoing noncardiac surgery: outcomes and effects on length of stay. *The American Journal of Medicine,* 115**,** 515-520.

Forster, A. J., Taljaard, M., Oake, N., Wilson, K., Roth, V. and Van Walraven, C. 2012. The effect of hospital-acquired infection with Clostridium difficile on length of stay in hospital. *Canadian Medical Association Journal,* 184**,** 37-42.

Fraley, C. and Raftery, A. E. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association,* 97**,** 611-631.

Freund, Y. and Schapire, R. E. Experiments with a new boosting algorithm. *Machine Learning International Workshop Then Conference*, 1996. 149-156.

Freund, Y. and Schapire, R. E. 1999. A short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence,* 14**,** 771-780.

Frick, U., Rehm, J., Krischker, S. and Cording, C. 1999. Length of stay in a German psychiatric hospital as a function of patient and organizational characteristics: a multilevel analysis. *International Journal of Methods in Psychiatric Research,* 8**,** 146-161.

Friedman, J., Hastie, T. and Tibshirani, R. 2000. Additive logistic regression: a statistical view of Boosting. *Annals of Statistics,* 28**,** 337-374.

Galski, T., Bruno, R. L., Zorowitz, R. and Walker, J. 1993. Predicting length of stay, functional outcome, and aftercare in the rehabilitation of stroke patients. The dominant role of higher-order cognition. *Stroke,* 24**,** 1794-1800.

Garg, L., McClean, S., Barton, M., Meenan, B. and Fullerton, K. An extended phase type survival tree for patient pathway prognostication. *Health Care Management (WHCM), 2010 IEEE Workshop on*, 18-20 Feb. 2010 2010. 1-6.

Garg, L., McClean, S., Meenan, B., El-Darzi, E. and Millard, P. 2009. Clustering patient length of stay using mixtures of Gaussian models and phase type distributions. *22nd IEEE International Symposium on Computer-Based Medical Systems.* Albuquerque, NM: IEEE.

Garg, L., McClean, S., Meenan, B. J. and Millard, P. 2011. Phase-type survival trees and mixed distribution survival trees for clustering patients' hospital length of stay. *Informatica,* 22**,** 57-72.

Gertman, P. M. and Restuccia, J. D. 1981. The appropriateness evaluation protocol: a technique for assessing unnecessary days of hospital care. *Medical Care,* 19**,** 855-871.

Gill, J. 2000. *Generalized linear models : a unified approach,* London, Sage Publications.

Goldstein, H. 2011. *Multilevel statistical models,* Hoboken, NJ., Wiley.

Goldstein, H., Browne, W. and Rasbash, J. 2002. Partitioning variation in multilevel models. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences,* 1**,** 223-231.

Gong, G. 1986. Cross-Validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression. *Journal of the American Statistical Association,* 81**,** 108-113.

Gorunescu, F., El-Darzi, E., Belciug, S. and Gorunescu, M. Patient grouping optimization using a hybrid self-organizing map and Gaussian mixture model for length of stay-based clustering system. *5th IEEE International Conference on Intelligent Systems (IS)*, 2010 London. IEEE, 173-178.

Gorunescu, F., McClean, S. I. and Millard, P. H. 2002. Using a queueing model to help plan bed allocation in a department of geriatric medicine. *Health Care Management Science,* 5**,** 307-312.

Graves, N., Weinhold, D., Tong, E., Birrell, F., Doidge, S., Ramritu, P., Halton, K., Lairson, D. and Whitby, M. 2007. Effect of healthcare-acquired infection on length of hospital stay and cost. *Infection Control and Hospital Epidemiology,* 28**,** 280-292.

Hagenaars, J. A. and McCutcheon, A. L. 2009. *Applied latent class analysis,* Cambridge, Cambridge University Press.

Hair, J. F. 2009. *Multivariate data analysis,* New Delhi, Pearson Education.

Harper, P. R. 2002. A framework for operational modelling of hospital resources. *Health Care Management Science,* 5**,** 165-173.

Harper, P. R. 2005. A review and comparison of classification algorithms for medical decision making. *Health Policy,* 71**,** 315-331.

Harper, P. R., Knight, V. A. and Marshall, A. H. 2011. Discrete conditional Phase-type models utilising classification trees: application to modelling health service capacities. *European Journal of Operational Research,* 219**,** 520-530.

