• U C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

MIRIAM RAQUEL SEOANE PEREIRA SEGURO SANTOS

Sistema de Apoio à Análise e ao Tratamento de Doentes com Carcinoma Hepatocelular

Dissertação apresentada à Universidade de Coimbra para
cumprimento dos requisitos necessários à obtenção
do grau de Mestre em Engenharia Biomédica

Orientadores:
Professor Doutor **Alberto Cardoso**
Professor Doutor **Pedro H. Abreu**

Coimbra, 2014

Este trabalho foi desenvolvido em colaboração com:

Centro de Informática e Sistemas da Universidade de Coimbra (CISUC)

CISUC

Centro Hospitalar e Universitário de Coimbra (CHUC)

CHUC
CENTRO HOSPITALAR
E UNIVERSITÁRIO
DE COIMBRA

Abstract

Liver cancer is the sixth most frequently diagnosed cancer and the third cause of cancer-related deaths worldwide. Hepatocellular Carcinoma (HCC) represents more than 90% of primary liver cancers and it's a major global health problem. Clinical guidelines aim to assist clinicians in their decision-making process, under the assumptions of Evidence-Based Medicine (EBM). However, clinical practice often deals with the mismatch between EBM and the desired Personalized Medicine (PM), adjusted to a given patient. In order to make a reasoned decision, clinicians frequently need to access the patient's information, which is a difficult quest in the great majority of hospital contexts. The patient's clinical files are often dispersed in physical files, subjected to loss and inconsistency. Furthermore, such scenario also makes patient's clinical data susceptible to missing data.

In this work, we present a Clinical Decision Support System (CDSS) for managing clinical data of HCC patients, and an Artificial Intelligence (AI) module to be integrated with the developed CDSS. We have conducted several clustering approaches to profile a HCC patients database with heterogeneous and missing data. Our analysis led to the patients division into two groups, G1 and G2, with statistically significant overall survivals. HCC stage C patients were present in both groups, which suggested some heterogeneity between these patients. We have also performed some classification studies in order to access group assignment for a new patient presented to our CDSS.

In brief, we have developed a framework that allows cancer data management in the HCC context. Our results show that it is possible to develop a CDSS for HCC patients which integrates clinical data management with AI techniques, targeting the treatment of these patients within the paradigms of PM. We have demonstrated that CDSSs allow the clinicians access to the patients' clinical data at all times, while supporting them in their daily decisions.

Keywords: Hepatocellular Carcinoma (HCC), Evidence-Based Medicine (EBM), Personalized Medicine (PM), Missing Data (MD), Imputation, Clinical Decision Support System (CDSS), Profiling Prognostic Groups, Cancer Data, Clustering, Artificial Intelligence (AI), clinical data

Resumo

O Cancro do fígado é o sexto cancro mais frequentemente diagnosticado e a terceira causa de morte por doenças relacionadas com cancro em todo o Mundo. O Carcinoma Hepatocelular (CHC) está na origem de mais de 90% dos tumores primários do fígado, sendo considerado um problema à escala global.

As *guidelines* clínicas, suportadas pela Medicina Baseada na Evidência (MBE), procuram auxiliar os clínicos no seu processo de tomada de decisão. No entanto, a prática clínica lida frequentemente com o desfasamento entre a MBE e a desejada Medicina Personalizada (MP), ajustada a um dado doente. De modo a poderem tomar decisões fundamentadas, os clínicos necessitam de ter a informação dos doentes disponível para consulta, a qualquer altura. Na maioria dos contextos hospitalares, a informação clínica do doente está muitas vezes registada em suporte físico (papel), distribuída por várias instalações. Isto torna os ficheiros igualmente susceptíveis a dados em falta.

Neste trabalho, apresentamos um Sistema de Apoio à Decisão Clínica, para a gestão de dados clínicos de doentes com CHC. É também apresentado um módulo de Inteligência Artificial a ser integrado no sistema. Vários métodos de análise de agrupamentos foram utilizados de modo a determinar grupos prognósticos com diferentes características, considerando dados heterogéneos e com valores em falta. A análise propiciou a divisão em dois grandes grupos, G1 e G2, com sobrevivências globais estatisticamente significativas. Os nossos resultados sugerem igualmente uma heterogeneidade entre os doentes no estádio avançado da doença. Foram ainda avaliados alguns métodos de classificação, de modo a desenvolver modelos preditivos para a atribuição do grupo mais correcto para um determinado doente.

Em resumo, este trabalho foca-se no desenvolvimento de uma ferramenta que alie a gestão de dados clínicos a um ”motor inteligente” de inferência que permita gerar recomendações úteis aos clínicos nas suas actividades diárias. O sistema integra algoritmos de Inteligência Artificial que permitem orientar os tratamentos dos doentes no âmbito da Medicina Personalizada.

Palavras-Chave : Carcinoma Hepatocelular (CHC), Medicina Baseada na Evidência (MBE), Medicina Personalizada (MP), Preenchimento de dados em falta, Sistema de Apoio à Decisão Clínica (SADC), Personalização de Grupos Prognósticos, Métodos de Agrupamento, Inteligência Artificial (IA), dados clínicos

Acknowledgements

I would like to express my sincere gratitude to my advisors, Prof. Alberto Cardoso and Prof. Pedro Henriques Abreu, for their continuous support, patience, caring, enthusiasm and availability throughout my thesis. Professor Alberto Cardoso, who has always provided me the opportunities and resources I needed. His kindness and motivational words in hard times gave me confidence to work in my own way and concretize my ideas. Professor Pedro Henriques Abreu, for his encouragement and fearless honesty. His unlimited willingness to give his time and knowledge so generously made him more of a mentor and friend than a professor, and for that I owe him my deepest respect, admiration and trust.

I must also acknowledge Prof. Armando Carvalho, Dr. Adélia Simão, and further CHUC's team members, Dr. Lurdes Correia, Dr. Pedro Correia and Dr. Raquel Silva, for the opportunity to work with them. This research would not have been possible without their hard work, wise assistance, insightful comments and hard questions. A special appreciation also goes to HEPATOMED - Associação para a Promoção da Hepatologia.

Throughout my academic journey I have been blessed with the truest and extraordinary friends. Inês Lopes and Inês Barroso, for standing by me no matter what. Marta Pinto, for believing in me until I learned to believe in myself. Sara Santos, Diana Capela, Carolina Queijo, Patrícia Santos, Sofia Prazeres and Joana Paiva, whose smiles always encouraged me to be myself. Mariana and Bruna Nogueira, Diogo Passadouro, Diogo Martins and Heloísa Sobral, for showing me that humility, kindness, hard work, honesty and courage are always rewarded. Last, but by no means least, a very special thanks to Bruno Andrade, for being π times weirder than me.

Finally, I am truly grateful to my loving family. My mother, for teaching me there is always a bigger treasure than the one we're sad to lose, and my sister, who is truly "my ray of sunshine on a rainy day".

”In an era when today’s truths become tomorrow’s outdated concept, an individual who is unable to gather pertinent information is almost as helpless as those who are unable to read and write.”

Breivik and Gee, 1989

Contents

Abbreviations

AI Artificial Intelligence

AL Average Linkage

ANN Artificial Neural Network

APEF Portuguese Association for the Study of the Liver

APMGF Portuguese Association of Family Medicine

AUC Area Under the Curve

Anti-HCV HCV Antibody

BCLC Barcelona-Clinic Liver Cancer

CDSS Clinical Decision Support System

CIS Clinical Information System

CL Complete Linkage

CP Child-Pugh

DT Decision Trees

EASL-EORTC European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer

EBM Evidence-Based Medicine

ECOG Eastern Cooperative Oncology Group

G1 Group 1

G2 Group 2

GA Genetic Algorithms

HBV Hepatitis B Virus

HBcAb Hepatitis B Core Antibody

HBeAb Hepatitis B e-Antibody

HBeAg Hepatitis B e Antigen

HBsAb Hepatitis B Surface Antibody

HBsAg Hepatitis B Surface Antigen

HCC Hepatocellular Carcinoma

HCVAg HCV Core Antigen

HCV Hepatitis C Virus

HEOM Heterogeneous Euclidean-Overlap Metric

HIV Human Immuno-deficiency Virus

IARC International Agency for Research on Cancer

INR International Normalized Ratio

KNN k-nearest neighbours

LDA Linear Discriminant Analysis

LD Listwise Deletion

LR Logistic Regression

MARS Multivariate Adaptive Regression Splines

MAR Missing At Random

MCAR Missing Completely At Random

MD Missing Data

MLP Multi-Layer Perceptron

ML Machine Learning

MNAR Missing Not At Random

NAFLD Non-alcoholic fatty liver disease

NASH Nonalcoholic Steatohepatitis

PACS Picture Archiving and Communication System

PAM Partition Around Medoids

PCA Principal Components Analysis

PD Pairwise Deletion

PEI Percutaneous Ethanol Injection

PG1 Prognostic Group 1

PG2 Prognostic Group 2

PM Personalized Medicine

PS Performance Status

RFA Radiofrequency Ablation

RI Regression Imputation

SI Statistical Imputation

SL Single Linkage

SOM Self-Organizing Maps

SPH Portuguese Society of Hepatology

SVMI Support Vector Machines Imputation

SVM Support Vector Machines

TACE Chemoembolization

WHO World Health Organization

WPGMA Weighted average distance

List of Figures

List of Tables

Chapter 1

Introduction

This project was developed in the Department of Informatics Engineering (DEI) of the Faculty of Sciences and Technology of the University of Coimbra, within the Biomedical Engineering Master's program. The work results from a collaboration with Coimbra Hospital and Universitary Centre (CHUC), more specifically at the Service of Internal Medicine A. The aim of this chapter is to provide an overview of our work. The first two sections focus on contextualization and motivation for this work. Its objectives and planning are stated in the third and forth sections. Finally, the thesis structure is presented.

1.1 Contextualization

For the past few years, we have been witnessing an exponential growth of cancer incidence and related deaths worldwide. Solely in 2012 were reported about 14,1 millions of new cancer cases and 8,2 millions of deaths, according to the statistics published by GLOBOCAN [1]. Liver cancer is the sixth most frequently diagnosed cancer and the third cause of cancer-related deaths worldwide, accounting for 7% of all cancers [2]. Hepatocellular Carcinoma (HCC) represents more than 90% of primary liver cancers and is a major global health problem [3].

In the last decade, liver cancer has been of great concern to Portuguese League Against Cancer, Portuguese Association for the Study of the Liver (APEF) and other entities of reference in Portugal, as the Portuguese Association of Family Medicine (APMGF) and the Portuguese Society of Hepatology (SPH). In 2010, SPH predicted an increasing number of liver cancer cases by approximately 70% by the end of 2015, seeking a greater national awareness regarding liver diseases [4]. Other several studies concerning this neoplasia have sought to define its dimension in Portugal. According to the work of Tato Marinho *et al.* [5], HCC patients' hospital admissions tripled from 1993 to 2005, with the overall costs of admission rising proportionally. Despite the significant growth of this disease in the last decades, the epidemiological data of HCC in Portugal are scarce and scattered [6,7], complicating the planning of health promoting activities such as vaccination and screening, but also compromising the patient's healing process, caused by the lack of information and case studies regarding this pathology.

1.2 Motivation

When treating patients, physicians are often faced with difficult decisions and considerable uncertainty regarding their options. They rely on clinical guidelines, professional experience, knowledge, previous decisions and observed outcomes to guide their decisions. Clinical guidelines are summarized consensus statements on best practice regarding a certain disease, and they intend to assist physicians and other healthcare professionals in the decision-making process, under the assumptions of Evidence-Based Medicine (EBM) [3]. However, these guidelines are not limited to a "cookbook" or a blind application of protocols. Guidelines have to be adapted to each hospital's regulations, team capacities, infrastructures and cost-benefit strategies. Moreover, the application of EBM to an individual patient may turn out to be an infeasible task. Clinical practice often deals with the mismatch between EBM and the desired Personalized Medicine (PM), adjusted to a specific patient [8]. Given the biological variability among patients, the applicability of a given therapeutic to a particular case must be evaluated by the clinician. In order to make a reasoned decision, it is fundamental that the patients' information is available for clinicians to consult at all times, which may not happen in most cases.

In the majority of hospital contexts, the patient's clinical information is dispersed in physical files [9], sometimes divided in multiple facilities, turning the access and share of existing information into a problematic issue. Every day, a large amount of clinical information is generated. Laboratory results, imaging findings, pathological information and several other patient variables evolving in time are managed by various people within the institutions, recorders in different times, formats and types of files. Without a proper registration system, these data are subjected to loss and inconsistency. This scenario also makes datasets compiled from patient's clinical information susceptible to missing data.

1.3 Objectives

In our work, we focus on the development of a web-based registration system to store relevant clinical information of HCC patients of CHUC. Our system can be accessed through a standard web browser and allows the clinician to access all patients information, inserting new information, editing the existing records and search for particular fields or cases, if necessary. Furthermore, a reporting system is included, in a way that it is possible to consult some aspects regarding the demographic and epidemiological characterization, risk factors, stage of tumours and survival analysis. However, we want this system to be more than a tool for data collection and storage, a HCC *recommendation* system that supports medical decision, based on case-based reasoning. Besides allowing the information retrieval and management, it should analyse the complete patients' clinical information and assess the best treatment choices that maximize the overall survival of each patient. Our main goals can be described as follows:

- To develop a web-based application for managing clinical data of HCC patients: a Clinical Decision Support System (CDSS). The system should be build so that data entry is constrained to a set of rules, in order to avoid inconsistency in patient's records and to enable automatized patient's data consultations. Thus, the entry fields are predefined, default values are settled when applicable and some data structures have to respect some constraints.

- To build a "data mining" module, that should be integrated with the web-application. This is intended to be an inference motor that can assist physicians in their daily activities, by analysing the available patients' information in the database and generating a set of appropriate recommendations.

1.4 Planning

In this section we present a visual comparison between the expected scheduling (the one defined at the beginning of the thesis) and the real schedule (during the development of the thesis) (Figure 1.1).

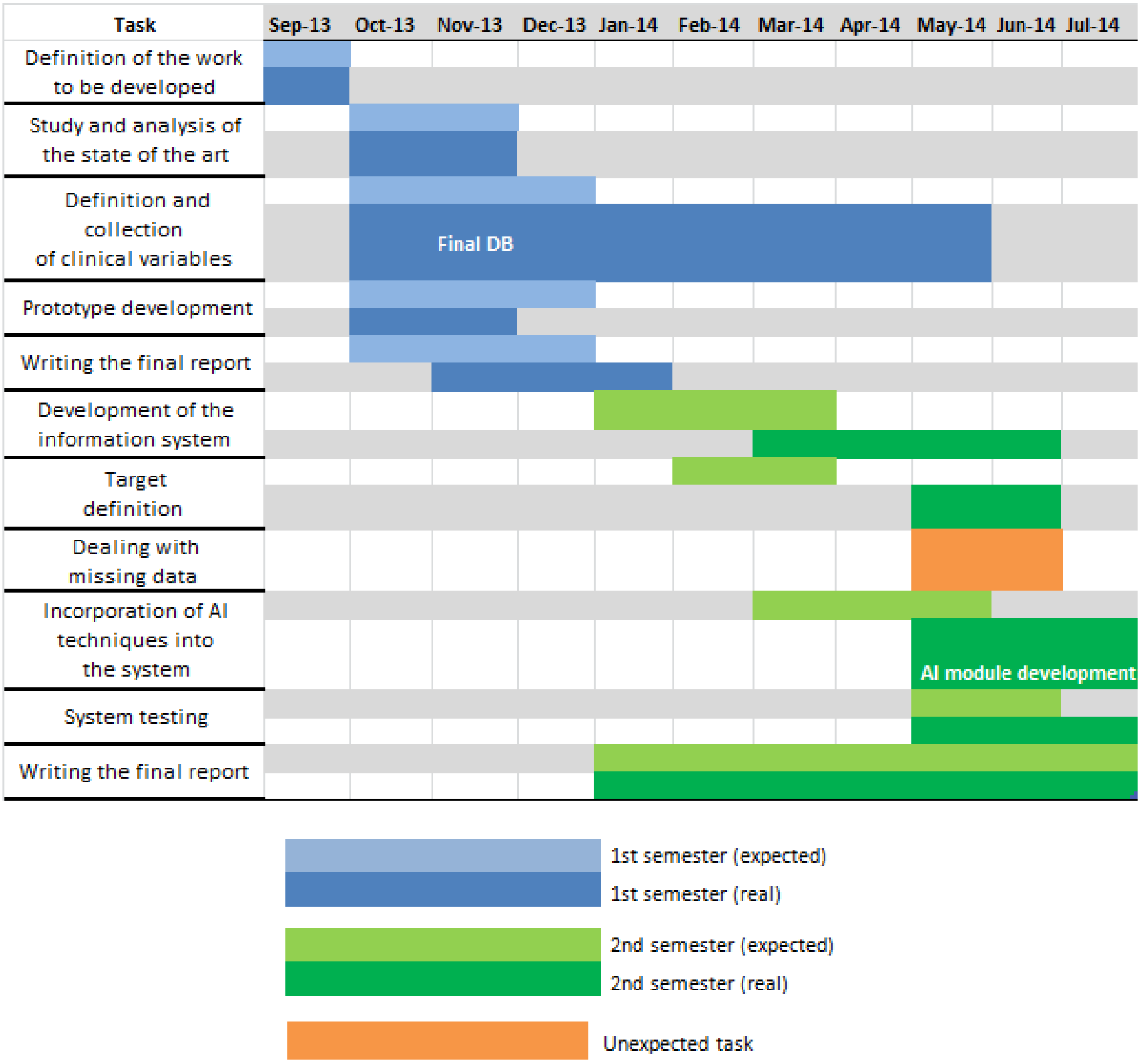


Figure 1.1: Project's expected vs. real work plan.

As analysed in Figure 1.1, the schedule was composed by 11 tasks:

- **Definition of the work to be developed:** At this phase, it was important to define the project's objectives, methodologies, scheduling and work plan. During this task, we started contacting CHUC's team to understand what are their needs and expectations towards the project. Getting in contact to their system, evaluating its flaws and suggest new approaches were the most important objectives performed in this task.

- **Study and analysis of the state of the art:** In order to develop an up-to-dated clinical information system, is was fundamental to study the state-of-the-art on recommendation systems, whether they were developed for HCC in particular or similar diseases. This task was mainly focused on the analysis of similar work, identifying the requisites that met our objectives and exposing their advantages and disadvantages.

- **Definition and collection of clinical variables:** For the purpose of establishing a complete and appropriate registry for HCC patients, it was required to review the state-of-the-art on the management of Hepatocellular Carcinoma, according to the current evidence on the matter. The validation of the variables was performed through continuous contact with CHUC's clinicians, in order to guarantee the consistency of our set of proposed variables. During this process, some variables were added to the initial set, while others were discarded. The experience and expertise of clinicians was essential to define our final set of features. The collection of data consisted in retrieving the patients' physical files, currently available at CHUC's Service of Internal Medicine, and gathering each patient's follow-up data. Each patient's file was reviewed by five clinicians which used a cross check validation in order to avoid error in the stored data. In February also took place the project's first intermediate presentation and a poster presentation at Congresso Português de Hepatologia, in the 17^{th} APEF's Annual Reunion.

- **Prototype development:** In this task, the system's requirements analysis was made through successive meetings with CHUC's team. We documented the list of our system's features and their priority. After gathering all the fundamental requirements, a prototype was developed, successfully validated by CHUC's team.

- **Writing the intermediate report:** The intermediate report was written, based on the state-of-the-art regarding clinical decision support systems and management of Hepatocellular Carcinoma.

- **Development of the clinical information system:** After the validation of the prototype, the next task was the development of the system. The final set of variables was defined, as well as the system's requirements including functional and non-functional requirements, database structure and other aspects related to the system interface design. In consequence, at this phase, the prototype was improved. These improvements included new application and forms layout and a reporting tab, web access and database implementation.

- **Target definition:** Target definition is an iterative process, in which we seek to identify the most influential factors to patient's personalization. At this point, the tumour stage, the performed treatment and the overall survival can be seen as target variables. However, the choice of target varies according to the data that is used. In our case, it may depend on the clinicians' needs, or timing of the analysis, i.e., risk factor analysis, first medical evaluation or other follow-up data.

- **Dealing with missing data:** This was not initially covered in the original project proposal. Patients' data contained a lot of missing values, which meant that a literature review of research works in the area of missing data was performed in order to overcome the problem. According to this review, we've selected the most appropriate approaches to overcome this issue regarding our dataset.

- **Incorporation of Artificial Intelligence (AI) techniques into the system:** This task was not completely fulfilled. The data mining module was fully developed, but was not integrated in the developed platform. The data collection process was very time consuming, the missing data issues were not expected, and thus there wasn't enough time to rewrite the code from MATLAB to PHP or JavaScript. For that reason, we called this task "AI module development", which consisted in a study of AI techniques to profile HCC patients according to their characteristics, aiming to achieve the fittest survival estimation function to each group. At the end of May took place the project's second intermediate presentation.

- **System Testing:** This task consisted in the validation of the defined systems requirements.
- **Writing the final report:** The writing of the final report concludes our work.

There is a clear difference between the expected and real work plan. This was mainly due to the delay in the data collection. Gradually changing the data affects our study. Some variables were added or discarded after data examination (pre-processing, correlation between variables, distance metrics). These frequent adjustments in the dataset made it even more susceptible to missing data and erroneous values. This slowed down the data pre-processing and the data importation to the system. Moreover, when different sets of patients are considered, the conclusions from the previous analysis can not be accepted. This changes forced us to update our study files and remake our analysis more often.

Updated patient's info are a more problematic issue. If a new patient is inserted, we need to add new information in the files. Or, if a patient is removed from the study, we simply remove his information. However, if something changes in the previous entered patient's file, this requires a closer examination. The clinicians could have entered variables that previously were missing, or delete them if they found out they had made a mistake in the previous registry.

As a final remark, in spite of all these issues, the majority of the project's goals were accomplished. The incorporation of the data mining module into the developed system was the only goal that wasn't met.

1.5 Document Structure

The remainder of this thesis is organized as follows: Chapter 2 presents some background regarding Hepatocellular Carcinoma. Chapter 3 exposes a brief review of the literature, considering Decision Support Systems. Chapter 4 deals with some aspects of Missing Data theory and Chapter 5 presents our software implementation and further details on our clinical decision support system. Finally, Chapter 6 reports the achieved results and Chapter 7 presents the conclusions and proposals for further studies.