Harrison, G. W. 1994. Compartmental models of hospital patient occupancy patterns. *In:* Millard, P. &McClean, S. (eds.) *Modelling hospital resource use: a different approach to the planning and control of health care systems.* 1994 ed. London: Royal Society of Medicine Press.

Harrison, G. W. 2001. Implications of mixed exponential occupancy distributions and patient flow models for health care planning. *Health Care Management Science,* 4**,** 37-45.

Harrison, G. W. and Escobar, G. J. 2010. Length of stay and imminent discharge probability distributions from multistage models: variation by diagnosis, severity of illness, and hospital. *Health Care Management Science,* 13**,** 268-279.

Harrison, G. W. and Millard, P. H. 1991. Balancing acute and long-term care: the mathematics of throughput in departments of geriatric medicine. *Methods of Information in Medicine,* 30**,** 221-228.

Hartigan, J. A. and Wong, M. A. 1979. Algorithm AS 136: a k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 28**,** 100-108.

Hashemi, S. and Trappenberg, T. 2002. Using SVM for classification in datasets with ambiguous data. *Sixth World Multiconference Systemics, Cybernetics, and Informatics.* Orlando, Florida.

Hastie, T. and Tibshirani, R. 1990. *Generalized additive models,* London, Chapman and Hall.

Hausman, J. and McFadden, D. 1984. Specification tests for the multinomial logit model. *Econometrica: Journal of the Econometric Society,* 52**,** 1219-1240.

Heavens, J. 1999. Casemix-The missing link in South African healthcare management: An overview of Casemix groupings such as DRGs and HRGs, their use for improved clinical and administrative healthcare management, and recommendations for a way forward in South Africa. The Health Informatics R&D Co-ordination Programme of the Informatics and Communications Group [Online]. Available: http://www.mrc.co.za/researchreports/casemix.pdf [Accessed 5 Sept 2009].

Hilbe, J. M. 2009. *Logistic regression models,* Boca Raton, CRC Press.

Hosmer, D. W. and Lemeshow, S. 2010. *Applied logistic regression,* New York, John Wiley.

Hougaard, P. 1999. Multi-state models: a review. *Lifetime Data Analysis,* 5**,** 239-264.

Iglesias, C., Nixon, J., Cranny, G., Nelson, E. A., Hawkins, K., Phillips, A., Torgerson, D., Mason, S. and Cullum, N. 2006. Pressure relieving support surfaces (PRESSURE) trial: cost effectiveness analysis. *BMJ,* 332**,** 1416.

Instituto de Salud Publica. 2006. *Encuensta Nacional de Salud y Nutricion*[Online]. Available: http://www.insp.mx/publicaciones-anteriores-2010/659-encuesta-nacional-de-salud-y-nutricion-2006.html [Accessed November 25 2008].

Irvine, V., McClean, S. and Millard, P. 1994. Stochastic models for geriatric in-patient behaviour. *Mathematical Medicine and Biology,* 11**,** 207.

Jain, A. K., Murty, M. N. and Flynn, P. J. 1999. Data clustering: a review. *ACM computing surveys (CSUR),* 31**,** 264-323.

Jiawei, H. and Kamber, M. 2001. *Data mining: concepts and techniques,* San Francisco, CA., Morgan Kaufmann.

John, G. H. and Langley, P. Estimating continuous distributions in Bayesian classifiers. *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, 1995. Morgan Kaufmann, 338-345.

Jong, J. D., Westert, G. P., Lagoe, R. and Groenewegen, P. P. 2006. Variation in hospital length of stay: do physicians adapt their length of stay decisions to what is usual in the hospital where they work? *Health Services Management Research,* 41**,** 374-394.

Jørgensen, B. 1997. *The theory of dispersion models,* London, Chapman & Hall.

Kaplan, E. L. and Meier, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association,* 53**,** 457-481.

Kass, G. V. 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics,* 29**,** 119-127.

Kleinbaum, D. G. 1998. Survival Analysis, a Self-Learning Text. *Biometrical Journal,* 40**,** 107-108.

Kleinbaum, D. G. and Klein, M. 2011. *Logistic regression : a self-learning text,* London, Springer.

Knaus, W. A., Wagner, D. P., Zimmerman, J. E. and Draper, E. A. 1993. Variations in mortality and length of stay in intensive care units. *Annals of Internal Medicine,* 118**,** 753-761.

Koh, H. C. and Tan, G. 2011. Data mining applications in healthcare. *Journal of Healthcare Information Management,* 19**,** 65.

Kohavi, R. 1996. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. *Second International Conference on Knowledge Discovery and Data Mining.* Portland, Oregon.

Lall, R., Campbell, M. J., Walters, S. J. and Morgan, K. 2002. A review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods in Medical Research,* 11**,** 49-67.

Landwehr, N., Hall, M. and Frank, E. 2005. Logistic model trees. *Machine Learning,* 59**,** 161-205.

Lee, A. H., Ng, A. S. and Yau, K. K. 2001. Determinants of maternity length of stay: a Gamma mixture risk-adjusted model. *Health Care Management Science,* 4**,** 249-255.

Leung, K. M., Elashoff, R. M., Rees, K. S., Hasan, M. M. and Legorreta, A. P. 1998. Hospital- and patient-related characteristics determining maternity length of stay: a hierarchical linear model approach. *American journal of public health,* 88**,** 377.

Leyland, A. H. and Goldstein, H. 2001. *Multilevel modelling of health statistics,* New York, Wiley.

Lim, T. S., Loh, W. Y. and Shih, Y. S. 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning,* 40**,** 203-228.

Liu, P., El-Darzi, E., Vasilakis, C., Chountas, P., Huang, W. and Lei, L. 2004. Comparative analysis of data mining algorithms for predicting inpatient length of stay. *In: Pacific Asia Conference on Information Systems.* Bangkok, Thailand.

Liu, W., White, A., Thompson, S. and Bramer, M. 1997. Techniques for dealing with missing values in classification. *Advances in Intelligent Data Analysis Reasoning about Data***,** 527-536.

Loh, W. and Shih, Y. 1997. Split selection methods for classification trees. *Statistica sinica,* 7**,** 815-840.

Long, J. 1997. *Regression models for categorical and limited dependent variables,* Sage Publications, Inc.

Long, J. S. and Freese, J. 2006. *Regression models for categorical dependent variables using Stata,* College Station, Texas, Stata Press.

Lowell, W. E. and Davis, G. E. 1994. Predicting length of stay for psychiatric diagnosis-related groups using neural networks. *Journal of the American Medical Informatics Association,* 1**,** 459.

Lunt, M. 2005. Prediction of ordinal outcomes when the association between predictors and outcome differs between outcome levels. *Statistics in Medicine,* 24**,** 1357-1369.

Lutz, R. W. Logitboost with trees applied to the WCCI 2006 performance prediction challenge datasets. *International Joint Conference on Neural Networks*, 2006 Vancouver, BC. IEEE, 1657-1660.

Maclachlan, G. J. and Krishnan, T. 1997. *The EM algorithm and extensions,* New York, Wiley.

Manning, W. G. 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics,* 17**,** 283-295.

Manning, W. G., Basu, A., Mullahy, J. and Manning, W. 2002. Modeling costs with generalized gamma regression. Available: http://citeseerx.ist.psu.edu/viewdoc/summary [Accessed 17 Feb 2009].

Manning, W. G. and Mullahy, J. 2001. Estimating log models: to transform or not to transform? *Journal of Health Economics,* 20**,** 461-494.

Marazzi, A., Paccaud, F., Ruffieux, C. and Beguin, C. 1998. Fitting the distributions of length of stay by parametric models. *Medical Care,* 36**,** 915-927.

Marshall, A., Vasilakis, C. and El-Darzi, E. 2005a. Length of Stay-Based Patient Flow Models: Recent Developments and Future Directions. *Health Care Management Science,* 8**,** 213-220.

Marshall, A. H., Burns, M. L. and Shaw, B. 2007. Patient activity in hospital using discrete conditional phase-type (DC-Ph) models. *In:* Skiadas, C. H. (ed.) *Recent Advances in Stochastic Modelling & Data Analysis.* Singapore: World Scientific Publishing.

Marshall, A. H. and McClean, S. I. 2003. Conditional phase-type distributions for modelling patient length of stay in hospital. *International Transactions in Operational Research,* 10**,** 565-576.

Marshall, A. H. and McClean, S. I. 2004. Using coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. *Health Care Management Science,* 7**,** 285-289.

Marshall, A. H., McClean, S. I., Shapcott, C. M. and Millard, P. H. 2002. Modelling patient duration of stay to facilitate resource management of geriatric hospitals. *Health Care Management Science,* 5**,** 313-319.

Marshall, A. H., Vasilakis, C. and El-Darzi, E. 2005b. Length of stay-based patient flow models: recent developments and future directions. *Health Care Management Science,* 8**,** 213-220.