Chapter 2

Hepatocellular Carcinoma

In order to design an appropriate CDSS for HCC patients, it's fundamental to understand some underlying aspects of this pathology. In this chapter we'll review some important concepts in HCC characterization, in particular its etiology and risk factors, staging system and treatment allocation.

The cell is the structural and functional unit of any living organism. The human body consists of trillions of cells. All of them have a useful lifetime. They grow, divide themselves and die when they become older or suffer irreparable structural damages. During the early years of someone's life, normal cells divide too quickly to allow the person to growth. However, when the individual reaches adulthood, most cells divide only to replace worn-out or damaged cells. Cancer arises when there is a proliferation of abnormal cells. The division process, that is usually controlled, goes wrong. New cells are formed without the body's need while the worn-out cells do not die. However, not all tumours are necessarily cancer - there are malignant and benign tumours. Only malignant tumours are cancer. Malignant tumours can invade surrounding tissues and organs, and even free themselves from the primary tumour and enter the bloodstream or lymphatic system, "travelling" to other distant organs. In this case, we are dealing with the process of metastasis: from the original cancer (primary tumour), new tumours are formed in other organs - these are called secondary tumours.

The human body is composed of four types of tissues: connective, nervous, muscular and epithelial. Epithelial tissue is widely distributed throughout the body because it is responsible for coating the skin and internal organs. Each organ has its own epithelial tissue, often consisting of more than one type of epithelial cell, each with a different function in the body. A Carcinoma is a type of cancer that arises when an epithelial cell undergoes a malignant transformation. Most cancer names derive from the origin of their primary tumour. Thus, when the source of cancer is an epithelial cell cancer of the liver, known as hepatocyte, the cancer is called hepatocellular carcinoma. HCC may have different growth patterns. Some malignant tumours begin as a single tumours that grow larger and only spread to other parts of the liver in later stages. A second pattern is described by the appearance of small cancerous nodules scattered throughout the liver. This pattern is particularly common in patients with cirrhosis, and the most frequently detected in Portugal.

2.1 Etiology and risk factors

Approximately 90% of HCCs are associated with a known underlying risk factor. The most frequent factors include chronic viral hepatitis (types B and/or C), alcohol intake and aflatoxin exposure. Worldwide, approximately 54% of cases are associated with Hepatitis B Virus (HBV) and 31% with hepatitis C Virus (HCV), leaving around 15% associated with other causes (Table

2.1).

Table 2.1: Geographical distribution of main risk factors for HCC worldwide. (Updated from [3], according to the International Agency for Research on Cancer (IARC) 2012 data [1]).

Geographic area	**HVC(%)**	**HBV(%)**	**Alcohol(%)**	**Others(%)**
Europe	60-70	10-15	20	10
America	50-60	20	20	10 (NASH)[1]
Asia and Africa	20	70	10	10 (aflatoxin)

2.1.1 Hepatitis

In simple terms, the word "hepatitis" means "liver inflammation". Hepatitis can be caused by bacteria, viruses, but also by the consumption of toxic substances (e.g. alcohol, certain drugs), and autoimmune diseases.

There are 5 main hepatitis viruses, referred to as types A, B, C, D and E. These viruses can be transmitted via contaminated water or food (hepatitis A and E), through contact with contaminated blood or infected body fluids (B, C and D) and also sexual contact (B and D). There are also autoimmune hepatitis, which are due to a disorder of the immune system. The body creates autoantibodies that attack the liver cells, rather than protecting them. However, viral hepatitis is the most common cause of hepatitis, and have become a matter of great concern in recent years due to its potential to become the largest current pandemic. Viral hepatitis can be acute or chronic. Acute hepatitis mostly heal themselves, but some can evolve to chronic hepatitis. In particular, hepatitis B and C are more likely to progress to chronic stages. Hepatitis is considered to be chronic if it is not healed after 6 months. They can lead to cirrhosis and, at later stages, to hepatocellular carcinoma.

2.1.1.1 Hepatitis B Virus (HBV)

HBV is usually transmitted via infected blood. It can be transmitted in medical and dental procedures where there are flaws in the sterilization process, by sharing needles or dirty syringes, unprotected intercourse and even saliva or other body fluids. HBV is only transmitted from human to human and it's more contagious than HIV or HCV.

Most individuals infected with HBV infection recover without realizing it. However, in less than 10 % of infected individuals, the immune system is unable to deal with the virus and the disease persists for more than 6 months, evolving to chronic hepatitis. Clinical manifestations and outcomes of HBV infection depend on the amount of virus present in the body and the strength of the body's immune system. The degree of virus activity can be determined by assessing the presence of certain viral components present in blood, the production of antibodies in response to these viral components and other clinical markers. Thus, the HBV serological tests involve the measurement of various antigens and specific antibodies of this virus. Antigens, as well as HBV-DNA, are parts of the virus, a sign that an individual is infected and can infect others. Antibodies are created by the immune system and their purpose is to "fight the virus". The major serological markers for HBV are:

- **Hepatitis B Surface Antigen** *(HBsAg)*: It is a part of the virus' surface. It appears between 2 and 6 months after infection and indicates that an individual has acute or

[1] Nonalcoholic Steatohepatitis

chronic hepatitis B. If HBsAg disappear and antibodies are produced (negative HBsAg and positive HBsAb), it is considered that the infection is healed.

- **Hepatitis B Surface Antibody** *(HBsAb)*: It is created by the immune system with the aim of destroying the virus. HBsAb is positive in the case of a ”cure” or in case of a successful vaccination against HBV. The HBs antibodies also make an individual immune to HBV, so that he/she can not be reinfected with the virus.

- **HBV-DNA** *(or viral load)*: It measures the virus's replication (virus production by the disease) and how infectious an individual really is. Some forms of hepatitis B produce only small quantities of virus in the body (low-replicative). Other forms of the disease produce the virus in very large amounts (high-replicative chronic hepatitis B). Low-replicative chronic hepatitis B is not usually associated with rapid disease progression. Most patients have normal results in liver function tests.

- **Hepatitis B Core Antibody** *(HBcAb)*: Similarly to HBsAb, HBcAb is produced by the immune system but its main objective is to destroy the core of HBV. When an individual is infected, HBcAb becomes positive and remains so forever, even if the infection is later cured or becomes chronic. However, HBcAb does not appear in healthy and vaccinated individuals. In brief, HBcAb allows to determine if the subject ever been (or is still) infected with HBV.

- **Hepatitis B e Antigen** *(HBeAg)*: HBeAg is an indirect marker of active virus replication. HBV-DNA is typically very high in case of a high-replicative hepatitis. However, there is always a vulnerable part of the virus, HBeAg. The immune system can create HBe antibodies to destroy it. This process does not qualify as a ”cure”, but means that the virus is being controlled by the body and is no longer able to replicate successfully.

- **Hepatitis B e-Antibody** *(HBeAb)*: This antibody is specialized in destroying HBeAg. It can ”sabotage” the virus' replication process and inhibit its growth during several years or even decades. Again, this situation is not considered a cure, but a body's control over the virus.

2.1.1.2 Hepatitis C Virus (HCV)

Similarly to HBV, hepatitis C virus (HCV) is generally spread by direct or indirect blood contact (parental transmission). It can also be spread by contaminated syringes or needles, as well as through open wounds, sharing razors or other sharp objects and toothbrushes. This virus can be transmitted in sexual contact, despite the risk of contracting the disease by infected subject's sexual partner is low. So far there is no record of transmissions through the skin (healthy) or saliva. Unlike hepatitis B, there is no vaccine for hepatitis C. HCV is considered a major public health problem by WHO [2], particularly dangerous for causing liver cirrhosis and hepatocellular carcinoma [10].

In most cases (60%-80% of subjects), the body's defences can not effectively resist the virus, and hepatitis C becomes chronic. However, in the other 20%-40% of cases, HCV is eradicated after 6 months from the onset of infection without treatment. HCV can be detected in the blood directly via its genetic information (RNA) or indirectly through the presence of antibodies formed by the patient's white blood cells. There are three main markers for this virus: HCV-RNA, HCV Core Antigen and HCV Antibody [12].

- **HCV Antibody** *(Anti-HCV)*: Determines if the person was ever exposed to HCV.

[2] World Health Organization

- **HCV-RNA**: HCV-RNA is a viral ribonucleic acid (RNA) which is created in the blood. Its presence is a reliable marker of active replication of HCV. In other words, it determines the amount of circulating virus in the body at the time of the test.

- **HCV Core Antigen** *(HCVAg)*: Detects the presence or absence of the virus.

2.1.2 Cirrhosis

In almost every study about hepatocellular carcinoma, cirrhosis is mentioned as its major risk factor. Overall, it is estimated that one third of patients with cirrhosis will develop HCC during their life time [3]. In chronic infections, hepatitis viruses increasingly damage the liver cells. The immune system responds to infection and white blood cells migrate to liver tissue, ensuring that dead liver cells are destroyed. Nevertheless, most of times they are unable to completely destroy the virus. Thus, dead liver cells keep accumulating and are later replaced by scar tissue. The spread of such tissue in the liver causes liver fibrosis and later on liver cirrhosis. This is quite a gradual process, but as more cells are damaged and die, with the formation of increasingly portions of scar tissue, the liver loses its ability to function normally.

There are several possible causes for cirrhosis. It can be induced by viral chronic hepatitis, abusive alcohol consumption, hereditary metabolic diseases such as hemochromatosis or Wilson's disease, and by Non-alcoholic fatty liver disease (NAFLD). NAFLD is a condition in which people who consume little or no alcohol develop a fatty liver, very common in obese people. NAFLD can be divided in the following stages:

Simple fatty liver (steatosis): "Steatosis" means "fatty liver". In this phase, excess fat build up in the liver cells, but is considered harmless. The accumulation of fat is relative small and does not lead to liver inflammation.

Non-alcoholic steatohepatitis (NASH): NASH is a more agressive form of NAFLD, where the liver has become inflamed, which suggests that the liver cells are being damaged and that some are dying. This stage is much more concerning that steatosis, since 20% of patients with NASH progress to cirrhosis.

Fibrosis: In this stage, persistent inflammation of the liver results in the generation of fibrous scar tissue around the liver cells and blood vessels. The scar tissue replaces some of the healthy liver tissue, though most of liver cells remain functioning normally.

Cirrhosis: This is the more severe stage, in which great parts of the liver present fibrosis. The liver shrinks and becomes lumpy, since regenerative nodules are formed to attempt to repair the damaged tissue.

The Child-Pugh (CP) score is used to assess the prognostic of chronic liver disease, such as cirrhosis. The score employs five clinical measures of liver disease : Total Bilirubin, Albumin, Encephalopathy, Ascites and Prothrombin Time or International Normalized Ratio (INR). Each one is scored from 1 to 3 points, with 3 indicating the most severe condition, as can can be seen in Table 2.2.

Table 2.2: Child-Pugh Classification for severity of cirrhosis.

Clinical and Lab Criteria	**1 point**	**2 points**	**3 points**
Encephalopathy	None	Grade I/II	Grade III/IV
Ascites	None	Moderate	Severe
Bilirubin (mg/dL)	< 2	2-3	> 3
Albumin (g/dL)	$> 3{,}5$	2,8-3,5	$< 2{,}8$
Prothrombin time **INR**	< 4 $< 1{,}7$	4-6 1,7-2,3	> 6 $> 2{,}3$

Bilirubin is the main product resulting of the destruction by the spleen of worn out or injured red blood cells. High levels of bilirubin in the blood might indicate the presence of some pathology which causes red blood cells destruction. On the other hand, billirubin may be in high levels because the liver is unable to eliminate it, causing its accumulation in the blood. Thus, bilirubin allows an evaluation of the overall status of liver function.

Albumin is the most abundant protein in the blood plasma, produced exclusively in the liver and extremely sensitive to liver disease. Its main function is to produce coagulation factors and its concentration decreases when the liver is injured. The analysis of the blood's coagulation level is made by assessing the time of prothrombin and is presented through a standardized measure known as INR (International Normalized Ratio). Basically, INR measures the speed of a particular pathway of coagulation, comparing it to the normal speed. If the INR is higher, it means that the blood is taking longer to clot than normal, and the synthesis of coagulation factors is being hindered. This is indicative of liver injury.

Ascites is the accumulation of fluid in the abdomen. This fluid may have different compositions, such as lymph, bile, pancreatic juice and others. At the context of liver diseases, ascites is the overflow of blood plasma to the interior of the abdominal cavity and indicates that the disease is advanced and related to the onset of other complications such as cirrhosis, the esophageal varices's bleeding or the encephalopathy.

Hepatic encephalopathy is a condition in which the brain function deteriorates due to the increase of toxic substances in the blood that should have been eliminated in the liver in a normal situation. Substances are absorbed across the intestine and they pass to the blood through the liver where the toxic ones are eliminated. In hepatic encephalopathy, this does not happen due to a decrease of the liver function. Thus, these toxic substances may reach the brain and affect its operation.

The evaluation of liver disease is made by adding the score of each criterion. According to this sum, the disease is assigned to one of three different classes: A (least severe liver disease), B (moderately severe liver disease), and C (most severe liver disease).

2.2 Staging System

Staging systems in HCC define the outcome prediction and treatment assignment, based in the main HCC prognostic variables: tumour stage (defined by number and size of the nodules, presence of vascular invasion, extrahepatic spread), liver function (defined by Child Pugh's class, bilirubin, albumin, portal hypertension, ascites) and performance status (general health-status, defined by ECOG [3] classification and presence of symptoms). The recommended staging system for HCC patients is BCLC [4] staging system [3]. Other systems applied alone or in combination

[3] Eastern Cooperative Oncology Group

[4] Barcelona-Clinic Liver Cancer

with BCLC are not recommended in clinical practice. The BCLC classification divides HCC patients in 5 stages (0, A, B, C and D), according to Performance Status (PS), Child-Pugh class, number and size of HCC nodules. The Performance Status evaluates how the disease affects the patient's daily activities (Table 2.3). Accordingly, HCC patients are staged as follows:

Very early HCC (stage 0) is defined as the presence of a single tumour < 2 cm of diameter without vascular invasion in patients with good health status (PS-0) and well-preserver liver function (Child-Pugh A class). Those who behave as carcinoma in situ are also defined as stage 0.

Early HCC (stage A) is defined in patients presenting single tumours >2 cm or nodules <3 cm of diameter, PS-0 and Child-Pugh class A or B.

Intermediate HCC (stage B) is defined in patients presenting multinodular asymptomatic tumours without an invasive pattern.

Advanced HCC (stage C) is present in patients with cancer related-symptoms (symptomatic tumours, PS 1-2), macrovascular invasion (either segmental or portal invasion) or extrahepatic spread (lymph node involvement or metastasis). The outcome varies according to the liver functional status (Child-Pugh A or B).

End-Stage HCC (stage D) patients have tutors leading to a very poor performance status (PS 3-4), similarly to Child-Pugh C patients.

Table 2.3: Performance Status Classification.

Performance Status Evaluation
Grade 0: Fully active, able to carry on all pre-disease performance without restriction.
Grade 1: Restricted in physically strenuous activity but ambulatory and able to carry out work of light or sedentary nature, e.g, light house work, office work.
Grade 2: Ambulatory and capable of all self-care but unable to carry out any work activities. Up and about more than 50% of waking hours.
Grade 3: Capable of only limited self-care, confined to bed or chair more than 50% of waking hours.
Grade 4: Completely disabled. Cannot carry on any self-care. Totally confined to bed or chair.
Grade 5: Dead.

2.3 Treatment Allocation

Treatment allocation is based on BCLC allocation system. Recommendations in terms of selection of different treatment strategies are based on evidence-based data in circumstances where all potential efficacious interventions are available.

2.3.1 Resection

Resection is the first-line treatment option for patients with solitary tumours and very well preserved liver function, defined as normal bilirubin with either hepatic venous pressure gradient $\leq$ 10 mmHg or platelet count $\geq$ 100 000. Tumour recurrence is the major complication of resection and influences the subsequent therapy allocation and outcome. In order to select the ideal candidates for resection, the assessment of liver function has moved from the gross determination of Child-Pugh class to a more sophisticated measurement of indocyanine green retention rate at 15 min (ICG15) or hepatic venous pressure gradient (HVPG) $\geq$ 10 mmHg as a direct measurement of relevant portal hypertension. Surrogate measures of portal hypertension include platelet count below 100 000 mm^{-3}, and it has been confirmed as an independent predictor of survival in resected HCC cases [3]. Anatomical resections are recommended and intraoperative US enables de detection of nodules between 0,5 and 1 cm and is considered the standard of care for discarding the presence of additional nodules and guide anatomical resections. The tumour extension, as said before, should be evaluated using last generation Computerized Tomography (CT) and Magnetic Resonance Imaging (MRI) scans. Considering the available information, the EASL-EORTC [5] panel does not recommend adjuvant interferon due to lack of significant patient number and partially conflicting data.

2.3.2 Liver Transplantation

Considered for patients with single tumours less than 5 cm and advanced liver dysfunction or tumours consisting in less than 3 nodules $\leq$ 3 cm (Milan criteria [3]) not suitable for resection. Patients within the Milan criteria while on the waiting list are treated with adjuvant therapies to prevent tumour progression. It is recommended to treat patients waiting for transplant with local ablation, and as a second choice with chemoembolization when waiting times are estimated to exceed 6 months. Extension of tumour limit criteria for liver transplantation has not been established. There is no clear upper limit for eligibility of downstaging. LDLT (Living Donor Liver Transplant) has associated risks of death and life-threatening complications for both donor and recipient and must be restricted to centers of excellence in hepatic surgery and transplantation. The policy adopted by the panel is that LDLT can be offered to patients with HCC if the waiting list exceeds 7 months.

2.3.3 Radiofrequency Ablation and Percutaneous Alcohol Injection

Local ablation with radiofrequency (RFA) or percutaneous ethanol injection (PEI) is considered for patients with BCLC 0-A tumours not suitable for surgery. The prime technique used is PEI, which induces coagulative necrosis of the lesion as a result of cellular dehydration, protein denaturation and chemical occlusion of small tumour vessels. RFA is the most widely assessed alternative to PEI for local ablation of HCC. The energy generated by RF ablation induces coagulative necrosis of the tumour producing a safety ring in the peritumoural tissue, which might eliminate small-undetected satellites. In tumours smaller than 5 cm, RFA is recommended as the main ablative therapy. PEI is recommended in cases where RFA is not feasible. In tumours $\leq$ 2 cm, BCLC 0, both techniques achieve complete responses in more than 90% of cases. Child-Pugh A patients are ideal candidates to RFA, but, at this point, there are no data to support RFA as a replacement of resection as the first-line treatment for patients with early HCC (BCLC A) stage.

2.3.4 Chemoembolization and transcatheter therapies

This procedure is recommended for patients with BCLC stage B, multinodular asymptomatic tumours without vascular invasion or extra hepatic spread. It is discouraged in patients with decompensated liver disease. Chemoembolization (TACE) is the most widely used primary

[5] European Association for the Study of the Liver - European Organisation for Research and Treatment of Cancer

treatment for unresectable HCC and the recommended first-line-therapy for patients at intermediate state of the disease.

2.3.5 Systemic therapies

Sorafenib [13] is the standard systemic therapy for HCC. It is indicated for patients with well-preserved liver function (Child-Pugh A) and with advanced tumours (BCLC C). There are no clinical or molecular biomarkers to identify the best response to Sorafenib, and there is no second-line treatment for patients with intolerance or failure to Sorafenib. In this setting, best supportive care or the inclusion in clinical trials is recommended. Patients at BCLC D should receive palliative support, but should not be considered for participating in clinical trials. HCC is recognized as among the most chemo-resistance tumour types, and Sorafenib emerged as the first effective treatment in HCC. It Is currently the standard-of-care for patients with advanced tumours. Other therapies, including chemotherapy, hormonal compounds, immunotherapy and several others showed inconclusive or negative results.

Figure 2.1 sumarizes the BCLC classification system and therapy allocation described in the previous sections.

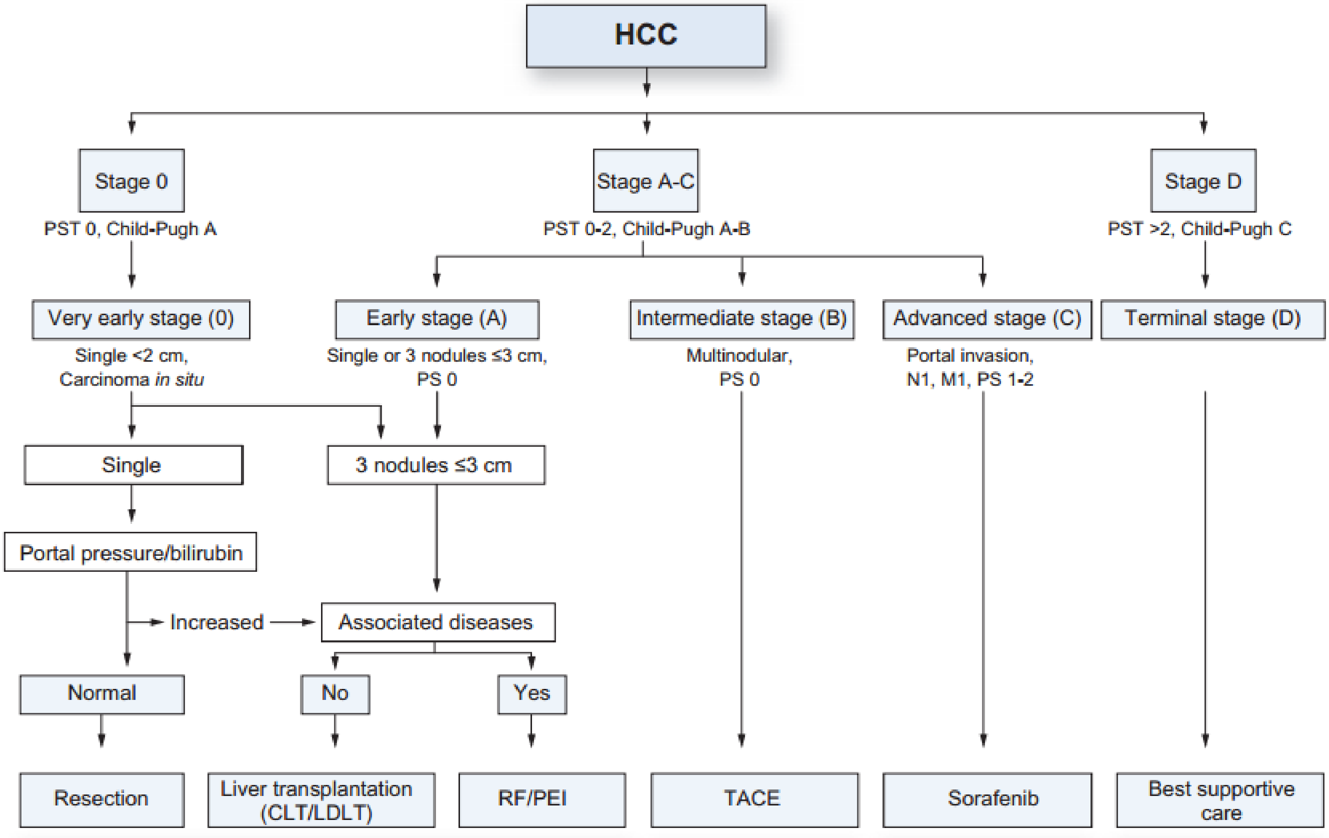


Figure 2.1: BCLC staging system and treatment strategy resume [3].

2.4 Conclusions

The main purpose of this chapter is to summarize the most recent medical evidence regarding HCC management. The study of HCC characterization, in terms of clinical variables, staging and allocation systems allowed us to define the requirements for the development of our CDSS, analysed in Chapter 5. Furthermore, in order to evaluate the results of the applied Artificial

Intelligence techniques (Chapter 6), one must be familiarized with the aspects of HCC discussed in this chapter.

Chapter 3

Clinical Decision Support Systems

In 1969, Goertzel introduced the concept of a Clinical Decision Support System (CDSS) as "a tool that assists in patient's clinical care, facilitating the acquisition of data and decision-making" [14]. Over the past four decades, many definitions have arisen. Musen defined a CDSS as "any software that processes information relating to a particular medical condition and produces inferences in the form of outputs that assist clinicians in their decision-making process, being considered a smart program on the part of their users" [15]. Miller and Geissbuher described a CDSS as "an algorithm to assist the clinician in one or more steps of the diagnostic process" [16]. Sim *et al.* consider that a CDSS "is a software developed with the aim of directly supporting the clinician in decision-making, in which the individual characteristics of a patient are compared with a computerized knowledge base so that it can make assessments and generate specific recommendations for that particular case, presenting them to the clinician or to the patient, as a basis for their decisions" [17].

Each one of these definitions reflect its authors' points of view, and thus can generate some discussion. However, regardless the definition that one considers more adequate, it is undisputed that all authors acknowledge the potential of such systems to provide benefits in healthcare quality and patients' healing process outcomes [18]. In our work, we will adopt Sim's definition. However, our system does not intend to generate recommendations to be presented to the patient. Our system intends to support only the clinician, in his daily activities.

3.1 Types of Clinical Decision Support Systems

Metzger *et al.* consider that CDSSs can be described according to their structure, behaviour and accessibility [19]. Regarding their structure, they differ in the timing at which they provide decision support: before, during or after the decision has been made. Concerning their behaviour, they are considered active or passive, according if the CDSS actively generates alerts and other warnings or only responds to the clinical inputs, respectively. According to their accessibility, they can provide general or specific/specialized information. Another categorization scheme of CDSSs is its differentiation into knowledge-based systems or non-knowledge-based systems. The majority of CDSSs are knowledge-based systems, composed essentially of three components: a knowledge base, the inference structure and the communication procedure [21]. The CDSSs lacking the first component (the knowledge base) are called "non-knowledge-based".

3.1.1 Knowledge-Based Systems

The knowledge-based systems are in some way similar to human reasoning. The knowledge base consists of a wide range of information about a particular domain, structured to be efficiently processed by the system. There are several schemes of information representation. Logical representation, where information is presented in the form of "if-then" statements, is the most frequently used and described in the literature. The efficiency of a CDSS depends on the quality of its knowledge base. The way it is exploited towards the development of rules for decision support is a major factor influencing the success of the recommendation system [15].

The "formulas" that combine these rules or associations constitute the second component of knowledge-based systems, the structure of inference. Essentially, these formulas involve the application of Artificial Intelligence (IA) techniques, able to analyse the existing information in the knowledge base and form new conclusions regarding a particular patient [21]. The inference mechanisms mentioned in the literature include the following [19]:

- **Rule-based reasoning**: These systems are based in "if-then" statements, which are seen as "standards". The inference engine seeks to associate the data under study with those known "standards". Rule-based systems "translate" the physicians' knowledge into expressions that can be evaluated as "rules". Therefore, they are often called "evidence-based systems" [22]. When acquired a considerable set of rules that support the knowledge base, the data under study are evaluated according to those rules (or their combination) until a conclusion is achieved. These type of systems are used for storing a large amount of information. However, its main disadvantage lies in the difficulty to translate the clinicians' experience and knowledge in simple and concrete rules.

- **Case-based reasoning**: These systems are mainly developed when it's not possible to model medical knowledge through formal methods of representation (such as Arden Syntax [20], for instance). The success of this approach is linked to the quality of the similarity metrics used to evaluate the existing cases and the efficiency of the methods chosen to discover and associate similar cases. Case-based reasoning are mostly used for subgroup analysis, and one of its great advantages is that analysis based in similar cases often produce more reliable and persuasive findings than the evidence-base medicine results. However, the assessment of similarity between cases may not prove to be a trivial process.

- **Model-based reasoning**: This method uses human pathophysiological models to define the dynamics of the body's biological processes. It is a promising and useful concept for application in CDSSs, frequently called "Patient Specific Modeling" [21]. The expected behaviour of a certain case according to these models is compared to the manifested behaviour. It is assumed that if the model is properly formulated, then the discrepancies between the predicted behaviour and the observed behaviour will not be significant. However, the major difficulty with this implementation arises when the validity of the model is not guaranteed. The more complex the system is, the more challenging it will be to design a model that accurately describes it [23].

- **Bayesian reasoning**: Bayesian Decision Theory is the core of these systems, establishing probabilistic relationships between the knowledge base's variables, for instance, symptoms and diseases, treatments and overall survival or medications and complications. These systems are based on Statistical Bayes classification, where a pattern is assigned to the most probable class, that is, the class with the maximum *a posteriori* probability. *A posteriori* probabilities are determined according to *a priori* probabilities, class conditional probabilities and Bayes rule. It is very useful to traduce disease progression over time or the relation between various diseases, assuming a cause-effect relationship between the variables under study. The main obstacle to its implementation is precisely the difficulty in specifying the cause, the effect and their relation in the clinical context, given its complexity [21].

- **Heuristic Reasoning**: Heuristics systems include statistical measures, and are used when there is no knowledge and/or computational resources to produce a "perfect answer". Heuristics methods reduce the problem's complexity; however, by definition, do not guarantee that the optimal solution is achieved. Heuristic methods are exploratory algorithms that seek to solve the problem, taking as a starting point a plausible solution and iterating through successive approximations aimed at an optimal solution. Commonly, the "best possible" solution is found, though not the "optimal solution". This approach may suggest a certain subjectivity or lack of precision. However, this is not necessarily a disadvantage, but a similar feature to human intelligence: we often use our personal experience to find solutions for everyday problems.

- **Semantic Networks**: A semantic network is a graphical way of representing knowledge, where the domain's concepts in question are represented by a set of "nodes" connected to each other through a set of arcs that describe the relationships between the existing nodes. The application of semantic networks in clinical inference is limited, since medical knowledge itself involves a plurality of concepts, making it particularly difficult to define a formal semantic framework able to translate it [15].

Finally, the communication mechanism is how information is entered into the system and the results (outputs) are returned to the user. In "stand-alone" systems, this information is often manually entered by the clinician. When CDSSs are integrated to other clinical management systems, the patient's information is incorporated in its electronic record, thus, containing data from several different services: laboratory, pharmacy or imaging. The output is then given to the physician in the form of recommendations and alerts [19].

3.1.2 Non-knowledge-based systems

Non-knowledge-based systems rely on machine learning techniques to produce useful inferences for decision making. Machine Learning is a branch of Artificial Intelligence that concerns the study and construction of systems that can learn from data. The system can learn from its past experiences and recognize patterns in clinical data. Artificial Neural Networks (ANN) and Genetic Algorithms (GA) are the most widely used approaches in the construction of such systems [19].

Neural networks are mathematical-computational models inspired by neuronal cells' functioning, simulating human reasoning, since they are a typical example of "example-based learning". Indeed, the structural units of ANN are called "neurons". A generic ANN model is composed by three layers: the input, output and processing layer (or hidden-layer). The input layer receives the data, while the output layer communicates the result. The hidden-layer is responsible for data processing and results' calculation. This type of structure has some similarities to knowledge-based systems, but in this case the knowledge-base is not derived from scientific literature nor clinical experience. ANN analyse existing patterns in the patient's information and derive associations between his input variables (symptoms, risk factors) and his output variables, for instance, his diagnosis or appropriate treatment strategy [22]. This is how the system "learns by example". The available information is studied and inferences are made about the most correct output for each input. These inferences are compared to the correct output (the targets, i.e., the actual results) and, based on the conclusion from these comparison, the system resets the associations between the input data and the previously determined output. This process continues iteratively until the correct result is achieved. Then, the system memorizes the model of such association between inputs and outputs in order to classify new cases. This iterative process is known as "training". The great advantage of this method is that it avoid the construction of "if-then" rules, and its definition by experts: as discussed, in medical contexts, these cause-effect rules may not be clearly defined *a priori*. Furthermore, ANN can more easily deal with missing data, because they can infer their values from the remaining set of complete data [22]. They also do not need a very large set of data to produce estimates, though the larger the "training" set is, the more accurate the results are. On the other hand, the training phase can be time consuming. However, the main disadvantage of this type of inference is its model's interpretation. This technique is often referred as a "black-box" inference model [15, 23], since the associations between data are complex and difficult to explain. For that reason, the use of these systems in the medical context is limited. Clinicians have the need to understand the mechanisms behind the system's recommendations. When these mechanisms become "less logical" and more complex, their confidence in system's responses considerably decreases [23].

GA are similar to ANN to the extent that derive their conclusions from patient's past information. GA are based in Darwin's Evolution's Theory, which explains the evolution of species through natural selection. As species evolve in order to adapt to their environment, GA also "reproduce" in various recombination in order to achieve the combination that best fits the data. When there is no specific knowledge about the domain under study, several sets of solutions are evaluated. The best sets (those that best fit the data) are then recombined ("mutated") to form the next set of possible solutions to be evaluated. The process continues iteratively until the optimal solution is reached. A "fitness function" determines which solutions should be kept and which should be eliminated [19]. The major difficulty here lies in the definition of "fitness", that is, what is considered a "good/poor" adjustment to data [15].

3.1.3 Clinical Decision Support System inference mechanism

As one can conclude from the above review, there is a wide range of available inference techniques for CDSSs development. Different inference engines have different advantages and disadvantages, and the appropriate choice of method (or combination of methods) for the implementation of an efficient inference engine, adequate to its application domain, is a delicate task. The main objective of a CDSS's inference engine is to analyse the data and "translate them" into useful conclusions. This process of data analysis is called "Data Mining". In literature, there are various applications of the discussed methods and data mining algorithms to distinct areas of Medicine [24]. In most cases, while studying a certain disease, various inference methods are used and their results compared. As an example, Soni *et al.* [25] compare Bayesian networks, case-based methods (Clustering algorithms) and rule-based methods (Decision Trees) and ANN applied to cardiovascular diseases diagnosis.

The selection of CDSS's type and a proper inference mechanism is dependent on the context of its application. Choosing a particular approach depends on the problem's domain, and on plenty other factors such as the cost of the system, the desired degree of efficiency and sensitivity and the amount of available data [22]. According to the "No Free Lunch Theorem" [26], no classification method is superior to all others in every context, i.e., there is no global "best classifier", superior to all others, whatever is the domain application under study. The selection of a CDSS inference model follows the "No Free Lunch Theorem", since the model itself is based on data mining techniques. This is the reason why various authors of the review articles mentioned above [18, 19, 22–24] suggest the evaluation of several inference techniques regarding the problem under consideration, in order to proceed with the selection of the most appropriate one. In conclusion, it's not possible to determine *a priori* the "greatest" CDSS type and inference model. This choice requires a thorough study of the domain in which the system is intended to operate, the type of data that will be analysed and the sort of recommendations that are intended to be generated.

3.2 Clinical Decision Support Systems in Healthcare

Information systems (ISs) in healthcare have been taking an increasing significance in the support provided to health professionals and patients themselves. In fact, the development of computerized systems for clinical data representation and management was instrumental in assisting the progress of clinical practice in recent decades [27]. Among the various applications of Information Technology to Healthcare are Clinical Information Systems (CISs) and Clinical Decision Support Systems (CDSSs).

The design and development of CISs is a key area of Medical Informatics and its main purpose is to improve the quality of health services, seeking to fulfil objectives such as allowing access to patient's information in all health facilities; return mechanisms for distributing and sharing that information among different health professionals; standardize clinical procedures and patient management services and also provide contextualized medical information to the patient himself, giving them personalized information about his health profile, clinical status and history [28]. Regarding the above objectives, a CIS should meet a set of requirements, through the registration and characterization of patients and their clinical information manipulation: management of medical consultations, integration of laboratory data in the context of diagnosis and therapy and statistical data of interest.

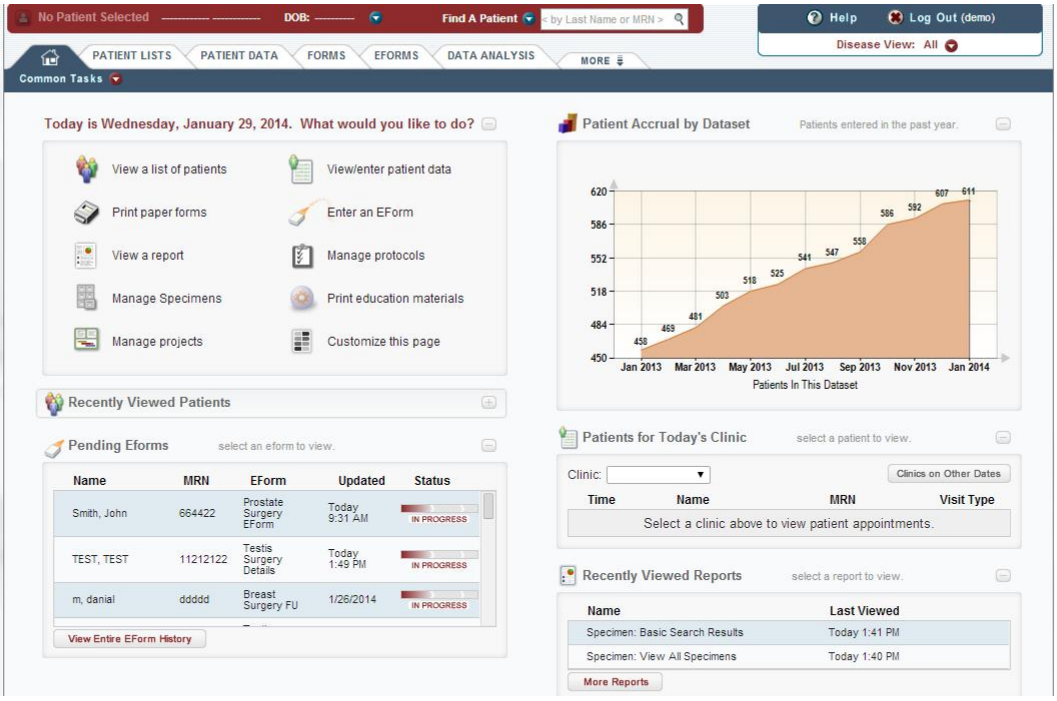
According to the World Health Organization (WHO), the amount of information in healthcare doubles every three years, affecting the clinical practice in various forms, with the emergence of new methods of diagnosis and therapy, innovations in the fields of molecular biology, genetics or chemistry and further studies on the effect of various drugs [29]. From this context arises the main motivation for the use of CDSSs. Taking advantage of computational resources, these systems have the ability to incorporate and represent an enormous amount of medical information and code selection strategies that produce useful responses to the process of decision making. According to this, a CDSS can be seen as a "information subsystem", associated with different medical specialities. They are developed in order to assist health professionals to make decisions that directly impact the patient's diagnosis or the management of processes that lead to diagnosis and thus their application, together with patient's contextual data, can help reduce the uncertainty associated to some clinical decisions. For instance, they may assist the physician in selecting the most suitable lab exam to validate a diagnosis, propose diagnostic or therapeutic strategies regarding a certain clinical condition and support the choice of the best treatment in order to control the progression of the disease, preventing unwanted drug interactions.

Health services involve a number of entities that need to share information to provide the best possible care to the patient. When an electronic record (EPR) is used to characterize the patient, it's necessary to consider the information flow related to the patient's follow-up. The process of decision making depends largely on how the patient's EPR is structured and properly updated. His medical record is of fundamental importance in the various steps of a medical decision, since that it consists in the knowledge base with which these actions will be taken. Thus, clinical activities, such as consultations, records of observations, diagnostic data, therapies and previous taken decisions must be duly registered in the CDSS in order to automate certain processes and define (and redefine) the system's learning and decision rules. Thus, we consider that there are two fundamental aspects in the development of a CDSS. On one hand, a good CIS that can collect, store and manage the access to healthcare information and patient's data - the knowledge base. On the other, the "introduction of intelligence" to the process, applying the knowledge base given by the CIS to build predictive models and decision rules to assist the clinician - the inference mechanism. In the following subsections, we will present some recent CISs and CDSSs used across several areas of Medicine.

3.2.1 Clinical Information Systems for sharing and managing clinical data

3.2.1.1 Caisis: Cancer Data Management

Caisis [30] is a web-application that combines cancer research with patient care (Figure 3.1). The main objective of this project, developed by BioDigital in 2002, is to improve the quality of cancer data so that they can be used in cancer research, while providing, in an organized and well structured way, every patient's history and relevant information, so that they can be managed by health professionals. Currently, Caisis is an academic application, mostly used as a tool to support research: patients' medical records are available for consultation and edition by the clinician, but they're part of a larger, standardized and "noiseless" dataset.



Today is Wednesday, January 29, 2014. What would you like to do?

Patient Accrual by Dataset — Patients entered in the past year.

Recently Viewed Patients

Pending Eforms — select an eform to view.

Name	MRN	EForm	Updated	Status
Smith, John	664422	Prostate Surgery EForm	Today 9:31 AM	IN PROGRESS
TEST, TEST	11212122	Testis Surgery Details	Today 1:49 PM	IN PROGRESS
m, danial	ddddd	Breast Surgery FU	1/26/2014	IN PROGRESS

Patients for Today's Clinic — select a patient to view.

Time	Name	MRN	Visit Type
Select a clinic above to view patient appointments.			

Recently Viewed Reports — select a report to view.

Name	Last Viewed
Specimen: Basic Search Results	Today 1:41 PM
Specimen: View All Specimens	Today 1:40 PM

Figure 3.1: Caisis Interface [30].

Caisis is open-source, runs on .NET Framework and is mostly written in C #, HTML and JavaScript. The requirements for the server include Windows Server 2000 or later, IIS 6 or later and the Microsoft .NET Framework 3.5 or 4.0. The used database is Microsoft's SQL Server 2008++. The client needs only to install one of Caisis current version's supported browsers: Internet Explorer 7+, Firefox, Safari 3+, or Chrome 12 +. Caisis is free to download and install under an open-source user license, referred to as the General Public License (GPL). GPL allows the user to download the application files and all source code, modify and distribute it, provided that such changes are shared with BioDigital and redistributed with the GPL.

The main features of this system are resumed in Table 3.1:

Table 3.1: Main Caisis features [30].

Patient Lists Allows the user to browse by patient groups (by last name, current status, referring physician) and find a particular patient.
Patient Data The user can enter and view the patient's clinical information.
Forms Printing paper forms, blank or filled, with patient's information.
E-forms Electronic forms allow computerized data entry.
Data Analysis Enables data exportation (Access or Excel format) by type of illness, level of privacy or objective. Also allows the user to select datasets for research and access to reports, clinical trials and other studies already conducted.

3.2.1.2 DOCgastro: A Clinical Information System for Gastroenterology

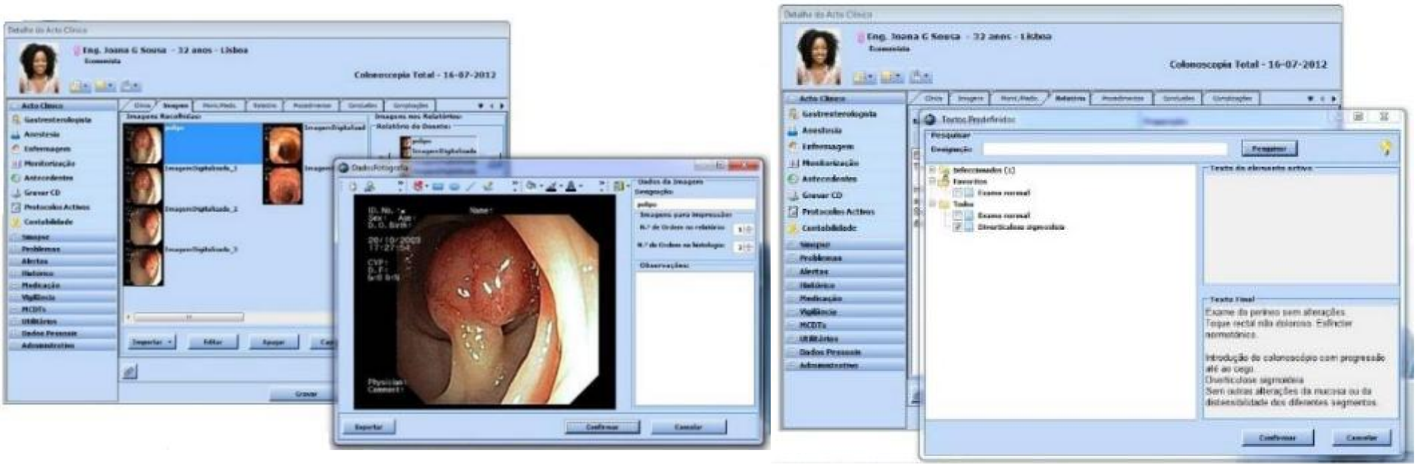
DOCgastro is currently implemented in North Lisbon Hospital Centre (NLHC). DOCgastro (Figure 3.2) is an Integrated Gastroenterology System, developed by Mobilware [1]. It was specially designed for Gastroenterology for gathering and storing information concerning this speciality exams. Allows video or photography capture during the exam, image editing and archiving the patient's record and its integration in the procedure reports. Such reports can be set previously in the system, in text or timely topics and changed if necessary. In addition to clinical information, DOCgastro ensures a complete record of the proceedings and consumables for proper accounting of resources. The application also allows the user to query specific tables for clinical procedure, conduct research and statistics on the database, the scheduling of examinations and their billing. DOCgastro can also be integrated with other hospital information systems such as hospital management systems, laboratory and pharmacy applications and Picture Archiving and Communication System (PACS).