Marshall, A. H. and Zenga, M. 2009. Simulating Coxian phase-type distributions for patient survival. *International Transactions in Operational Research,* 16**,** 213-226.

Marshall, A. H. and Zenga, M. 2010. Experimenting with the Coxian phase-type distribution to uncover suitable fits. *Methodology and computing in applied probability,* 14**,** 71-86.

Martin, S. and Smith, P. 1996. Explaining variations in inpatient length of stay in the National Health Service. *Journal of Health Economics,* 15**,** 279-304.

McClean, S. I., Faddy, M. and Millard, P. H. Markov model-based clustering for efficient patient care. *18th IEEE Symposium on Computer-Based Medical Systems*, 2005. IEEE, 467-472.

McClean, S. I., Garg, L., Barton, M. and Fullerton, K. Using mixed phase-type distributions to model patient pathways. *23rd International Symposium on Computer-Based Medical Systems (CBMS)*, 2010 Perth, WA. IEEE, 172-177.

McClean, S. I., McAlea, B. and Millard, P. H. 1998. Using a Markov reward model to estimate spend-down costs for a geriatric department. *Journal of the Operational Research Society,* 49**,** 1021-1025.

McClean, S. I. and Millard, P. H. 1993. Patterns of length of stay after admission in geriatric medicine: an event history approach. *The Statistician,* 42**,** 263-274.

McClean, S. I. and Millard, P. H. 1995. A decision support system for bed-occupancy management and planning hospitals. *Mathematical Medicine and Biology,* 12**,** 249-257.

McClean, S. I. and Millard, P. H. 1998. A three compartment model of the patient flows in a geriatric department: a decision support approach. *Health Care Management Science,* 1**,** 159-163.

McClean, S. I. and Millard, P. H. 2006. Where to treat the older patient? Can Markov models help us better understand the relationship between hospital and community care? *Journal of the Operational Research Society,* 58**,** 255-261.

McLachlan, G. J. and McGiffin, D. C. 1994. On the role of finite mixture models in survival analysis. *Statistical Methods in Medical Research,* 3**,** 211.

McLachlan, G. J. and Peel, D. 2000. *Finite mixture models,* New York, Wiley.

Meira-Machado, L., De Uña-Álvarez, J., Cadarso-Suárez, C. and Andersen, P. K. 2009. Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research,* 18**,** 195.

Mihaylova, B., Briggs, A., O'Hagan, A. and Thompson, S. G. 2011. Review of statistical methods for analysing healthcare resources and costs. *Health Economics,* 20**,** 897-916.

Millard, P. H. 1988. *Geriatric medicine: a new method of measuring bed usage and a theory for planning.* MD thesis, University of London.

Millard, P. H. 1994. Current measures and their defects. *In:* Millard, P. &McClean, S. (eds.) *Modelling hospital resource use: a different approach to the planning and control of health care systems.* London: Royal Society of Medicine Press.

Millard, P. H. and Tooting, L. S. W. 1992. *Flow rate modelling: a method of comparing performance in departments of geriatric medicine.* PhD thesis, University of London.

Morgan, J., Daugherty, R., Hilchie, A. and Carey, B. 2003. Sample size and modeling accuracy of decision tree based data mining tools. *Academy of Information and Management Science Journal,* 6**,** 71-99.

Nelder, J. A. and Wedderburn, R. W. M. 1972. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General),* 135**,** 370-384.

Neuts, M. F. 1994. *Matrix-geometric solutions in stochastic models : an algorithmic approach,* New York, Dover Publications.

Newburger, J. W., Wypij, D., Bellinger, D. C., Du Plessis, A. J., Kuban, K. C. K., Rappaport, L. A., Almirall, D., Wessel, D. L., Jonas, R. A. and Wernovsky, G. 2003. Length of stay after infant heart surgery is related to cognitive outcome at age 8 years. *Journal of Pediatrics,* 143**,** 67-73.

Nguyen, J. M., Six, P., Antonioli, D., Glemain, P., Potel, G., Lombrail, P. and Le Beux, P. 2005. A simple method to optimize hospital beds capacity. *International Journal of Medical Informatics,* 74**,** 39-49.

Nist/Sematech. 2003. *Engineering statistics handbook* [Online]. [Gaithersburg, Md.]: NIST. Available: http://purl.fdlp.gov/GPO/gpo5766 [Accessed September 16 2011].

O'Connell, A. 2006. *Logistic regression models for ordinal response variables*, Sage Publications, Inc.