3.2.2 Clinical Decision Support Systems and Nomograms used in Healthcare

3.2.2.1 MyRisk: Support System for Cancer Diagnosis

MyRisk prototype [32], developed at the Polytechnic Institute of Castelo Branco, Portugal, is a CDSS used to calculate cancer risk for each individual patient. Its graphical interface (Figure 3.3) is very intuitive and simple, where the user can get expert information about pathological characteristics, risk factors and behaviours associated with certain cancers, namely breast cancer, skin cancer and uterine cancer. The application also provides specialized warnings,

[1] www.mobilware.com



Exame do períneo sem alterações. Toque rectal não doloroso. Esfíncter normotónico.

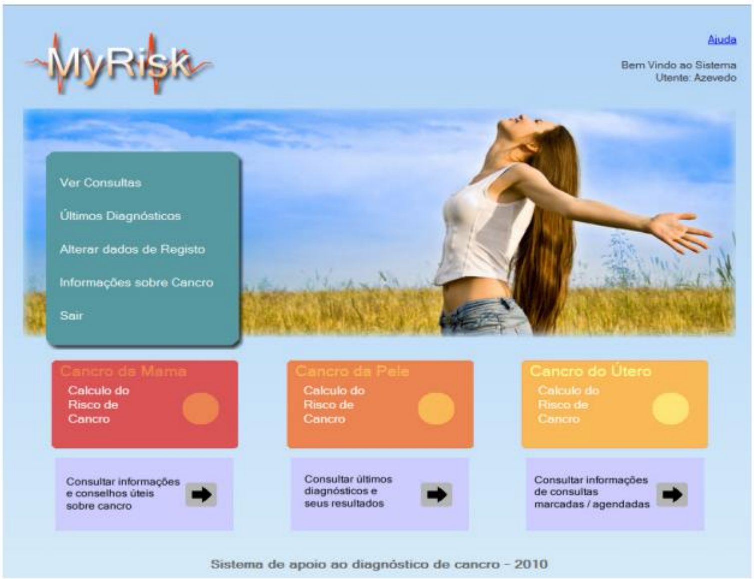
Introdução do colonoscópio com progressão até ao cego.
Diverticulose sigmoideia
Sem outras alterações da mucosa ou da distensibilidade dos diferentes segmentos.

(a)

(b)

Figure 3.2: DOCgastro's Interface for exam registration (a) and patient's information management (b).

according to each disease and type of risk, giving some information about necessary procedures for appointments or recommendations to adopt.



Bem Vindo ao Sistema
Utente: Azevedo

Consultar informações e conselhos úteis sobre cancro

Consultar últimos diagnósticos e seus resultados

Consultar informações de consultas marcadas / agendadas

Sistema de apoio ao diagnóstico de cancro - 2010

Figure 3.3: MyRisk Interface [32].

The application has three access levels: two for users (registered or unregistered) and a third for administrators and health professionals (physicians). The system's functionalities are described in Table 3.2.

Table 3.2: MyRisk main features [32].

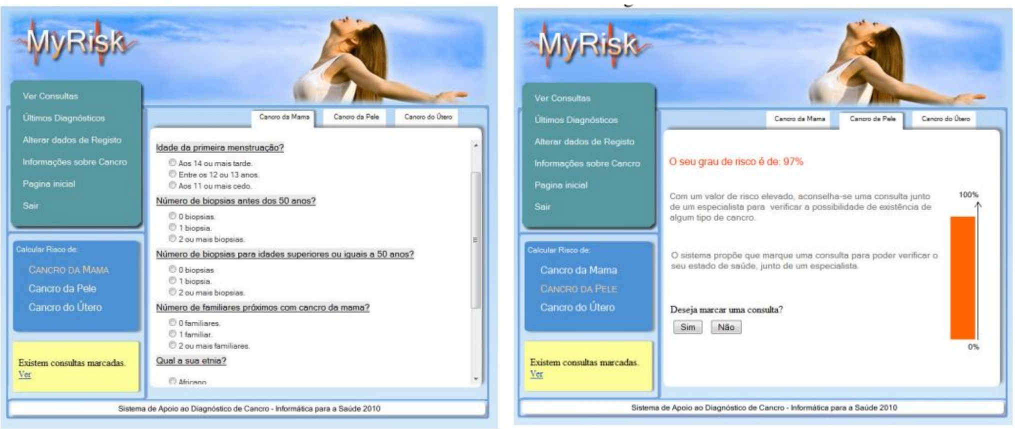
Unregistered Users	Consultation of useful information concerning the three types of cancer; To take advantage of other features, the user **must be registered**;

Registered Users	Personal Information Management;
	Calculation of cancer risk;
	Book appointments;
	Query answered questionnaires;

Physicians	Consultation of appointments' schedules;
	Definition of pathologies evaluation's parameters;
	Appointments' management;
	Conducted diagnosis consultation;
	Definition of cancer risk degree that implies an appointment's suggestion;

Administrators	Users Management Management of information and useful tips about cancer; Questionnaires management;

The calculation of cancer risk is based on filling a form prepared for this purpose. Each form (Figures 3.4 and 3.5) is composed of a set of questions. These issues can be changed depending on the considerations of the physician face to advances in investigations of the different types of cancer. Each question is associated to a 0-100% percentage, depending on the totality of survey questions and the degree of importance given by health professionals to each one of them. Likewise, for each question, the answers have an associated percentage that depends also on the number of possible responses and their degree of importance. Based on the percentage of each question and response, the cancer risk is calculated: Low (i), Medium (ii) and High (iii). The physician can change the percentages corresponding to each level and also the minimum percentage suggestive of an appointment.



Idade da primeira menstruação?
○ Aos 14 ou mais tarde.
○ Entre os 12 ou 13 anos.
○ Aos 11 ou mais cedo.

Número de biopsias antes dos 50 anos?
○ 0 biopsias.
○ 1 biopsia.
○ 2 ou mais biopsias.

Número de biopsias para idades superiores ou iguais a 50 anos?
○ 0 biopsias
○ 1 biopsia.
○ 2 ou mais biopsias.

Número de familiares próximos com cancro da mama?
○ 0 familiares.
○ 1 familiar.
○ 2 ou mais familiares.

Qual a sua etnia?
○ Africano

Existem consultas marcadas.

Sistema de Apoio ao Diagnóstico de Cancro - Informática para a Saúde 2010

O seu grau de risco é de: 97%

Com um valor de risco elevado, aconselha-se uma consulta junto de um especialista para verificar a possibilidade de existência de algum tipo de cancro.

O sistema propõe que marque uma consulta para poder verificar o seu estado de saúde, junto de um especialista.

Deseja marcar uma consulta?

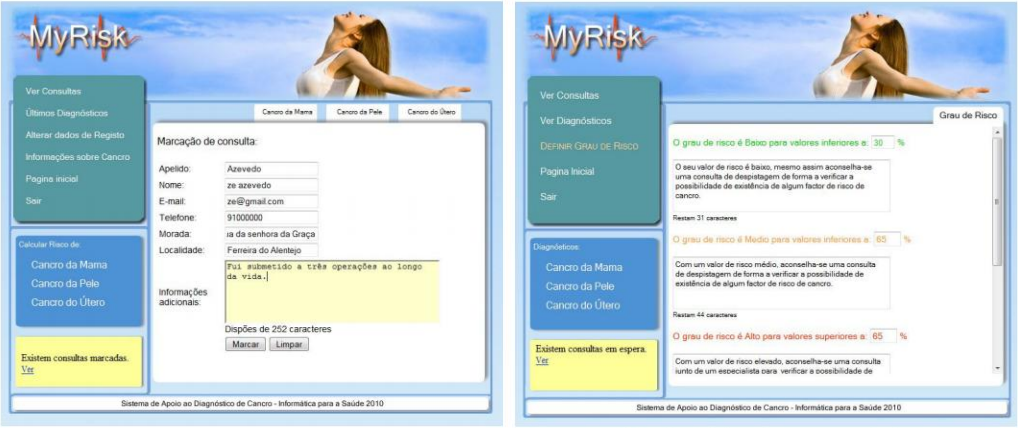
Existem consultas marcadas.

Sistema de Apoio ao Diagnóstico de Cancro - Informática para a Saúde 2010

(a) (b)

Figure 3.4: Example of a form (a) and cancer risk percentage calculation (b) [32].

This prototype was developed using exclusively open-source tools, namely PHP and MySQL for the business logic and data storage, respectively, and HTML, CSS and JavaScript to design the user interface.



Marcação de consulta:

Apelido: Azevedo
Nome: ze azevedo
E-mail: ze@gmail.com
Telefone: 91000000
Morada: ıa da senhora da Graça
Localidade: Ferreira do Alentejo
Informações adicionais: Fui submetido a três operações ao longo da vida.

Dispões de 252 caracteres

Existem consultas marcadas. Ver

Sistema de Apoio ao Diagnóstico de Cancro - Informática para a Saúde 2010

(a)

O grau de risco é Baixo para valores inferiores a: 30 %

O seu valor de risco é baixo, mesmo assim aconselha-se uma consulta de despistagem de forma a verificar a possibilidade de existência de algum factor de risco de cancro.

Restam 31 caracteres

O grau de risco é Medio para valores inferiores a: 65 %

Com um valor de risco médio, aconselha-se uma consulta de despistagem de forma a verificar a possibilidade de existência de algum factor de risco de cancro.

Restam 44 caracteres

O grau de risco é Alto para valores superiores a: 65 %

Com um valor de risco elevado, aconselha-se uma consulta junto de um especialista para verificar a possibilidade de

Existem consultas em espera. Ver

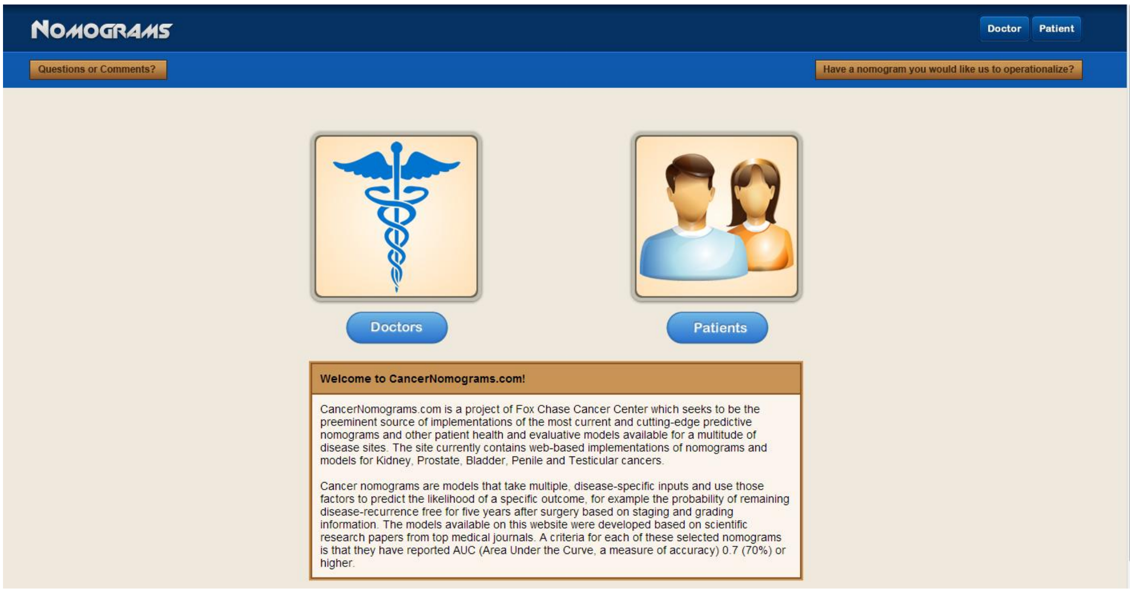
Sistema de Apoio ao Diagnóstico de Cancro - Informática para a Saúde 2010

(b)

Figure 3.5: Appointment's form (a) and definition of cancer risk percentages (b) [32].

3.2.2.2 CancerNomograms.com

The CancerNomograms.com [33] is a project developed by Fox Chase Cancer Center, which currently includes nomograms' web-applications for kidney, prostate and bladder cancer (Figure 3.6). The implemented predictive models were developed based on published scientific articles in prestigious medical journals. The criteria used for selecting the used algorithms was an Area Under the Curve (AUC) of 0.7 or higher.



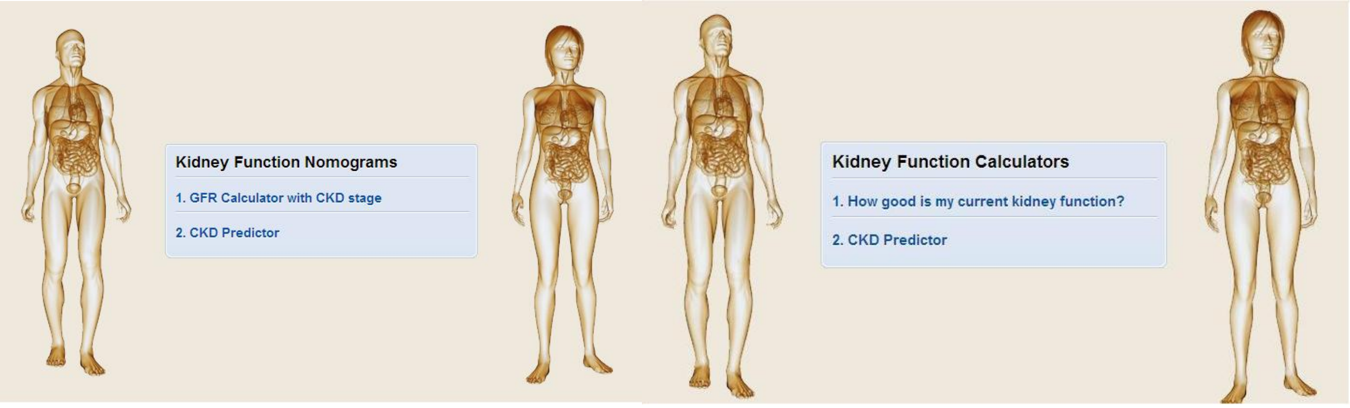
Welcome to CancerNomograms.com!

CancerNomograms.com is a project of Fox Chase Cancer Center which seeks to be the preeminent source of implementations of the most current and cutting-edge predictive nomograms and other patient health and evaluative models available for a multitude of disease sites. The site currently contains web-based implementations of nomograms and models for Kidney, Prostate, Bladder, Penile and Testicular cancers.

Cancer nomograms are models that take multiple, disease-specific inputs and use those factors to predict the likelihood of a specific outcome, for example the probability of remaining disease-recurrence free for five years after surgery based on staging and grading information. The models available on this website were developed based on scientific research papers from top medical journals. A criteria for each of these selected nomograms is that they have reported AUC (Area Under the Curve, a measure of accuracy) 0.7 (70%) or higher.

Figure 3.6: CancerNomograms interface [33].

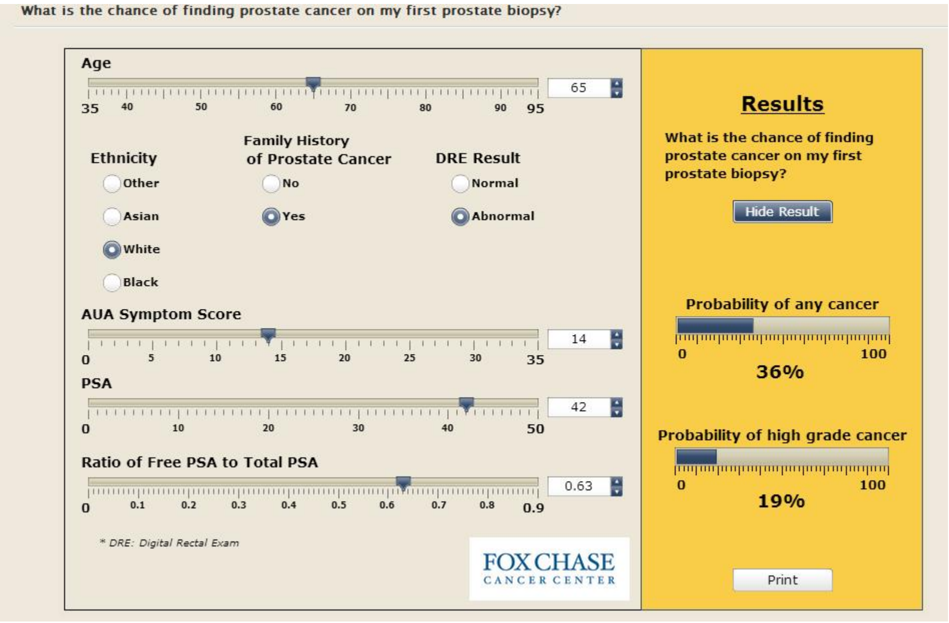
The application provides two access levels: for physicians and patients. However, the available information to each is exactly the same, the only thing that changes is the forms submission's format. For the physician, the menus (Figure 3.7a) are written in a more formal way, with acronyms and familiar clinical concepts. For the patient, the menus (Figure 3.7b) are adapted so that the actions become intuitive and understandable to a layperson in the clinical context. In most cases, the forms are presented through suggestive questions, in a way that is easier for the patient to select the information he wants, choosing the questionnaire for which he wants to know the results, or, in other words, "choosing the question that he wants to see answered."


Kidney Function Nomograms
1. GFR Calculator with CKD stage
2. CKD Predictor
Kidney Function Calculators
1. How good is my current kidney function?
2. CKD Predictor

(a) (b)

Figure 3.7: CancerNomograms - doctor's menu (a) and patient's menu (b) for kidney cancer nomograms (Kidney Cancer Predictive Tools) [33].

The variables' collection to evaluate the nomogram is done using a simple form (Figure 3.8). The answers to each question are predefined, so filling out the form is done by selecting the appropriate answer to each patient's condition. The risk calculation result is then returned on a scale of 0 to 100%.



What is the chance of finding prostate cancer on my first prostate biopsy?

Age: 65

Ethnicity: ○ Other ○ Asian ● White ○ Black

Family History of Prostate Cancer: ○ No ● Yes

DRE Result: ○ Normal ● Abnormal

AUA Symptom Score: 14

PSA: 42

Ratio of Free PSA to Total PSA: 0.63

* DRE: Digital Rectal Exam

Results

What is the chance of finding prostate cancer on my first prostate biopsy?

Hide Result

Probability of any cancer: 36%

Probability of high grade cancer: 19%

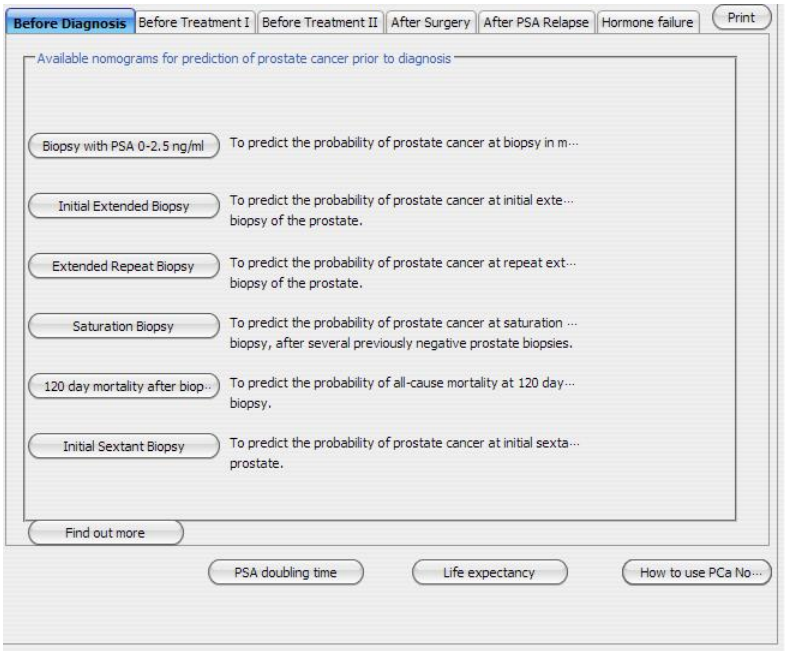
Print

Figure 3.8: CancerNomograms - Form and results for prostate cancer risk [33].

3.2.2.3 Nomogram.org

Nomogram.org [34], developed by Cancer Prognostics and Health Outcomes Unit, University of Montreal, offers nomograms to assist clinicians and patients based on personalized information. Its main objective is to facilitate the process of decision making by both assisting the physician in choosing the best diagnostic and therapeutic methods and offering the patient reliable information, enabling him to form a reasoned opinion about his treatment's options. Up to date, there are nomograms for prostate, kidney, bladder, greater urinary tract, penis and adrenal cancer. The interface does not distinguish between users, whether they are healthcare professionals or patients.

Prostate cancer's nomogram (Figure 3.9) was the first to be developed hence is the most complete. The pathology's related risks can be calculated from the pre-diagnosis to a more advanced stage of the disease, also accounting for intermediate stages. Accordingly, the physician (or patient) may query the application for predictions at any step of the treatment (Figure 3.12).



Available nomograms for prediction of prostate cancer prior to diagnosis

To predict the probability of prostate cancer at biopsy in m…

To predict the probability of prostate cancer at initial exte… biopsy of the prostate.