Pan American Health Organization.*Country profile: Mexico*. Available: http://www.paho.org/English/DD/AIS/cp_484.htm [Accessed April 04 2009].

Pearson, K. 1894. Contributions to the Mathematical Theory of Evolution.*Philosophical Transactions of the Royal Society of London,* 185, 71-110.

Pendergast, J. F. and Vogel, W. B. 1988. A multistage model of hospital bed requirements. *Health Services Management Research,* 23**,** 381.

Pofahl, W. E., Walczak, S. M., Rhone, E. and Izenberg, S. D. 1998. Use of an artificial neural network to predict length of stay in acute pancreatitis. *The American surgeon,* 64**,** 868-872.

Quantin, C., Sauleau, E., Bolard, P., Mousson, C., Kerkri, M., Lecomte, P. B., Moreau, T. and Dusserre, L. 1999. Modeling of high-cost patient distribution within renal failure diagnosis related group. *Journal of Clinical Epidemiology,* 52**,** 251-258.

Quinlan, J. R. 1987. Simplifying decision trees. *International journal of man-machine studies,* 27**,** 221-234.

Quinlan, J. R. 1993. *C4.5 : programs for machine learning*, San Mateo, CA., Morgan Kaufmann Publishers.

Quinlan, J. R. 1996. Bagging, Boosting, and C4.5. *National Conference on Artificial Intelligence,* 1**,** 725-730.

Ramakrishnan, B. S. R. 2012. Generalized robust statistics method for estimating average length of stay in hospitals. *Indian Journal of Science and Technology,* 5**,** 1859-1862.

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

Rauner, M. S., Zeiles, A., Schaffhauser-Linzatti, M. M. and Hornik, K. 2003. Modelling the effects of the Austrian inpatient reimbursement system on length-of-stay distributions. *OR Spectrum,* 25**,** 183-206.

Rezanková, H. 2005. *Cluster analysis and categorical data,* Prague, Oeconomica.

Ridley, S., Jones, S., Shahani, A., Brampton, W., Nielsen, M. and Rowan, K. 1998. Classification trees: a possible method for iso-resource grouping in intensive care. *Anaesthesia,* 53**,** 833-840.

Ross, S. M. 2010. *A first course in probability,* Upper Saddle River, NJ, Pearson Education International.

Rotter, T., Kinsman, L., James, E., Machotta, A., Gothe, H., Willis, J., Snow, P. and Kugler, J. 2010. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. *Cochrane database of systematic reviews* [Online]. Available: http://apps.who.int/rhl/reviews/langs/CD006632.pdf [Accessed 25 Oct 2010].

Ruffieux, C., Paccaud, F. and Marazzi, A. 2000. Comparing rules for truncating hospital length of stay. *CaseMix Quarterly,* 2**,** 3-11.

Saltzman, J. R., Tabak, Y. P., Hyett, B. H., Sun, X., Travis, A. C. and Johannes, R. S. 2011. A simple risk score accurately predicts in-hospital mortality, length of stay, and cost in acute upper GI bleeding. *Gastrointestinal Endoscopy,* 74**,** 1225-1229.

Secretaria de Salud. 2007. Programa Nacional de Salud 2007-2012. Mexico.

Sayers, S. L., Hanrahan, N., Kutney, A., Clarke, S. P., Reis, B. F. and Riegel, B. 2007. Psychiatric comorbidity and greater hospitalization risk, longer length of stay, and higher hospitalization costs in older adults with heart failure. *Journal of the American Geriatrics Society,* 55**,** 1585-1591.

Scheaffer, R. L. and Young, L. J. 2010. *Introduction to probability and its applications,* Boston, MA, Brooks/Cole.

Singh, C. H. and Ladusingh, L. 2010. Inpatient length of stay: a finite mixture modeling analysis. *The European Journal of Health Economics : HEPAC,* 11**,** 119-126.

Snijders, T. a. B. and Bosker, R. J. 1999. *Multilevel analysis: An introduction to basic and advanced multilevel modeling,* London, Sage Publications.

Steel, F. 2009. *Module 7: Multilevel models for binary responses* [Online]. University of Bristol, Centre for Multilevel Modelling. Available: http://www.cmm.bris.ac.uk/lemma/course [Accessed December 12 2011].

Stineman, M. G., Escarce, J. J., Tassoni, C. J., Goin, J. E., Granger, C. V. and Williams, S. V. 1998. Diagnostic coding and medical rehabilitation length of stay: their relationship. *Archives of Physical Medicine and Rehabilitation,* 79**,** 241-248.