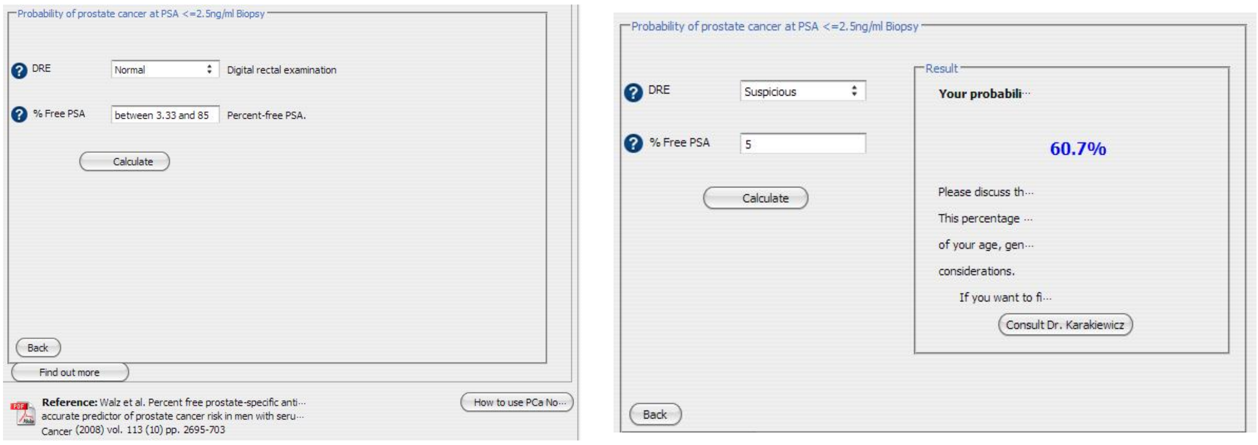
To predict the probability of prostate cancer at repeat ext… biopsy of the prostate.

To predict the probability of prostate cancer at saturation … biopsy, after several previously negative prostate biopsies.

To predict the probability of all-cause mortality at 120 day… biopsy.

To predict the probability of prostate cancer at initial sexta… prostate.

Figure 3.9: Nomogram.org - prostate cancer related nomograms [34].



Probability of prostate cancer at PSA <=2.5ng/ml Biopsy

DRE: Normal — Digital rectal examination

% Free PSA: between 3.33 and 85 — Percent-free PSA.

Reference: Walz et al. Percent free prostate-specific anti… accurate predictor of prostate cancer risk in men with seru… Cancer (2008) vol. 113 (10) pp. 2695-703

(a)

Probability of prostate cancer at PSA <=2.5ng/ml Biopsy

DRE: Suspicious

% Free PSA: 5

Result

Your probabili…

60.7%

Please discuss th…
This percentage …
of your age, gen…
considerations.
If you want to fi…

(b)

Figure 3.10: Nomogram.org - Form to calculate the probability of prostate cancer risk (a) and the nomogram's results (b) [34].

Table A.1 (Appendix A) summarizes the main characteristics of each application presented in Sections 3.2.1 and 3.2.2.

3.2.3 Clinical Decision Support Systems and Nomograms applied to Gastroenterology

3.2.3.1 Leeds Abdominal Pain

The first successful CDSS in gastroenterology was developed in the late 60's, specifically applied to the diagnosis of acute abdominal pain: Leeds Abdominal Pain System. It became operational in 1971 at Leed's University Hospital, UK, achieving high rates of success in real time diagnosis of seven different pathologies: appendicitis, diverticulitis, perforated ulcer, colitis, small bowel obstruction, pancreatitis and unspecified abdominal pain. The system was based on the communication between a KDF9 English Electric computer, located in the Computer Laboratory of Electronics, University of Leeds, and a Westrex 33 ASR terminal located in the Department of Surgery of the University Hospital in Leeds, about 800 meters. The system's creators wrote a FORTRAN program which integrated Bayes's Probability Theory, and based on previously entered patient's data, generated the "diagnosis" for a new patient.

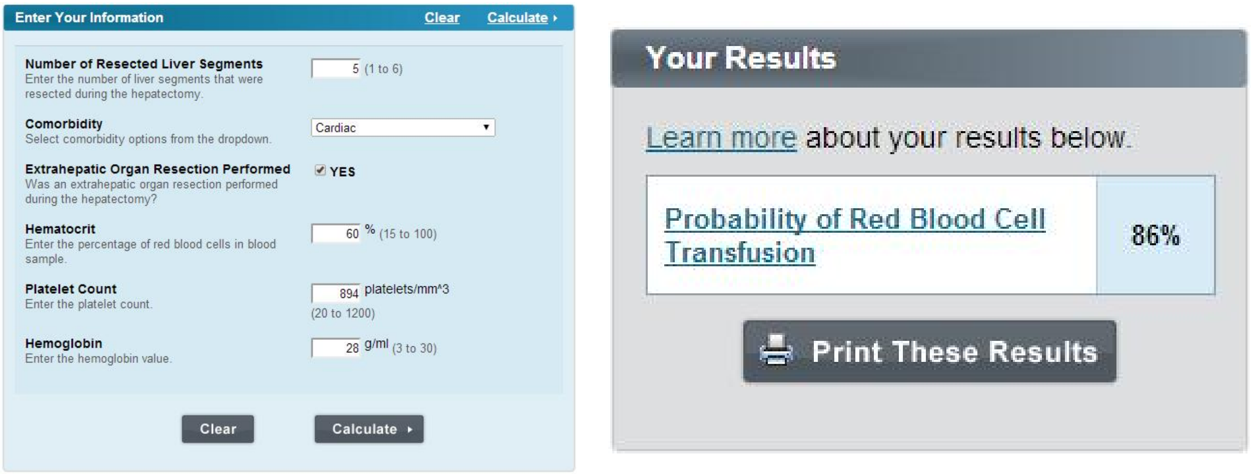
The collection of clinical data was done by filling in a form created for the purpose. This introduced some noise into the system to the extent that it was unavoidably subject to the "inter-observer variance", i.e., to differences in completing the questionnaire from physician to physician. The authors sought to minimize the influence of this factor through the use of training on patient's clinical information registration for clinicians. Instead of inserting all the hand-written patient's clinical history, each patient's variables (sex, age, pain location, among others), were represented by 3-digit codes, reducing the computational burden of later analysis. The use of such simple codes also allowed the data entry by some family member or other person with access to those codes. Therefore, the clinician is not required to have any direct contact with the computer or even with the terminal. In fact, once the the form is complete, no one needs to access the system until the diagnosis is achieved and returned by the terminal.

Ideally, given a certain set of clinical data, the computer would return a diagnosis based on the known characteristics of various diseases. Unfortunately, as evidenced in this study, it is necessary to assign each patient to a particular category. Thus, the systems selects the "database" related to the group where the patient falls, stored in disk. Then, a Bayesian analysis is computed and the resulting probabilities are stored. The response algorithm takes into account the request made to the system. It examines all cases that can be used in the analysis and when there are no more cases to include, the results are presented. The achieved diagnosis can also be compared to the one made by the clinician. If they do not match, the system selects patient's informations that may be responsible for the discrepancy, and presents them as a suggestion for further verification. If the probabilities returned by Bayes analysis are unsatisfactory (the results' accuracy is not enough to confidently "ensure" any of the considered diagnosis), the system suggests a list of rare diseases, which can help the clinician in less common cases.

This system does not make any recommendations concerning treatment strategies, its "responsibility" is only limited to (a) return the diagnosis probabilities for a set of pre-established diseases and (b) recommend, if necessary, the acquisition of additional information. The same team of researchers conducted a study from Jan 1st to Dec 1st (1971) seeking to compare the diagnostic efficacy with and without the use of their system. The accuracy rate obtained by this CDSS reached 91,8%, considering a total of 304 cases examined during this period, a value much higher than the rate of correct diagnoses mentioned by doctors, ranging between 65% and 80%.

3.2.3.2 Memorial Sloan Kettering Cancer Center Prediction Tools for Cancer Care

Researchers from Memorial Sloan Kettering Cancer Center have been pioneers in the development of nomograms for predicting the risk of cancer and treatment outcomes. The evaluation of these parameters is done according to the patient's characteristics and pathology. The nomograms available online include bladder, gastrointestinal tract, breast, colorectal, endometrial, melanoma, ovarian, prostate, renal, pancreatic, thyroid, sarcoma, uterine leiomyosarcoma, lung and liver cancer. In the particular case of liver cancer, the nomogram is used to predict the need for red blood cells transfusion before, during or after an hepatectomy - a surgical procedure in which part of the liver is removed. The test's results allow the physician a better monitoring and guidance of his patient (Figure 3.11).



Enter Your Information

Number of Resected Liver Segments
Enter the number of liver segments that were resected during the hepatectomy.
5 (1 to 6)

Comorbidity
Select comorbidity options from the dropdown.
Cardiac ▾

Extrahepatic Organ Resection Performed
Was an extrahepatic organ resection performed during the hepatectomy?
☑ YES

Hematocrit
Enter the percentage of red blood cells in blood sample.
60 % (15 to 100)

Platelet Count
Enter the platelet count.
894 platelets/mm^3 (20 to 1200)

Hemoglobin
Enter the hemoglobin value.
28 g/ml (3 to 30)

Your Results

Learn more about your results below.

Probability of Red Blood Cell Transfusion	86%

(a)

(b)

Figure 3.11: Liver Cancer Nomogram - Form that assesses the need for blood transfusion (a) and results presented by the system (b).

3.2.3.3 Other Clinical Decision Support Systems applied to Gastroenterology

In the original article by Horrocks *et al.* [35], some important questions concerning CDSSs in gastroenterology arose: are they really useful for physicians? Can they offer a measurable advantage in diagnostic/therapeutic decision?

Seeking to answers these questions, a review article published in the Journal of Health Informatics sought to describe the most recent experiences regarding the implementation of CDSSs in gastroenterology, in order to establish the level of development, testing and advantages in medical practice associated to the introduction of these software [36]. In this paper, CDSSs are evaluated according to the following parameters: concerned clinical issue/disease to which the CDSS is applied, system's architecture, integrated Artificial Intelligence (IA) tools, sizes of the used samples (number of clinical cases), achieved results, comparison of such results with the expert reviews, user feedback, evidence of improvement in clinical practice and encountered critical problems. After an exhaustive search in PubMed, LILACS [2] and ISI Web of Knowledge databases, 9 of 104 publications were selected. Excluded articles did not meet the inclusion criteria: to be a computerized CDSS in gastroenterology and provide the full text.

The study conducted by Das *et al.* [37] consisted in the development and validation of an experimental model to predict the need for an endoscopic treatment. The study by Chu *et*

[2] Literatura Latinoamericana y del Caribe en Ciencias de la Salude

al. [38] was based on the development of a predictive model to determine the source of bleeding, need for blood transfusion, urgent endoscopy or predisposition to acute gastrointestinal bleeding, with the aim of assisting clinical practice in an emergency situation. Berner *et al.* [39] created a recommendation system for safe medication prescribing. Farion *et al.* [40] described a CDSS for patients' triage, through their clinical history's analysis, physical examination and laboratory tests, using notebook computers. Sadeghi *et al.* [41] developed a system based on a Bayesian network for the purpose of automating the screening of patients with non-traumatic abdominal pain. Lin [42] divided his project into two phases: the first with the aim of distinguishing between healthy individuals and individuals with liver disease; the second to identify the pathology within the group of sick individuals. Finally, Aruna *et al.* [43] designed a system for gastrointestinal disorders's diagnosis, DIAGNET.

Table A.2 (Appendix A) summarizes the main characteristics of each CDSS described in these articles, according the outlined parameters referred above.

3.2.4 Clinical Decision Support Systems for Hepatocellular Carcinoma

3.2.4.1 Information Technology Systems in Personalized Medicine A clinical use-case for Hepatocellular Carcinoma

In the work [44], the authors seek to understand how the current evidence present in guidelines, clinical practice and the requirements of a Personalized Medicine based solution can be conciliated with the development of an information management and recommendation system, regarding the particular case of HCC. The authors propose to identify the factors that reflect the patient's clinical condition as well as relating them to the tumour's nature, individual patient response and results of therapeutic strategies. All these variables (which are given the name of ”Information Entities” - IEs) would then be used for general ”Digital Patient Models” (DPMs), customized models for each patient, through MultiEntity Bayesian Networks - (MEBNs). According to the authors, this structure of standard clinical information of a HCC patient, together with structured information about the disease itself and the several clinical approaches, would enable the creation of a statistical model, able to produce reliable diagnosis, prognosis and personalized treatment's recommendations. This model could then be used to build a decision support system, to which the authors call MBME - Model-Based Medical Evidence.

Until today, this system is no more than a proposal. The authors have reviewed the literature regarding HCC's epidemiology, etiology, risk factors, biomarkers, and therapeutic strategies, identifying the essential IEs, and trying some MEBNs for data mining and decision support. However, these algorithms are not presented nor described in [44]. Furthermore, their results are not clear. The authors attempt to justify these flaws through the lack of available information, identifying the need for more clinical cases to develop a larger amount of models, and more detailed ones, in order to validate the criteria used in the algorithms' modelling. However, in their opinion, it is very clear that the understanding, prevention and treatment of HCC will benefit from the construction of such a recommendation system that emphasizes the patient's individual characteristics and his personal medical history, providing a new paradigm of Evidence Based Medicine: the use of specific models for patients individuals, i.e., subgroup analysis [44].

3.2.4.2 A database for cirrhotic patients for early detection of Hepatocellular Carcinoma

Cirrhosis is present in over 80% of HCC cases, being clearly identified as the main precursor lesion of this pathology. In this study, the authors address the main features of e-Hepar III, a support tool for the diagnosis of liver disorders [45]. This system is integrated with a database of 200 patients. Each clinical case is described using 170 variables, such as patient's demographics, physical examination's results, laboratory tests, and histopathological diagnosis. e-Hepar III provides a set of statistical methods that enables data analysis regarding patients' diagnosis and prognosis, assessing liver cirrhosis evolution.

The support rules for diagnosis and prognosis are based on diagnostic maps, case-based reasoning and regression models. Each patient has multivariate data, that is, each clinical case is described by a set of variables that compose multidimensional patterns. In diagnostic maps, these variables have to be transformed so that they can be represented in only two dimensions. This allows the "translation" of each clinical case as a point and the representation of all patients as "points" on graph. A symbol is assigned to each disease and thus these "diagnostic maps" show all points (patients) represented by symbols according to their pathology. In this way, the differences between the various diseases are visually highlighted. In the authors' opinion, this graphical representation is important since it allows the clinicians to better understand the processes that lead the system to generate recommendations based on patients' characteristics, and thus increasing their interest and involvement in this "assisted decision-making process". Rather than simply receiving a response from the system, the clinician can understand the response's underlying reasons. It is the case-based reasoning that enables decision support in selecting diagnostic and therapeutic strategies. The system uses information regarding past experiences (similar cases) to solve a new decision problem. e-Hepar's regression models are used to find patients at high risk of liver cancer, indicating its prognosis based on the evolution of the disease. This paper describes in slightly shallow way a data mining tool that identifies common patterns in the collected data and uses them in the decision-making process.

The authors express their interest in publishing more details about the system and its performance in terms of accuracy in the early diagnosis of HCC in patients with cirrhosis but so far they do not describe the algorithms/techniques used for assessing the similarity between cases, nor the regression models used. Furthermore, initially there were only 2 out of 200 cirrhosis patients with an HCC diagnosis. This number rose to 10 in the two-year follow-up that followed. As seems clear to us, these numbers are not sufficient to validate the system's performance. Any preliminary results of the system would be inconclusive, so the added value of this study relates to the most interesting variables selected to define each patient's clinical condition.

3.2.4.3 Disease-Free Survival after hepatic resection in Hepatocellular Carcinoma patients

Ho *et al.* attempted to establish a model to describe free survival disease at 1, 3 and 5 years after hepatic resection in a study population of 482 patients with HCC [46]. Three prediction models were tested: ANNs, logistic regression and decision trees. According to the authors, the conclusions driven from a comparison between different models may help in the selection of the best method to be integrated into a CDSS for this pathology. The existing patients in the database constructed for this study were divided into 3 groups according to their disease-free survival. In each group, patients were labelled as disease-free hepatic resection survivors if no death or recurrence occurred during the period considered in the three survival models (1, 3 or 5-years). The selected clinical cases were reviewed, in order to collect information concerning each patient demographics, risk factors, clinical variables regarding laboratory tests, tumour stage and others associated with the results after resection and with the surgical procedure itself.

After collecting the data, the variables suffered some transformations before the models could be developed. In particular, continuous variables were categorized to minimize the effects of extreme values and increase the algorithms' computational efficiency. The correlation between the variables was also found, keeping only the statistically significant variables.

To construct ANNs models and decision trees the authors used Waikato Environment Knowledge Analysis (WEKA), while to implement the logistic regression models Statistical Package for the Social Sciences (SPSS) was used. From each of the three groups, 80% of the cases was selected to train the models and the remaining 20% for validation. The comparison between the models' performance was done by evaluating the respective area under the curve (AUC) values. ANNs outperformed the other models in the great majority of training and validation groups. Accordingly, the authors consider that ANNs have shown encouraging potential in CDSSs regarding this particular context: using HCC patient's clinical records to predict their disease-free survival after resection. According to the authors' interpretation, "physicians may also consider machine-learning methods as a supplemental tool for clinical decision-making and prognostic evaluation." [46]

This is an interesting work, but with limited potential as regards our objectives. In the first place, its area of application boils down to the prognosis of patients who have received hepatic resection. The "inclusion criteria" are very strict, which means that patients treated with transplantation and ablation, patients with histological evidence of benign tumours, patients in advanced stages of the disease or patients for which the tumour was not completely removed are automatically discarded. The same with patients with incomplete data, which does not reflect the reality of most clinical contexts. In addition, the study also does not take into account the patient's clinical evolution, and his prognosis is constrained to a dichotomous state: "free-disease survivor" or "non-free-disease survivor/dead". Thus, this work may be seen as a classification task, where a set of clinical variables are evaluated and a binary classification is produced, indicating whether or not the patient is free of disease in the considered interval (1, 3 or 5 years). Of extreme importance is to notice that the prognosis is made after resection, which means that the model will not be very useful as regards the decision-making process, since the decision has already been taken.

3.2.4.4 Mortality Prediction for Hepatocellular Carcinoma patients after hepatic resection

From the same authors of [46], this study compares the performance of ANN and logistic regression models to predict mortality of HCC patients who underwent liver resection [47]. The methodology is very similar to the previous study, however, the variables' selection is made in a different way. For each model and each group of survival (1, 3 and 5 years), the selected variables vary. Another difference is that recurrence is also considered as an input variable, in addition to those described in [46], being recognized as an important predictor of mortality in patients with HCC.

The only relevant difference between the two studies is the response of the algorithms - one seeks to predict disease-free survival and the other only intends to predict if the patient is alive or dead in the considered periods, may he be disease-free or not. Thus, the same limitations as [46] may be encountered.

3.2.5 Interactive decision support in hepatic surgery

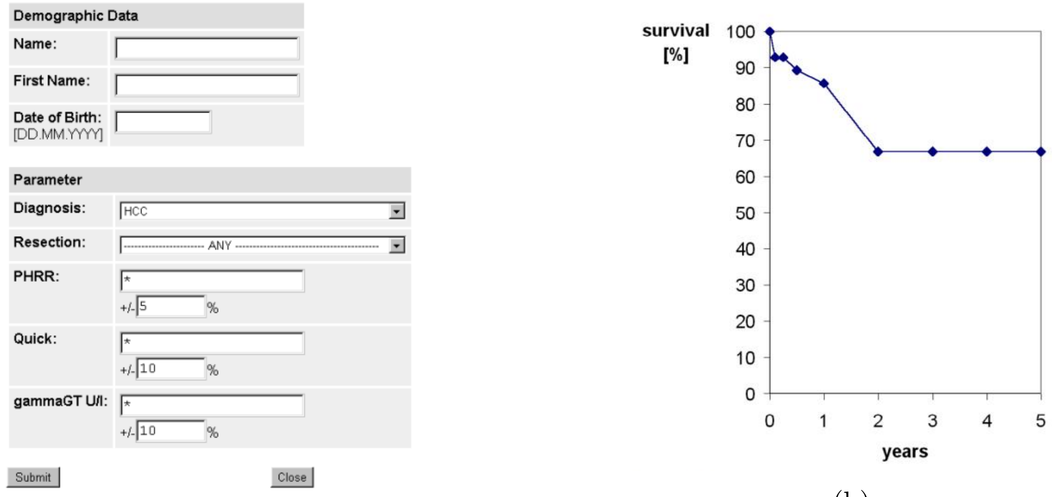
Hepatic surgery covers a set of complicated operations with significant perioperative and post-operative risks for the patient. Researchers from University of Munich developed a web-based risk assessment tool that collects and analysis patient's data and determines what kind of patients do benefit from specific procedures based on survival and complication rates [48]. The basic idea is to find similar cases to a given patient. The similarity criteria is quite simple: a case is similar to a certain patient if all considered predictive parameters correspond with a given level of tolerance. Similar cases are displayed to the physician, so that he can verify the analysis, excluding the cases he finds inappropriate, if necessary. The prognosis of matching cases is then aggregated and taken as an estimate for the risk of an individual patient. The risk is visualized as a Kaplan-Meier plot, the standard for visualizing survival data in Medicine [48].

The risk assessment tool is written in PERL, running on a Linux machine providing Apache web server, and a PostgreSQL database. Data entry is performed with a standard web browser. The authors developed a software tool for "rapid prototyping of highly adaptive web forms and management of data transformations" [48], similar the UltraDev extension of Macromedia Dreamweaver [3] , but adapted to the needs of medical databases, that is, with more specific templates. This tool allows an interactive definition of database tables. A preview of the forms is generated and shown to the physicians, and once the structure is defined, all PERL programs and database tables are generated. Each item in the data structure has a set of attributes: type of item (text, pulldown menu, checkbox, radio button, textarea, date, time), default values, constraints, layout and a unique object ID, so that data transformations can be easily made if the data structure is updated. The database itself consists in eight tables (demographics, medical history, volumetrics, surgical documentation, histology, laboratory values, complications and outcome), with an overall number of 451 items (numerical and categorical) that can be stored. This high number of items makes avoiding missing data an impossible task. However, according to the authors, the similarity search also includes records which have missing values, though they do not explain the search processes in these cases. Furthermore, the research database provides a set of specific reports, e.g. the number of patients per diagnostic category or a list of patients with lost-followup [48]. Other functions of the system include user authentication and access control (to secure patient information) and tools for data export, in XML format.

When the physicians access the application, a form is presented (Figure 3.12), requiring patient's demographic data for whom a suggestion is needed. Additionally, five clinical relevant parameters have to be specified, namely diagnosis, type of planned resection, partial hepatic resection (PHRR), prothrombin activity (Quick) and gamma-GT. After submitting the form, the system connects to the database to retrieve the appropriate results, computes the Kaplan-Meier estimates and generates a web page displaying the plot and the underlying data. By simply clicking on a similar case, the physician can go directly to the database and verify the source information and decide whether that case is appropriate or not. If considered inappropriate, it can be excluded from the analysis by selecting an "exclude" button. Accordingly, the analysis are then recalculated, if necessary.

Of all the studied applications, this is the closest to our objectives. The system is not limited to strict criteria as in [46] or [47], being able to find similar cases to a larger set of HCC patients. Unlike [46], this system considers patients with and without liver resection, and with or without liver transplantation. However, it considers the overall survival, and thus, it does not give information about free-disease survival. The Kaplan-Meier plot is a superior approach for survival analysis than [47]. Using this method, the physician has more information than a simple binary classification (dead or alive). He can get an estimate of how long will that

[3] www.macromedia.com



Demographic Data

Name:	
First Name:	
Date of Birth: [DD.MM.YYYY]	

Parameter

Diagnosis:	HCC
Resection:	------------------- ANY -------------------
PHRR:	* +/- 5 %
Quick:	* +/- 10 %
gammaGT U/l:	* +/- 10 %

(a)

(b)

Figure 3.12: Risk assessement form (a) and an example of a Kaplan-Meier plot for HCC patients (b) [48].

patient be alive, according to the chosen surgical procedure. Moreover, he gets involved into the analysis and has the ability to verify and adjust it for an individual patient according to his expertise. However, one might argue that the similarity research is quite simplistic. If a set of 451 variables is stored, why use solely 5 parameters in similarity search? The authors argue that "a risk assessment tool must be fast and easy to use", justifying the choice of only 5 parameters, shown to be predictive for patient outcomes. The exportation format may also be questioned. XML is a standard integration format; however, it may be difficult to interpret by physicians, with no knowledge in the subject.

Table A.3 (Appendix A) summarizes the main characteristics of each study exposed in Section 3.2.4.

Chapter 4

Dealing with Missing Data

Missing or incomplete data are a part of almost every study involving collected information, a common drawback that researchers need to deal with when solving real-life classification tasks. There are a number of alternative ways to deal with missing data. However, the choice of an appropriate alternative must result of a careful missing data process analysis. Thus, any discussion of missing data should begin with the question of why is data missing in the first place. Missing data occurs in a variety of application domains, for several different reasons. Data could be missing for perfectly simple reasons, such as equipment malfunction, because a participant was on vacation or the data was incorrectly entered due to misinformation or human error. On the other hand, data could be missing on the basis of either the participant's observed values on the dependent variable or any of the independent variables. Understanding the reasons of missing data is fundamental to determine how those data will be treated.

Healthcare is a particularly problematic domain regarding missing data. Every day, a large amount of clinical information is collected from multiple sources and stored in database systems. Patients' data are managed by various people within the institutions, recorded in different times and formats, thus making datasets compiled from patients' clinical information very susceptible to missing data. Accordingly, modelling and predicting clinical outcomes may turn out to be a difficult quest. Survival prediction, as an example, plays an important role in end-of-life decisions, as it helps to determine which treatments should be attempted. Therefore, it is extremely important that the accuracy of this prediction is neither biased or weak in terms of statistical power. However, survival prediction models are trained with clinical datasets frequently containing missing values. In the last few decades, missing data became an attractive area of statistics, with growing studies proposing and comparing strategies for achieving the best possible result solution for missing data drawbacks, neither losing records from the database or distorting the results with the introduction of bias in the prediction process.

4.1 Missing Data mechanisms

The most two conventional approaches used for managing missing data are to delete or impute values. However, this is not an easy fix, since the latter can cause bias, while the former causes both bias and loss of statistical power [57]. This drawback can be attenuated by classifying the underlying data missing mechanism. Basically, the missing mechanism can be seen as the process underlying the generation of incomplete datasets.

Most authors agree with the taxonomy of missingness presented by Rubin and colleagues [58] [59], inferring three different explanations for missing data. Accordingly, data can be missing completely at random (MCAR), missing at random (MAR) or missing completely not at random (MNAR).

When the probability that an observation is missing is unrelated to the value of such observation or to the value of any other variables in the same study, data are MCAR. For instance, in a survey, data would not be considered MCAR if obese subjects were less likely to report their weight than individuals with normal weight - the probability that the dependent variable "weight" is missing is unquestionably related to the the value of such variable. Moreover, if women are less likely to report their weight than men, data cannot be considered MCAR, since missingness would clearly be correlated to gender. However, if a participant's data were missing for reasons that are in no way related to the study, such as a doctor's appointment, schedule difficulties, a flat tire, that patient's would be MCAR. MCAR values can also be generated by others, besides the participants. For instance, if the person responsible for filling the data misplaces or misreads documents or information. In MCAR the probability of missing data is a constant, i.e., any observation on a variable is as likely to be missing as any other.

Data are MAR if the probability of missing data on a variable is correlated with values from other variables in the study, but not with the values that would have been present in that variable, had them not been missing. The word random in "Missing at Random" makes the concept more difficult to grasp. A real life example would be people who are depressed being less likely to report their weight. The variable "weight" would be correlated to depression. If, in addition, depressed people had a lighter weight in general, the probability of missing would be correlated with the dependent variable as well, the weight itself: with a high rate of missing data among depressed people, the existing mean weight may be lower than it would be without missing data. However, if within depressed subjects the probability that reported weight is missing was unrelated to the values of weight itself (imagine that the weight varies among depressed individuals as much as among normal weighted ones), then data would be considered MAR, though not MCAR.

The third type of missing data, MNAR or Nonignorable Missing Data (NIMD), occurs when the probability that an observation is missing is correlated with the values of the other variables in the study and, in addition, directly related with the value of such observation. Following the previous examples, this would be the case if people with higher weights (obese people) are in fact more reluctant to report their weight when compared to people with normal weights. Data is not missing at random. The average weight obtained with the available data is clearly biased when compared to the mean that would be obtained with the complete data. As another example, a participant may fail to answer a question either by shame or lack of comfort: some people simply do not feel comfortable about revealing personal information, for instance. And although the information that lead to the lack of response may or may not be in the study, this doesn't make it neither random or ignorable.

Regardless of the domain, nearly every study in this field agrees with Rubin's definition of missingness patterns. However, Cismondi *et al.* [60] consider that it might not be correct to focus only on finding the appropriate imputation method according to the classification of missing data into one of the three categories described above, especially when it comes to medical databases. In some cases, missing data are generated by virtue of the sampling frequency of the study design. A good example is given in [60]: for instance, blood pressure may be sampled hourly, and lab tests 4 hourly. Considering a gridding template with 1h frequency, lab tests will show many periods of missing data. However, data is only missing because of the choice of such sampling frequency, rather than lab test not being done. According to this line of thought, not every missing datum is a "true missing", and both deletion and imputation may actually lead to wrong conclusions. Following the examples in [60], a patient with normal blood pressure has a lower blood pressure sampling frequency, when compared to another that has a blood infection, requiring, for instance, an hourly monitoring. In the case on normal blood pressure, if other variables are deleted when blood pressure is not measured, information loss may occur. Similarly, if a patient has been periodically connected and disconnected from a ventilator, there are only records of some segments of data. Imputing values for the "disconnected" segments would not be correct, since it would suggest the patient was always under monitoring, which is false, thus biasing predictive results.

Despite this considerations, previous studies have accepted that missing data are related to some missing mechanism without attempting to discriminate if absent values are created by the study design. In this review, we'll make the same assumptions. Instead of analysing if missing data should be imputed or not, and distinguish between "recoverable" and "non-recoverable" missing values [60], we'll survey some studies that lay emphasis on comparing several imputation techniques, according to the characteristics of incomplete datasets, particularly with regard to the type of illness, mechanism of missing data, number of samples, variables and percentage of missing values in the dataset.

4.2 Strategies for Missing Data imputation

Some authors distinguish between "traditional" or "conventional" treatments for missing data and "modern approaches" for dealing with missing data [57, 61, 62]. However, for sake of simplicity, we'll distinguish between "Case Deletion methods" and "Imputation methods". Case Deletion methods consist on case elimination techniques while "Imputation methods" refer to the process of replacing missing data with substitute values. Regarding the "Imputation methods", we'll further divide them in "statistical methods" and "machine learning methods", since this is the common terminology used in most recent publications [63, 64, 66].

4.2.1 Case Deletion Methods

By far, the most common approach to missing data is the elimination of cases [59]. Omitting these cases and running the analysis on what remains is the most basic of case deletion methods. Following Howell's example, if 5 subjects in the study have missing scores in one or more variables, the study is 5 observations short. [57]. This approach is known as Listwise Deletion (LD) or Complete Case Analysis. As the name implies, LD consists in eliminating cases with missing values so that only complete cases remain for analyses. The advantage of LD is allowing the application of standard analysis techniques, since the remaining data are complete. Under the assumption that data are MCAR, it leads to unbiased parameter estimates [57]. However, with data containing a great amount of missing values, LD often results in a decrease in the sample size. This leads to a loss of statistical power, even if data are MCAR. Moreover, when this assumption is incorrect, the results may be biased.

Pairwise Deletion (PD) consists in removing cases on an analysis-by-analysis basis. In other words, the cases are evaluated according to the variables they are related to. If, those cases have missing values in the considered variables, they are removed. For instance, if one participant report his weight and gender, but not his age, then he is included in the analysis involving weight and gender, but not in the analysis involving age. The problem with this approach is that the parameters of the models constructed under these method's assumptions will be based on different datasets, with different sample sizes, which lead to bias. Furthermore, similarly to LD, PD also shares the assumption that data are MCAR. As mentioned, this may lead to biased estimates when that assumption is incorrect.

4.2.2 Imputation Methods

Imputation is the process of replacing a missing datum with a substitute value. There are several imputation approaches, according to the method used to determine such "substitute" value for each absent observation. Following the definitions of most recent papers in the subject [63,64,66], we will also distinguish between "statistical" and "machine learning" methods. Both statistical and Machine Learning methods use the available complete information to impute absent values. This is an advantage compared to discarding incomplete cases, since imputing missing values provide additional information that can enhance the classification performance [65].

Statistical methods consist on the substitution of a missing value with a meaningful estimate. Typical statistical methods are based on replacing the missing values with the most similar among existing data point, without the need of constructing a predictive model to evaluate "similarity". Roughly, it consists on the application of heuristics to achieve "plausible" estimates. Statistical imputation methods include mean imputation, hot-deck imputation and multiple imputation [66]. Imputation methods based on machine learning are more complex procedures. They consist in constructing a predictive model based on the complete available data to estimate values for substituting those that are missing.

4.2.2.1 Statistical Imputation Methods

A once-common method for Statistical Imputation (SI) was hot-deck imputation [57, 61, 62], where a missing value was imputed from a randomly selected similar record. For instance, suppose an obese young female, resident in Coimbra refused to participate in a depression survey. The researches might simple get a record that came from an obese, young woman in Coimbra from another database and use it to substitute the missing record and continue their studies.

Replacing missing values with the mean of the corresponding variable, known as Mean Imputation, is the most common of the SI techniques. Though there are more sophisticated procedures, Mean Imputation is used in almost every study concerning missing data [63–67, 69, 70]. The mean is calculated using only the complete cases for the variable whose observations are missing. There are a few issues with this approach: it adds no new information to the analysis and it leads to an underestimate of error, as pointed by Little [59]. As stated in [68], this underestimation derives from two sources [57]. In the first place, from the loss of the natural variation in the data. Secondly, from the smaller standard errors produced: no new information is added, although the sample size increases, increasing the denominator in standard error's calculation, thus reducing it. Moreover, as shown in [59], Mean Imputation can attenuate the overall correlation estimate between variables.

Regression Imputation (RI) is another SI approach for handling missing data [58]. In RI, the existing variables are used to make a prediction, using a regression equation, and the predicted value is used as a substitute of the missing datum. As Little describes it, "in a bivariate analysis with missing data on a single variable, the complete cases are used to estimate a regression equation where the incomplete variable serves as the outcome and the complete variable is the predictor" [59]. The imputed value is in some way related to other information that we have about the subject or sample. In fact, as seen in [59], the imputed values will have a correlation of one with the values from the variable used in their prediction. Thus, although RI can be considered a step forward regarding the previously described methods, it can lead to an overestimation of the correlation between variables. Furthermore, the imputed values lack variability and thus the standard error of classification performance may be underestimated.

4.2.2.2 Machine Learning Imputation Methods

Missing data imputation through machine learning-based methods has recently attracted much attention. They consist in creating a predictive model to estimate the absent values from complete available information in the dataset. Some well-known learning algorithms have been applied to missing data handling, namely the Multi-Layer Perceptron (MLP) [64,65], K-Nearest Neighbours (KNN) [63,67,69], Self-Organizing Maps (SOM) [66], Decision Trees (DT) [70] and Support Vector Machines (SVM) [64, 70].

A Multi-layer perceptron is a modification of the standard linear perceptron and can distinguish non-linearly separable data [65]. It consists of multiple layers of nodes interconnected in a feed-forward way. A MLP model is trained using only the complete cases as a regression model. Given D input features, each incomplete attribute is learned (used as target) by using the other $D-1$ attributes as inputs. When several attributes are missing, several MLP schemes have to be designed, one per missing variables combination, as described in [66]. This method has some disadvantages. First of all, though MLP can solve non-linear problems, it cannot use missing data for training directly, the incomplete cases are not considered for training. Thus, when a considerable percentage of input vectors are incomplete, the results achieved by this algorithm may lead to biased learning [65]. Another downside is that when missing values appear in several combinations of attributes in a high-dimensional problem, many MLP models have to be implemented.

KNN is a classification algorithm in which the k nearest neighbours (samples or subjects) are chosen from the complete set of cases, found by minimizing a similarity measure. After finding those k closest examples in the feature space, the missing value is determined according to the type of data [66]. A majority voting of its neighbours can be used for discrete data and the mean for continuous data. Another alternative for continuous data is to weight the contribution of each k-neighbour according to its distance to the incomplete pattern [69]. This way, a greater contribution is given to the closest neighbours. It has been shown that this method provides a robust procedure for missing data estimation [65, 71]. However, its major drawback is related to the fact that KNN is a lazy learning algorithm. That is, it does not use the training data to do any generalization. Whenever the algorithm looks for the most similar neighbours, it has to search the entire dataset. This is especially problematic for large databases. Another issue is finding the optimal number of neighbours (value of k) and the most appropriate distance metric to be used. This requires a careful study of the dataset and the developement of several KNN models, in order to achieve the best results [69–71].

Self-Organizion Maps (SOM), as described in [66], are a type of artificial networks that use unsupervised learning that describe a mapping to a lower dimensional space. Basically, SOM consists of nodes placed in a d-dimensional array, where each node has a d-dimensional weight vector associated. Like most ANNs, SOM performs ”training and testing”, or in this case, ”training and mapping”. In the ”training” phase, SOM build the map using input examples. A vector in data space is placed onto the map by finding the node with the closest weight vector. Thus, nodes that are spatially close in the map have similar weight vectors. For each training input vector, the neuron that has the most similar vector is called the Best Matching Unit (BMU). The ”mapping” phase classifies a new input vector, according to the distances between the vector and the nodes. The ”distance function” is called *neighbourhood function*, explained in detail in [66]. When an incomplete vector is used as input to SOM, the missing observation are ignored during the selection of the BMU. The incomplete values are imputed with the values of the BMU in the missing dimensions. In other words, each missing value is imputed based on the weight vector of the BMU in the incomplete attributes.

Decision Trees (DT) are a well-known data mining algorithm, expressed as a ”recursive partition of the instance space" [70]. Their main advantage is that they are self-explanatory and can handle both continuous and nominal data, missing data and datasets that may have errors. With a reasonable number of leaves, DT can be compacted and converted to a set of rules, which are an easy-to-grasp representation of data [72]. One of DT's disadvantages is that some DT algorithms require that the target attribute has only discrete values, which could be problematic to input continuous variables.

The Support Vector Machine (SVM) is a state-of-the-art approach to pattern classification and regression, due to its ability to deal with high-dimensional data and flexibility in modelling diverse sources of data [73]. SVM can provide a good generalization performance since they tackle the principle of structural risk minimization [74] by balancing the model's complexity against its success at fitting the training data. They provide a good tradeoff between the flexibility of the model and the error in training data [75]. Thus, SVMs satisfy the Occam's Razor Principle: among competing solutions, with similar results, the one with the fewest assumptions should be chosen. SVMs belong to the general category of kernel methods. A kernel methods can operate in high-dimensional spaces, since they depend on the data only through dot-products. This has two main advantages: it allows to generate non-linear decision boundaries and enables the classification of data that have no obvious fixed-dimensional vector space representation [76, 77]. SVMs are known for excellent classification performance [70]. However, they require a comprehensive understanding of how they work. When training SVMs, researchers have to face several decisions: how to preprocess data, which kernel function to use and setting the SVM and kernel parameters. Uninformed decisions may lead to reduced performance, and thus, the use of SVM requires a comprehensive understanding of these choices, which can be considered a disadvantage. In Support Vector Machines Imputation (SVMI), the SVM model is trained using all examples that have no missing values. After achieving the optimal SVM parameters, the model is used to impute missing values. Absent attributes are treated as targets, using the remaining complete attributes as inputs.

4.3 Conclusions

In several papers in the literature [63–67,69,70], the authors evaluate the performance of several statistical and machine learning imputation methods, to investigate how different imputation methods can overcome the missing data problem. They all reach the same conclusion: machine learning techniques outperform statistical methods. However, as stated in [66], imputation techniques "depend on the available data and the prediction model used", and thus they have to be adapted according to the context, that is, the best imputation technique found for a particular dataset may not generalize well to different datasets. In our approach, we intend to impute values according to case-similarity (an instance's missing values should be imputed according to its most similar instance). In addition, we looked for methods fairly simple to explain to clinicians, and that did not require a high computational effort, that could prejudice our CDSS's performance. Therefore, we have chosen Mean Imputation, Logistic Regression and KNN to impute our dataset's missing values, as discussed in Chapter 6.

Chapter 5

Clinical Information System Development

In this chapter, we'll present our clinical system in detail, through the main steps of its development: requirement analysis, use cases definition, architecture, technological choices, prototype and final software platform.

5.1 Requirements Analysis

The software requirements specification is fundamental to delineate the boundaries of our clinical information system design and functionality. The Software Requirement Specification (SRS) will define and illustrate the overall project and its requirements - both functional and non-functional. In addition, the SRS will also define the users and their respective characteristics as well as any constraints to the system development the team has identified.

Functional requirements describe the behaviour of the system as it relates to the system's functionality. According to [79], they are "statements of services the system should provide, how the system should react to particular inputs, and how the system should behave in particular situations". Non-functional requirements elaborate the performance's characteristics of the system. Typically, non-functional requirements fall into areas such as accessibility, efficiency, extensibility, privacy and maintainability, among others.

In sections 5.1.1 and 5.1.2, we will list the identified requirements. They are presented with a requirement id, a brief decription and priority category, according to the MoSCOW method [80]:

M - MUST: Describes a requirement that must be satisfied in order to the final solution to be considered a success.

S - SHOULD: Represents a high-priority item that should be included in the solution, if possible.

C - COULD: Describes a requirement that is considered desirable but not necessary. This type of requirement is included if time and resources permit.

W - WOULD: Represents a requirement that stakeholders have agreed will not be implemented in a given release, but may be considered in the future.

5.1.1 Functional Requirements

This section presents a list of the functional requirements, classifier according the MoSCOW method. These requirements are aggregated according to their context. Thus, we have considered Filtering, Consultation, Importation, Edition, Creation, Data Exportation, Reporting and Deletion requirements.

Filtering requirements concern the user's filter searches to the system. A user must (M) be allowed to search data by patient's name or ID.

Table 5.1: Filtering Requirements.

F-1	**Filtering**	**Category**
F-1.1	Patient's filtering by name	M
F-1.2	Patient's filtering by Patients ID (PID)	M

Consultation requirements describe the mandatory need to consult clinical data. The clinicians must (M) be able to see any patient's medical evaluation, exams, risk factors or performed treatments.

Table 5.2: Consultation Requirements.

C-2	**Consultation**	**Category**
C-2.1	Patient's medical evaluation consultation	M
C-2.2	Patient's exams consultation	M
C-2.3	Patient's risk factors consultation	M
C-2.4	Patient's treatments consultation	M

Importation is a fundamental requirement of our system. It is mandatory (M) that the system can import .xls files (the format required by CHUC's team). Since .csv files are also commonly used within the institution to manipulate and share data, they should (S) be imported as well, if possible.

Table 5.3: Importation Requirements.

I-3	**Importation**	**Category**
I-3.1	Importation of .xls files	M
I-3.2	Importation of .csv files	S

Editing patient's data is a major system's functionality. A user must (M) be able edit any type of patients' records, whether they are risk factors, medical evaluations, exams or treatments.

Table 5.4: Edition Requirements.

E-4	**Edition**	**Category**
E-4.1	Edition of patient's risk factors	M
E-4.2	Edition of patient's medical evaluation	M
E-4.3	Edition of patient's exams	M
E-4.4	Edition of patient's treatments	M

Without the creation of patients or patient's records, the systems has no use. Thus, there are clearly mandatory (M) requirements. The system mus enable the creation of all types of patient's data (risk factor forms, medical evaluations, exams and treatments).

Table 5.5: Creation Requirements.

CR-5	**Creation**	**Category**
CR-5.1	Creation of a new patient	M
CR-5.2	Creation of a new patient's risk factors	M
CR-5.3	Creation of a new patient's medical evaluation	M
CR-5.4	Creation of a new patient's exams	M
CR-5.5	Creation of a new patient's treatments	M

CHUC's team manifested the need to export system's data. In particular, they required .png files (M). Other formats such .pdf, .svg and .xls are also a priority, and they should (S) be covered by the system. These formats should be included in future releases. .jpeg could (C) be included, but it is not a absolute necessity.

Table 5.6: Data Exportation Requirements.