Strate, L. L. and Syngal, S. 2003. Timing of colonoscopy: impact on length of hospital stay in patients with acute lower intestinal bleeding. *The American journal of gastroenterology,* 98**,** 317-322.

Tang, X., Luo, Z. and Gardiner, J. C. 2012. Modeling hospital length of stay by Coxian phase-type regression with heterogeneity. *Statistics in Medicine* [Online], 31. Available: http://onlinelibrary.wiley.com/doi/10.1002/sim.4490/pdf [Accessed 10 Mar 2012].

Taylor, G. J., McClean, S. I. and Millard, P. H. 1997. Continuous-time Markov models for geriatric patient behaviour. *Applied Stochastic Models and Data Analysis,* 13**,** 315-323.

Taylor, G. J., McClean, S. I. and Millard, P. H. 2000. Stochastic models of geriatric patient bed occupancy behaviour. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 163**,** 39-48.

Ting, K. and Zheng, Z. 1999. Improving the performance of boosting for naive Bayesian classification. *Methodologies for Knowledge Discovery and Data Mining***,** 296-305.

Trappenberg, T. and Back, A. A classification scheme for applications with ambiguous data. *IEEE International Joint Conference on Neural Networks*, 2000. IEEE, 296-301.

Trybula, W. J. 1997. Data mining and knowledge discovery. *Annual Review of Information Science and Technology (ARIST),* 32**,** 197-229.

Tu, J. V. and Guerriere, M. R. 1992. Use of a neural network as a predictive instrument for length of stay in the intensive care unit following cardiac surgery. *Computers and Biomedical Research,* 26**,** 220-229.

Ture, M., Kurt, I., Turhan Kurum, A. and Ozdamar, K. 2005. Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications,* 29**,** 583-588.

Ture, M., Tokatli, F. and Kurt, I. 2009. Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4. 5 and ID3) in determining recurrence-free survival of breast cancer patients. *Expert Systems with Applications,* 36**,** 2017-2026.

Urbach, D. R. and Austin, P. C. 2005. Conventional models overestimate the statistical significance of volume–outcome associations, compared with multilevel models. *Journal of Clinical Epidemiology,* 58**,** 391-400.

Vasilakis, C. and Marshall, A. H. 2005. Modelling nationwide hospital length of stay: opening the black box. *Journal of the Operational Research Society,* 56**,** 862-869.

Vittinghoff, E. 2004. *Regression methods in biostatistics linear, logistic, survival, and repeated measures models,* New York, Springer.

Wang, L. M., Li, X. L., Cao, C. H. and Yuan, S. M. 2006. Combining decision tree and Naive Bayes for classification. *Knowledge-Based Systems,* 19**,** 511-515.

Weiss, J. 2010. *Statistical Methods in Ecology* [Online]. University of North Carolina. Available: http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures.html [Accessed August 07 2011].

Williams, R. 2006. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *STATA Journal,* 6**,** 58-82.

Wilson, J., Woods, I., Fawcett, J., Whall, R., Dibb, W., Morris, C. and McManus, E. 1999. Reducing the risk of major elective surgery: randomised controlled trial of preoperative optimisation of oxygen delivery. *BMJ,* 318**,** 1099-1103.

Witten, I. and Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques,* Amsterdam, Morgan Kaufmann.

World Health Organization. 2007. *Mexico: Country Cooperation Strategy at glance*[Online]. Available: http://www.who.int/countries/mex/en/ [Accessed Febuary 25 2009].

Xiao, J., Lee, A. H. and Ram Vemuri, S. 1999. Mixture distribution analysis of length of hospital stay for efficient funding. *Socioeconomics Planing Sciences,* 33**,** 39-60.

Xie, H., Chaussalet, T. J. and Millard, P. H. 2005. A continuous time Markov model for the length of stay of elderly people in institutional long-term care. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168**,** 51-61.

Yao, Z., Liu, P., Lei, L. and Yin, J. R-C4. 5 Decision tree model and its applications to health care dataset. *International Conference on Services Systems and Services Management,* 2005. IEEE, 1099-1103 Vol. 1092.

Yau, K. K. W., Lee, A. H. and Ng, A. S. K. 2003. Finite mixture regression model with random effects: application to neonatal hospital length of stay. *Computational Statistics & Data Analysis,* 41**,** 359-366.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. and Ruzzo, W. L. 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics,* 17**,** 977-987.