DE-6	**Data Exportation**	**Category**
DE-6.1	Exportation in .pdf format	S
DE-6.2	Exportation in .svg format	S
DE-6.3	Exportation in .png format	M
DE-6.4	Exportation in .xls format	S
DE-6.5	Exportation in .jpeg format	C

The CHUC's team has expressed a great interest in a reporting functionality. This must (M) be included. The user must be able to query the systems according to a predefined set of options. More elaborate queries, such as results per group or filter are also desired and thus the system should (S) meet this requirements, if possible. Other types of queries (such as per filter and group) are not a priority, and could (C) if the time constraints allow.

Table 5.7: Reporting Requirements.

R-7	**Reporting**	**Category**
R-7.1	Reporting results per filter	S
R-7.2	Reporting results per group	S
R-7.3	Reporting results per filter and group	C
R-7.4	Reporting results per option	M

The user must (M) be able to remove any of the inserted patient clinical data: risk factors, medical evaluations, exams or treatments.

Table 5.8: Deletion Requirements.

D-8	**Deletion**	**Category**
D-8.1	Deletion of patient's risk factors	M
D-8.2	Deletion of patient's medical evaluation	M
D-8.3	Deletion of patient's exams	M
D-8.4	Deletion of patient's treatments	M

Authentication is an important requirement that must (M) be verified in our system. The patients' data protection has to be guaranteed through a unique password per clinician. Each clinician's credentials must (M) be verified every time the clinician accesses the system. The user's passwords should (S) meet some complexity rules. The need to change passwords periodically and to lock accounts in case of multiple login failures is to be accessed in future releases.

Table 5.9: Authentication Requirements.

A-9	**Authentication**	**Category**
A-9.1	Each user must have his own password	M
A-9.2	User credential are verified each time the user accesses the system	M
A-9.3	Require a minimum password of at least 8 characters	S
A-9.4	Require passwords with Lowercase, Uppercase, Numbers and Special characters	S
A-9.5	Require users to choose new passwords at least 90 days and prevent the reuse of a password for 1 year	C
A-9.6	Lock acess to accounts if there are 30 failed authentication attempts within 5 minutes	W

As our system is intended to be a recommendation system, the integration of an AI module is also considered to be a functional requirement. The system should use the patients data to provide meaningful information regarding treatment options and/or survival prognosis.

Table 5.10: Artificial Intelligence Module Requirements.

AIM-10	**Artificial Intelligence Module**	**Category**
AIM-10.1	Classify a given patient into a prognostic group	S
AIM-10.2	Predict overall survival according to a patient's characteristics	S
AIM-10.3	Update currently existing patient profiles	W
AIM-10.4	Recommend the most appropriate treatment according to a patient's similar cases	C

5.1.2 Non-Functional Requirements

Non-functional requirements relate to the system's performance characteristic. They may also describe aspects of the system that do not relate to it's execution, but rather to it's evolution over time. We have identified Implementation and Documentation requirements.

Implementation requirements include a user-friendly interface (a requisite especially emphasised by CHUC's team), system's extensibility, availability and usability. Extensibility is the system's capability to grow, that is, to incorporate new functionalities without affecting its internal structure and data flow. System's availability concerns the system's capability to work as required whenever the user needs. Usability includes metrics of effectiveness (if the users can successfully achieve their goals), efficiency (users' effort to achieve those goals) and satisfaction (users' experience feedback). System's documentation could (C) also be helpful for future users and developers.

Table 5.11: Implementation Requirements.

IM-11	**Implementation**	**Category**
IM-11.1	User-friendly Interface	M
IM-11.2	System's Extensibility	W
IM-11.3	System's Availabilty	M
IM-11.4	System's Usability	M

Table 5.12: Documentation Requirements.

DC-12	**Documentation**	**Category**
DC-12.1	System's features documentation	C
DC-12.2	System's accessibility documentation	C

Some users may not be used to deal with web-applications and related technologies. Therefore, some aspects of the application may not be so intuitive as we planned to be. Thus, a help section could (C) be useful to users that have some doubts about the system's functionalities and usability.

Table 5.13: Help Section Requirements.

H-13	**Help Section**	**Category**
H-13.1	Help section in main menu (filter options, insert new data, visualize reports)	C
H-13.2	Help section in each secondary menu (edit and delete previous entered data)	C
H-13.3	Help section in Reporting tab (available filter options, available reports, exporting options)	C

Navigation is a key component of a web-application. Navigation is the gateway into different sections of content, and needs to be very easy and intuitive. It must (M) be organized, with tabs for general actions and sub menus for specific actions. It must (M) use obvious section names, so that the user can quickly find what he is looking for (general and ambiguous words should be avoided). Once a user clicks into a application section, the system should (S) remind him ”where he is”, using a consistent methods to highlight the section the user is in, such as a change in color or appearance. Drop-down menus that break down top-level buttons into sub-sections should (S) be considered. Also, one should avoid too many separate navigation bars. The application must (M) be consistent, maintaining the same style, type and colors, to enable the users to get used to the application and feel comfortable browsing it.

Table 5.14: Navigation Requirements.

N-14	**Navigation**	**Category**
N-14.1	Main tabs for major actions and sub menus for secondary tasks	M
N-14.2	Obvious Section Names	M
N-14.3	Highlight the section the visitor is in	S
N-14.4	Few navigation buttons	S
N-14.5	Maintain the same style, type and color in all the menus	M

Data visualization should help the used to discern relationships in the data. Thus, the type of display choices should be chosen in such way that they do not distort reality, contain the necessary information and are presented in a way that the clinician understands.

Table 5.15: Visualization Requirements.

V-15	**Visualization**	**Category**
V-15.1	Enable data visualization through bar charts (e.g, risk factors distribution)	M
V-15.2	Enable data visualization through Kaplan-Meier Curves (e.g Survival Analysis)	M
V-15.3	Enable data visualization through simple tables (reporting results)	M
V-15.4	Use data-driven visualization libraries for data presentation layer	S
V-15.5	Allow user interaction with the graphs (see a particular point in the graph or label information)	S

5.2 Use Cases - UML Diagram

In this section we present the Use Case Diagram of the system using UML. The objective of this diagram is to illustrate the system's actors and their roles (Figure 5.1). Each use case has an associated ID, which will be used to identify and describe each one in the following section.

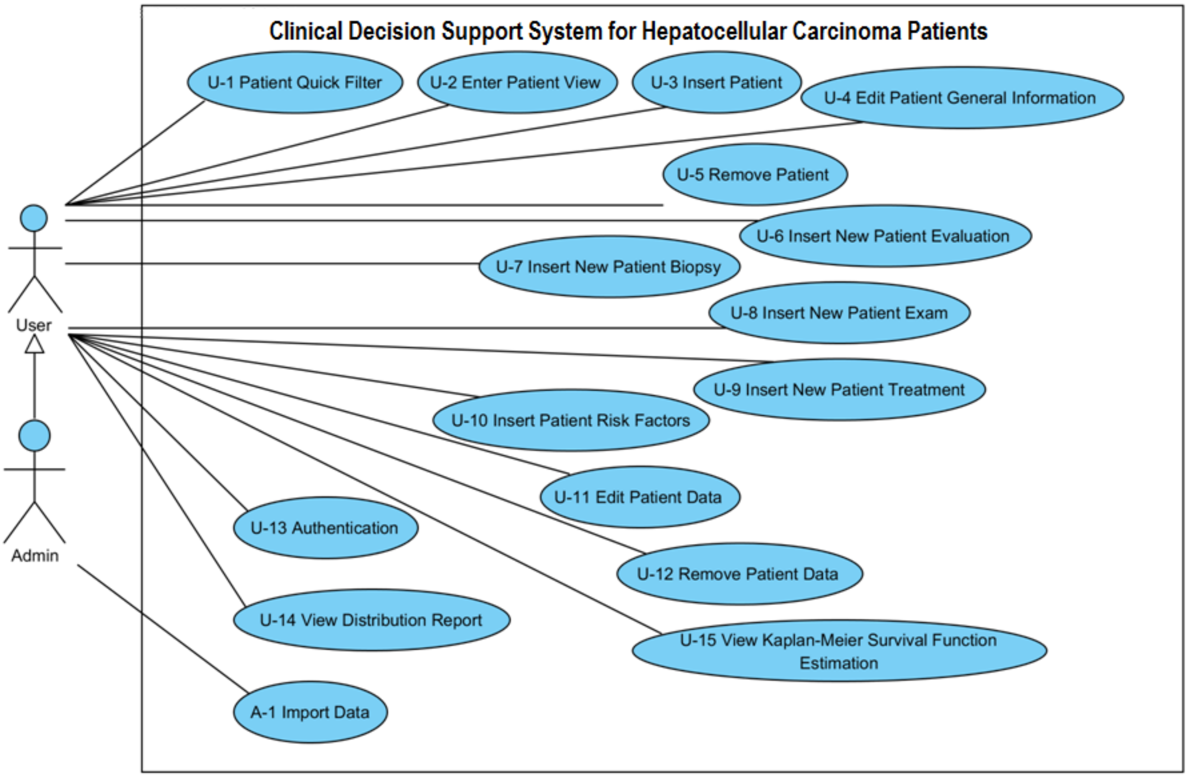

Clinical Decision Support System for Hepatocellular Carcinoma Patients
U-1 Patient Quick Filter
U-2 Enter Patient View
U-3 Insert Patient
U-4 Edit Patient General Information
U-5 Remove Patient
U-6 Insert New Patient Evaluation
U-7 Insert New Patient Biopsy
U-8 Insert New Patient Exam
U-9 Insert New Patient Treatment
U-10 Insert Patient Risk Factors
U-11 Edit Patient Data
U-12 Remove Patient Data
U-13 Authentication
U-14 View Distribution Report
U-15 View Kaplan-Meier Survival Function Estimation
A-1 Import Data
User
Admin

Figure 5.1: Use Cases Diagram.

Actors are divided in two major groups: User and Admin. The User is the general application user, for whom the application was intended. After authentication, he has access to all the application's features, except for data importation. The Admin is a more specialized user, usually the application's developer. He has access to the User's functionalities in addition with data importation. Admin also manages the users' accounts.

5.2.1 Brief Description of Use Cases

This section presents a brief description of each use case. After illustrating the general environment in Figure 5.1, we elaborate this analysis with more detailed information.

Table 5.16 lists the different use cases, indicating their IDs, actors and names. Each one of the following tables is related to a single use case, identified with its own ID and a short expression that represents its name. After indicating the actors involved, the use case is defined in a small description. In some cases, there are other characteristics in the table, namely the trigger, the use case's preconditions and postconditions, normal and alternative flow or special requirements. In particular, U-2 has the indication of the assumptions, and A-1 the frequency of use. Some tables have also the indication of notes and uses.

Table 5.16: Use Cases List.

Use Case ID	Primary Actor	Use Cases
U-1	User	Patient Quick Filter
U-2	User	Enter Patient View
U-3	User	Insert Patient
U-4	User	Edit Patient General Information
U-5	User	Remove Patient
U-6	User	Insert New Patient Evaluation
U-7	User	Insert New Patient Biopsy
U-8	User	Insert New Patient Exam
U-9	User	Insert New Patient Treatment
U-10	User	Insert Patient Risk Factors
U-11	User	Edit Patient Data
U-12	User	Remove Patient Data
U-13	User	Authentication
U-14	User	View Distribution Report
U-15	User	View Kaplan-Meier Survival Function Estimation
A-1	Admin	Import Data

A full description of the use cases in terms of their description, triggers, normal and alternative flows, notes and related issues can be consulted in Appendix B.

5.2.2 Entity-Relationship Diagram

The Entity-Relationship (ER) diagram presented in Figure 5.2 illustrates the logical structure of the developed database.

Our database is composed by seven entities (Users, Patients, Medical Evaluations, Risk Factors, Biopsy, Exams and Treatments) that relate to each other with different cardinalities:

Users: This entity describes the Users' information to be stored. Each User has an unique id, a username and password, a type (general user or admin), and a date of his last login and last activity in the platform. The Users entity has a 1-to-N relationship with all the other entities, except Patients entity. That is, a user can be associated to N exams, risk factors, medical evaluations, biopsies and treatments. In this context, ”to be associated” simply means each user can has access and can insert/edit records from all the other entities.

Patients: Basically, Patient's entity describe the patient's essential attributes: id, name, date of birth, sex, age at the diagnosis, among others. Patients entity also has a 1-to-N relationship with the remaining entities (except Users), since each patient may have recorded data concerning each one of the other entities. In other words, for each patient, there may have N recorded exams, medical evaluations, risk factors and so on.

Medical Evaluation: This entity aggregates information related to each medical evaluation - date, blood tests variables and results of physical examination. Each medical evaluation is related to a patient and is created by a user (a clinician). There can be N exams,

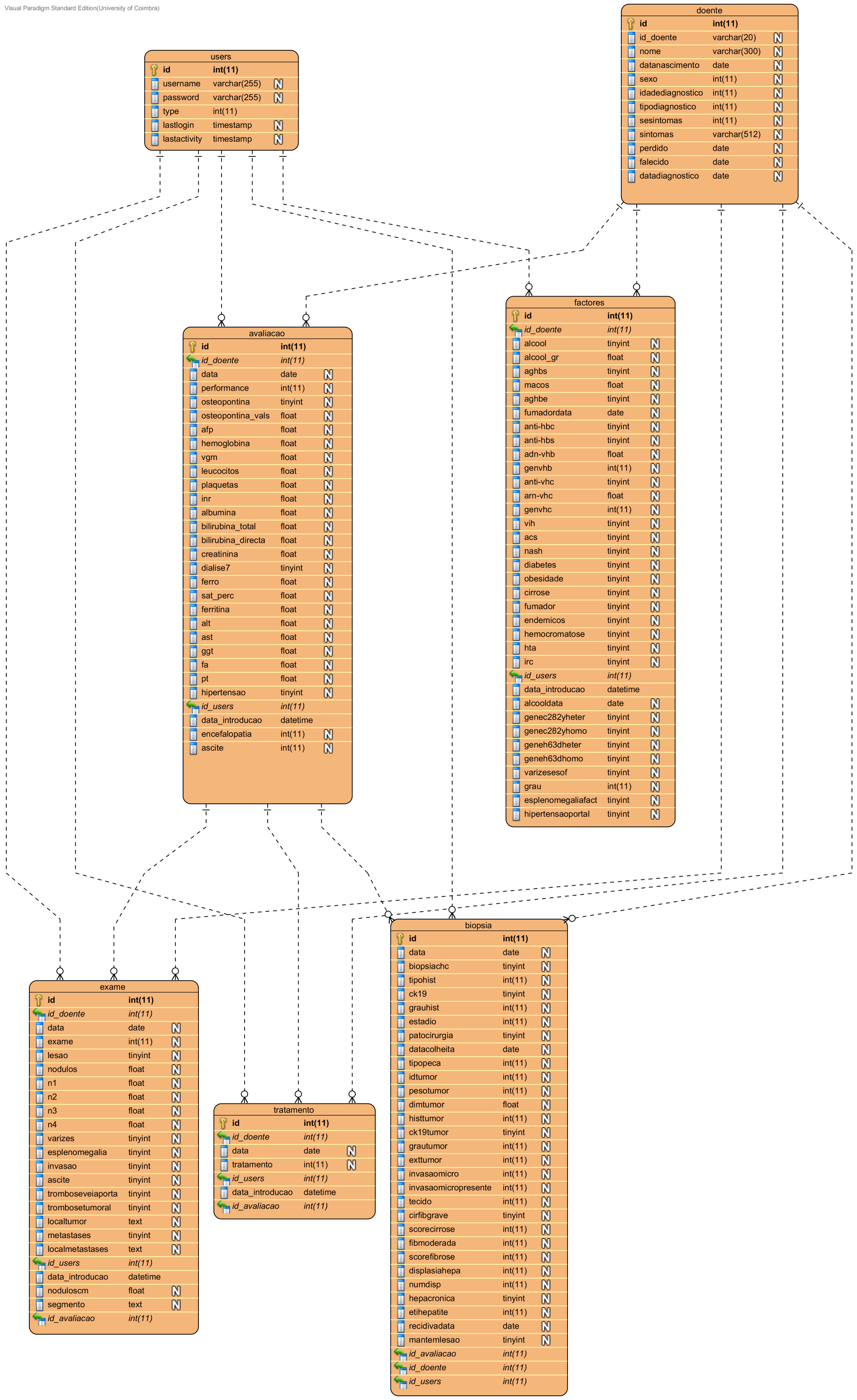


Figure 5.2: Entity-Relationship Diagram.

treatments or biopsies associated to each medical evaluation: for instance, sometimes the diagnosis requires several imaging exams or a biopsy to be performed. As another example, a medical evaluation may suggest the need for several treatments, such as a sequence of radiofrequency ablations, or a transplantation followed by chemoembolization.

Risk Factors: Risk factors entity is self-explanatory. It includes all risk factors that might be verified for each patient. It only relates to two other entities: Users (the actors that insert these information in the system) and Patients (to whom the information refers).

Exams, Treatments and Biopsy: These three entities encompass clinical data regarding exams, treatments and biopsies. They all relate to Patients, Users and Medical Evaluation entities: Users insert these Patients' information in the system; Medical Evaluations may include N exams, treatments and biopsies as previously explained.

5.3 Framework

Adopting a Web Application instead of a Desktop Application is a choice that is becoming more and more frequent over the years in software development. This is mainly because web technologies have advanced to such a point where the behaviour of the interface in terms of usability and animation rival with the Desktop Applications. Moreover, the ubiquity of web browsers allow to cross any platform boundaries without the need of additional code. On the other end, a Web Application allows us to fill other requests like centralization, multi-user support and real time access to updated information from the Institution internal network, or any computer with access to the internet, without any additional configuration (Figure 5.3). However, Web Applications also have disadvantages. In our case, they relate to the need of an internet connection in order to access the application. However, in our work context, this did not constitute an important issue.

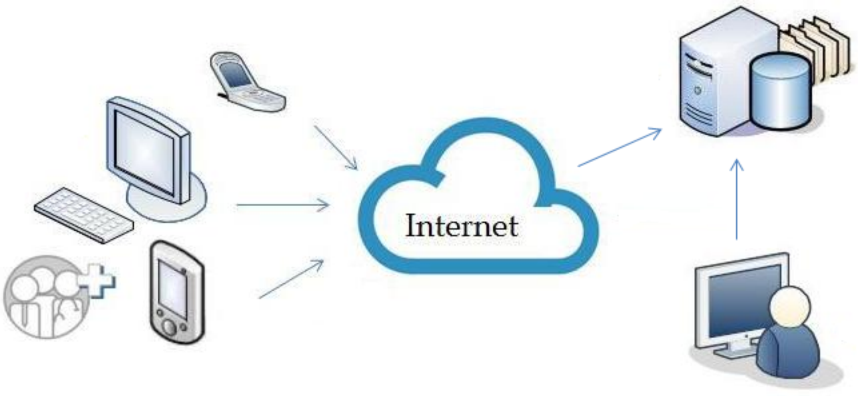


Figure 5.3: System's external interaction diagram.

Clinicians access the database through our web-application, using the Institution computers and internal network, or any other devices (personal computers, smartphones, tablets) as long as they are connected to the Internet. The Admin has local access to the system's database, so that he can manage, configure and update it, as shown in Figure 5.3.

5.3.1 Technologies

The application was implemented using PHP 5 as the server side scripting language. Running on a Apache 2.X Web server, it was supported by a MySQL 5.5 database. On the client side we took full advantage of HTML 5 features and associated technologies, CSS3 for styling, Ajax requests for in-page loading of content and JavaScript for interface management and control.

In order to avoid unnecessary development time we've chosen several frameworks and modules that would fit our system's features and behaviours. We have also used the following third party libraries:

1. MooTools [1]JavaScript framework (*mootools-core-1.5.0.js and mootools-more-1.5.0.js*)
2. MooTools Plugins:
   - History (*mootools.history.js*)
   - Auto-completer (*Autocompleter.js*)
   - Date picker (*Picker.js*)
3. D3.js JavaScript library (*d3.js*)
4. Dimple Charts (*dimple.js*)
5. Canvg SVG to Canvas converter (*canvg.js*)
6. PHPExcel v1.8

Quoting the MooTools developers, "MooTools is a compact, modular, object-oriented JavaScript framework designed for the intermediate to advanced JavaScript developer. It allows to write powerful, flexible, and cross-browser code with its elegant, well documented, and coherent API". MooTools's API is similar to some extent to the more popular API jQuery, and was an indispensable tool in terms of easing the manipulation of DOM [2] objects in order to provide the user with a simple-to-use, yet rich application.

5.3.2 Prototype

This section describes the first steps in the construction of our CDSS. Our prototype is a less detailed initial release, developed to validate some user requirements and preferences. The prototype's architecture is fairly simple (Figure 5.4). The patients' data are entered into an .xls file and parsed to a XML file. PHP reads the XML file, processes the data, and creates the web pages. HTML is used to structure the web pages while CSS is used for styling. JavaScript and jQuery are used for HTML manipulation, event handling and animation. The prototype [3] was developed in Portuguese, according to the CHUC's preferences.

The prototype's functionalities are shown in Figures 5.5 to 5.14. When users first try to access the application from a web browser, an HTML login page appears prompting the users for a username and password. The user's authentications, were used to determine the his role: admin or clinician. However, the same information was available for both types of users, since the only difference in their permissions was the authorization to data importation or not. This detail was not contemplated in the prototype.

When the user's login credentials are validated, the application presents a list of all the existing patients (Figure 5.6). Only the most relevant attributes for patient's identification are

[1] http://mootools.net/

[2] Document Object Model (DOM) is an application programming interface (API) for valid HTML and well-formed XML documents. It defines the logical structure of documents and the way a document is accessed and manipulated.

[3] The prototype may be explored by accessing http://chucdb.dei.uc.pt/login.php

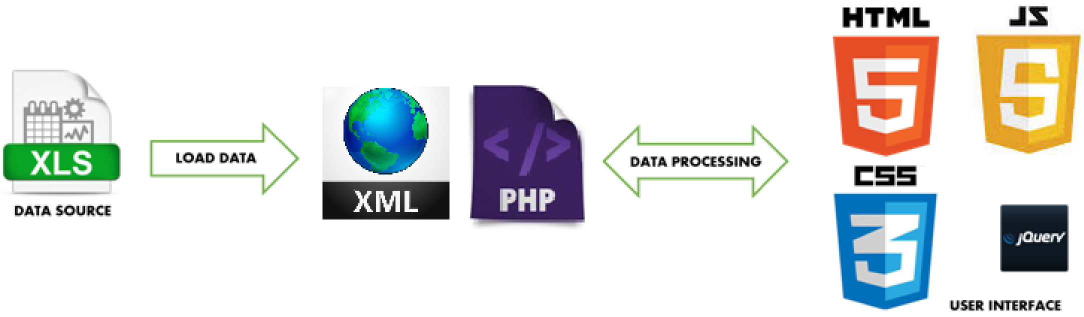
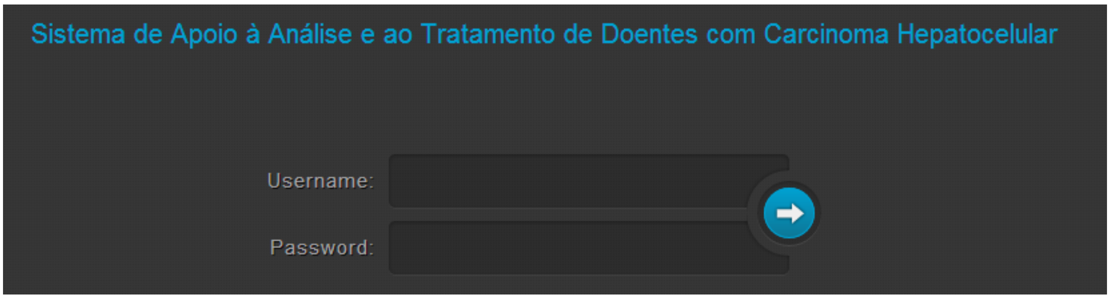


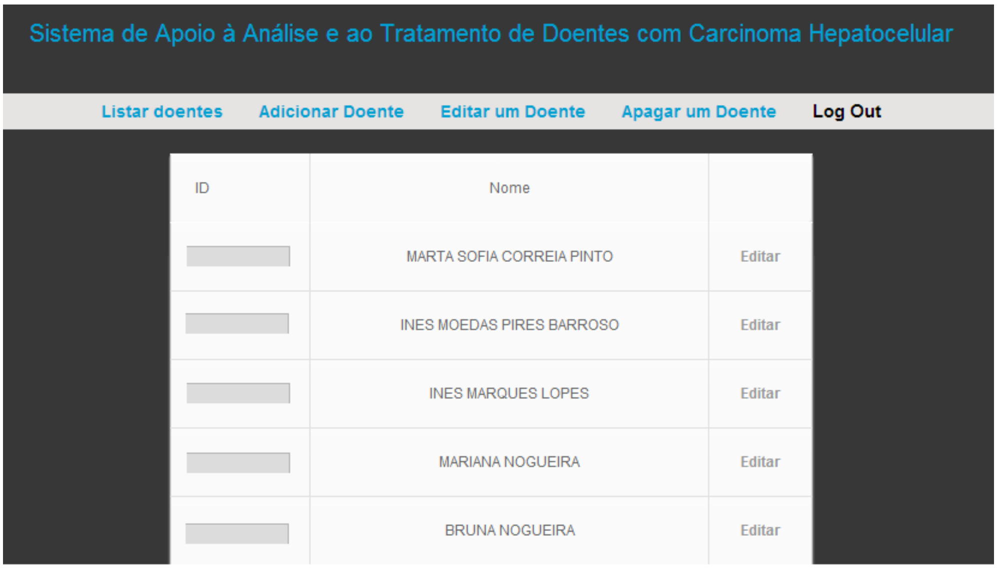
Figure 5.4: Prototype's architecture and technologies.



Sistema de Apoio à Análise e ao Tratamento de Doentes com Carcinoma Hepatocelular

Figure 5.5: Prototype's login page.

presented: ID, Name, Date of Birth, Gender and Age at Diagnosis. The user has to scroll down in order to see all of them, since the application of filters is not covered. Each patient's complete set of clinical data can be consulted by clicking over the patient's name (Figure 5.7). Non-existing information is identified by blank spaces in text fields, no filling in radio buttons or check boxes and a pre-defined option in drop-down list, such as "No information" (Figure 5.8).

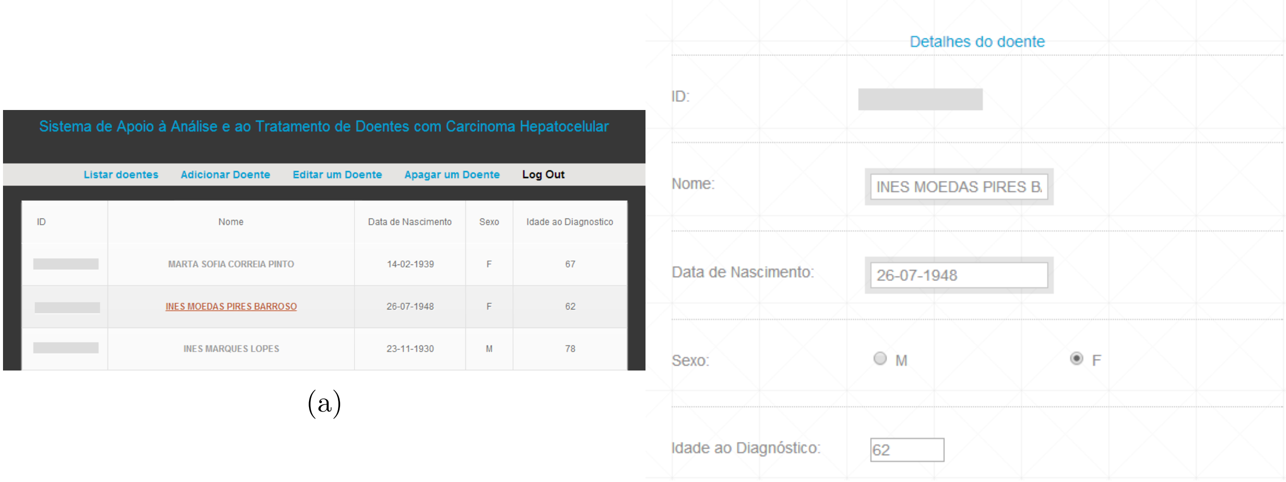


Sistema de Apoio à Análise e ao Tratamento de Doentes com Carcinoma Hepatocelular

ID	Nome	
[redacted]	MARTA SOFIA CORREIA PINTO	Editar
[redacted]	INES MOEDAS PIRES BARROSO	Editar
[redacted]	INES MARQUES LOPES	Editar
[redacted]	MARIANA NOGUEIRA	Editar
[redacted]	BRUNA NOGUEIRA	Editar

Figure 5.6: Prototype's list of patients.

The application's horizontal menu, on the top of the page, contains an array of options, namely "List Patients", "Add Patient", "Edit Patient", "Delete Patient" and "Log Out". By clicking in a chosen menu item, the user opens a defined part of the web application.



Sistema de Apoio à Análise e ao Tratamento de Doentes com Carcinoma Hepatocelular

ID	Nome	Data de Nascimento	Sexo	Idade ao Diagnostico
[redacted]	MARTA SOFIA CORREIA PINTO	14-02-1939	F	67
[redacted]	INES MOEDAS PIRES BARROSO	26-07-1948	F	62
[redacted]	INES MARQUES LOPES	23-11-1930	M	78

(a)

Detalhes do doente

ID: [redacted]

Nome: INES MOEDAS PIRES B.

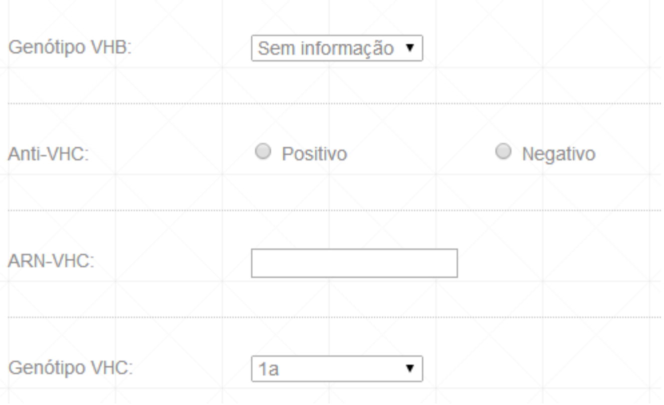
Data de Nascimento: 26-07-1948

Sexo: ○ M ● F

Idade ao Diagnóstico: 62

(b)

Figure 5.7: By clicking over the patient's name (a), his clinical data may be consulted (b)
.



Genótipo VHB: Sem informação ▾

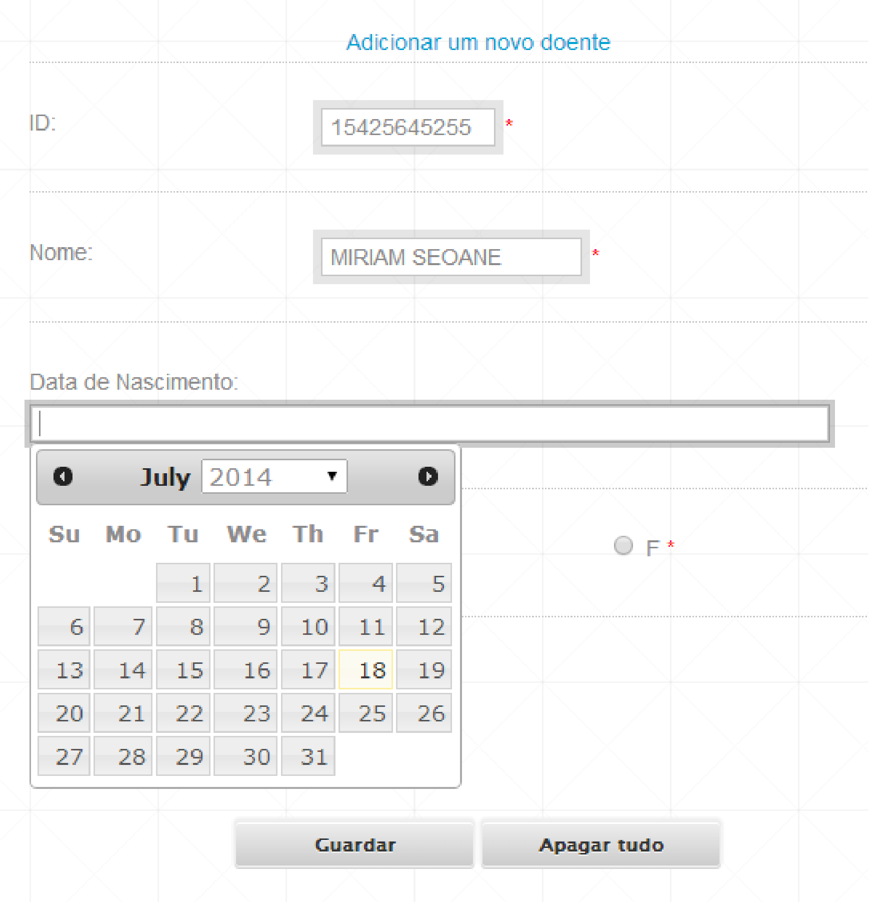
Anti-VHC: ○ Positivo ○ Negativo

ARN-VHC:

Genótipo VHC: 1a ▾

Figure 5.8: Non-existing information is identified in various ways.

When the option "Add Patient" is chosen, a web form appears, allowing the user to insert a new patient and fill his demographics (Figure 5.9), risk factors (Figure 5.10), exams (Figure 5.11) and medical evaluation data (Figure 5.12).



Adicionar um novo doente

ID: 15425645255 *

Nome: MIRIAM SEOANE *

Data de Nascimento:

July 2014

Su	Mo	Tu	We	Th	Fr	Sa
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

○ F *

Figure 5.9: Prototype's demographics page.

Some fields are mandatory (identified with a red asterisk) and are validated to avoid inconsistency errors (Figure 5.13). A validation message is shown to inform the user about the cause of the error.

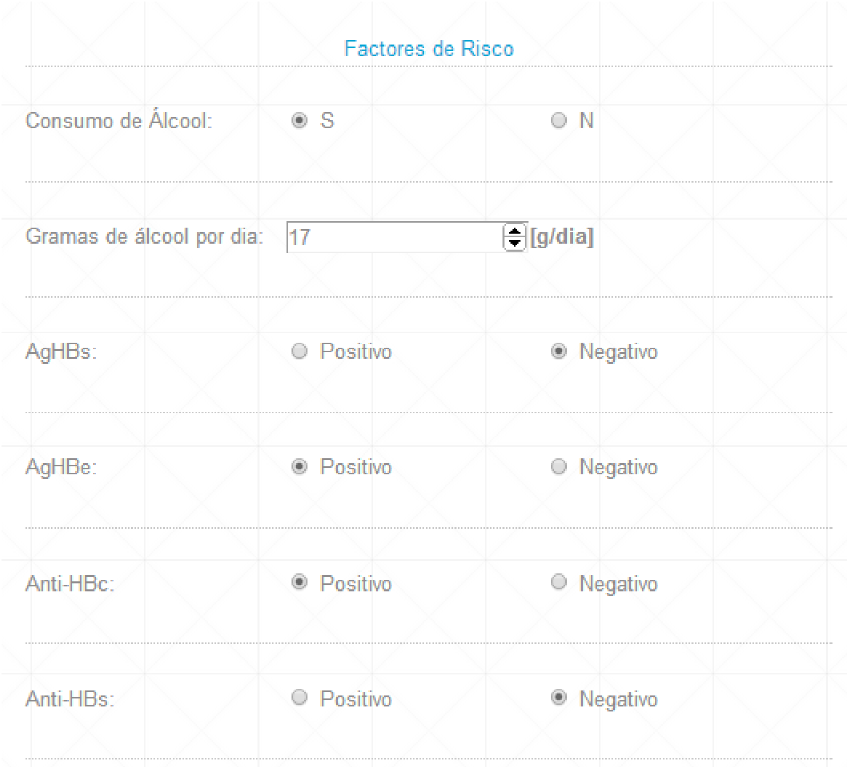
"Edit Patient" form shows all the patient's information, which can be altered by the user (Figure 5.14). Any modification to the data are saved by pressing the "Save" button.

The system's final version included several other features that were not initially covered, such as a reporting tab, filters to query the database, importation tab (for admins) and several compact forms for data entry, that do not require the user to scroll the page. In the next section, we will describe these functionalities in greater detail.

5.3.3 Final Version

Considering the requirements gathered from previous phases of the development, and having in mind the lack of technological experience of the final user (and thus the need to build an intuitive interface), the prototype was re-built, resulting in this new final version [4]. It consists of a common frame, carrying CHUC's logo and name, and similarly to the prototype, the application's name. It is important to state that the final version of the system's was developed

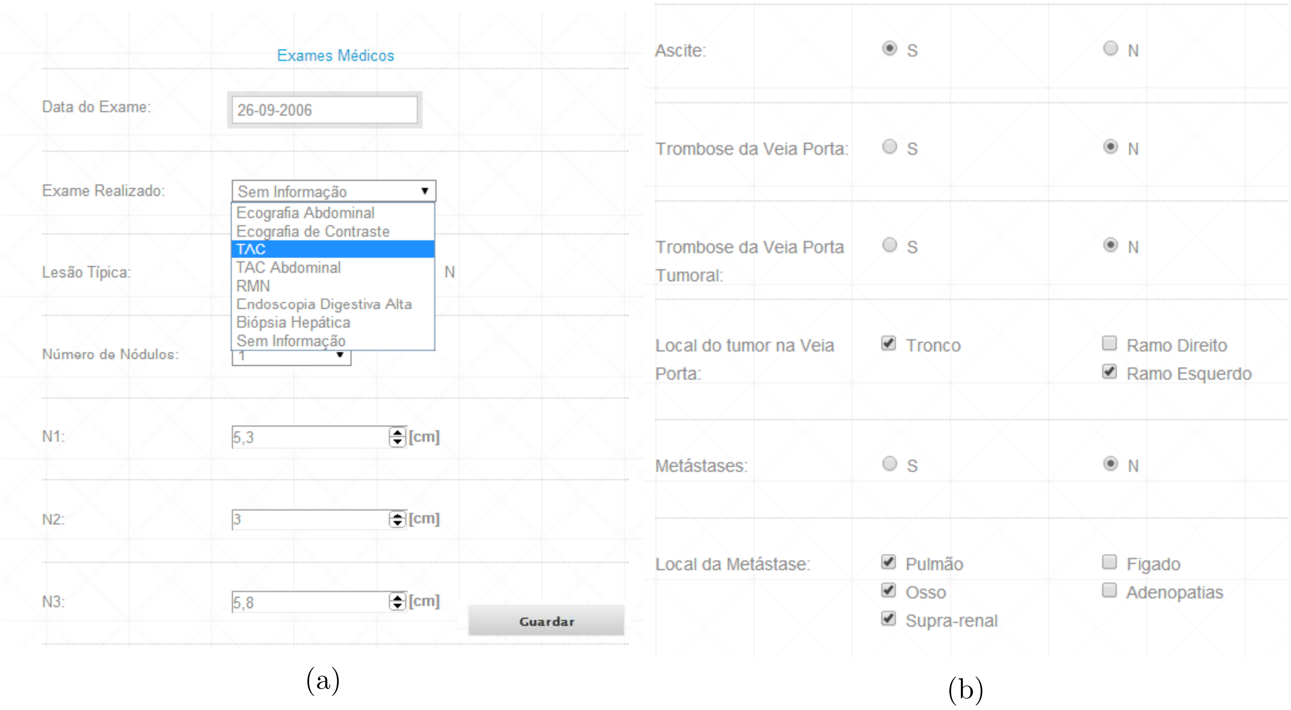
[4] The system's final version may be explored by accessing http://chucdb.dei.uc.pt/index.php



Factores de Risco

Consumo de Álcool:	● S	○ N
Gramas de álcool por dia:	17 [g/dia]	
AgHBs:	○ Positivo	● Negativo
AgHBe:	● Positivo	○ Negativo
Anti-HBc:	● Positivo	○ Negativo
Anti-HBs:	○ Positivo	● Negativo

Figure 5.10: Prototype's risk factors form.



Exames Médicos

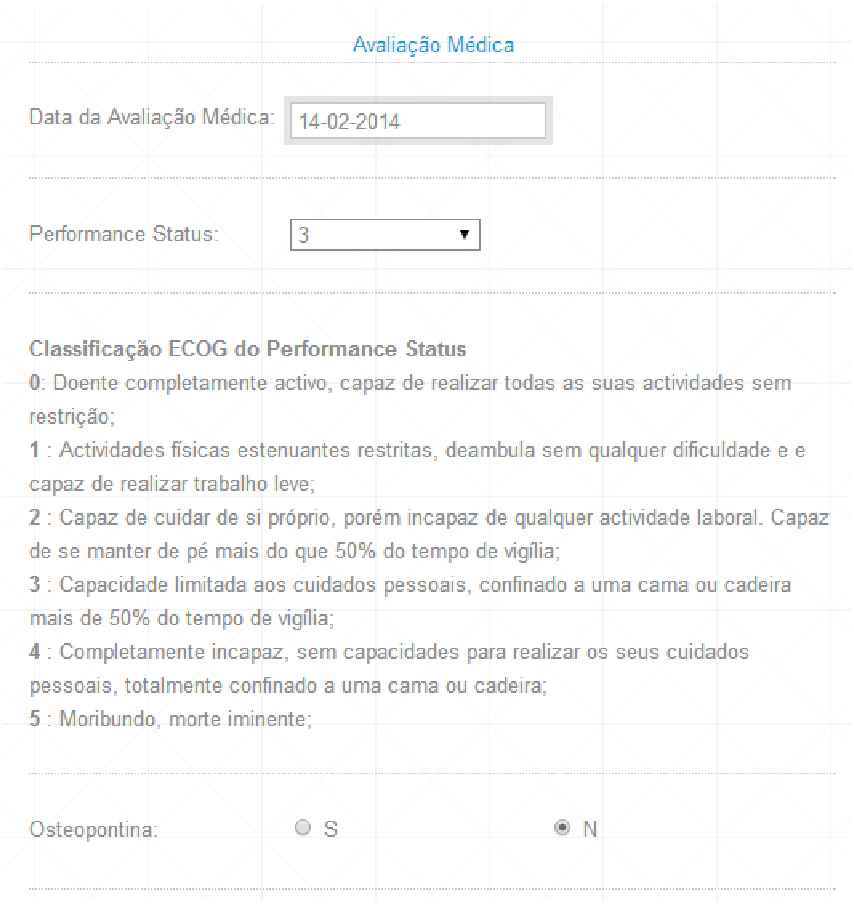
Data do Exame:	26-09-2006
Exame Realizado:	Sem Informação ▾ Ecografia Abdominal Ecografia de Contraste TAC TAC Abdominal RMN Endoscopia Digestiva Alta Biópsia Hepática Sem Informação
Lesão Típica:	N
Número de Nódulos:	1 ▾
N1:	5,3 [cm]
N2:	3 [cm]
N3:	5,8 [cm]

(a)

Ascite:	● S	○ N
Trombose da Veia Porta:	○ S	● N
Trombose da Veia Porta Tumoral:	○ S	● N
Local do tumor na Veia Porta:	☑ Tronco	☐ Ramo Direito ☑ Ramo Esquerdo
Metástases:	○ S	● N
Local da Metástase:	☑ Pulmão ☑ Osso ☑ Supra-renal	☐ Figado ☐ Adenopatias

(b)

Figure 5.11: Prototype's exams form: type of exam and findings (a) and conclusions (b).



Avaliação Médica

Data da Avaliação Médica: 14-02-2014

Performance Status: 3 ▾

Classificação ECOG do Performance Status

0: Doente completamente activo, capaz de realizar todas as suas actividades sem restrição;

1 : Actividades físicas estenuantes restritas, deambula sem qualquer dificuldade e e capaz de realizar trabalho leve;

2 : Capaz de cuidar de si próprio, porém incapaz de qualquer actividade laboral. Capaz de se manter de pé mais do que 50% do tempo de vigília;

3 : Capacidade limitada aos cuidados pessoais, confinado a uma cama ou cadeira mais de 50% do tempo de vigília;

4 : Completamente incapaz, sem capacidades para realizar os seus cuidados pessoais, totalmente confinado a uma cama ou cadeira;

5 : Moribundo, morte iminente;

Osteopontina: ○ S ● N

Figure 5.12: Prototype's medical evaluation form.

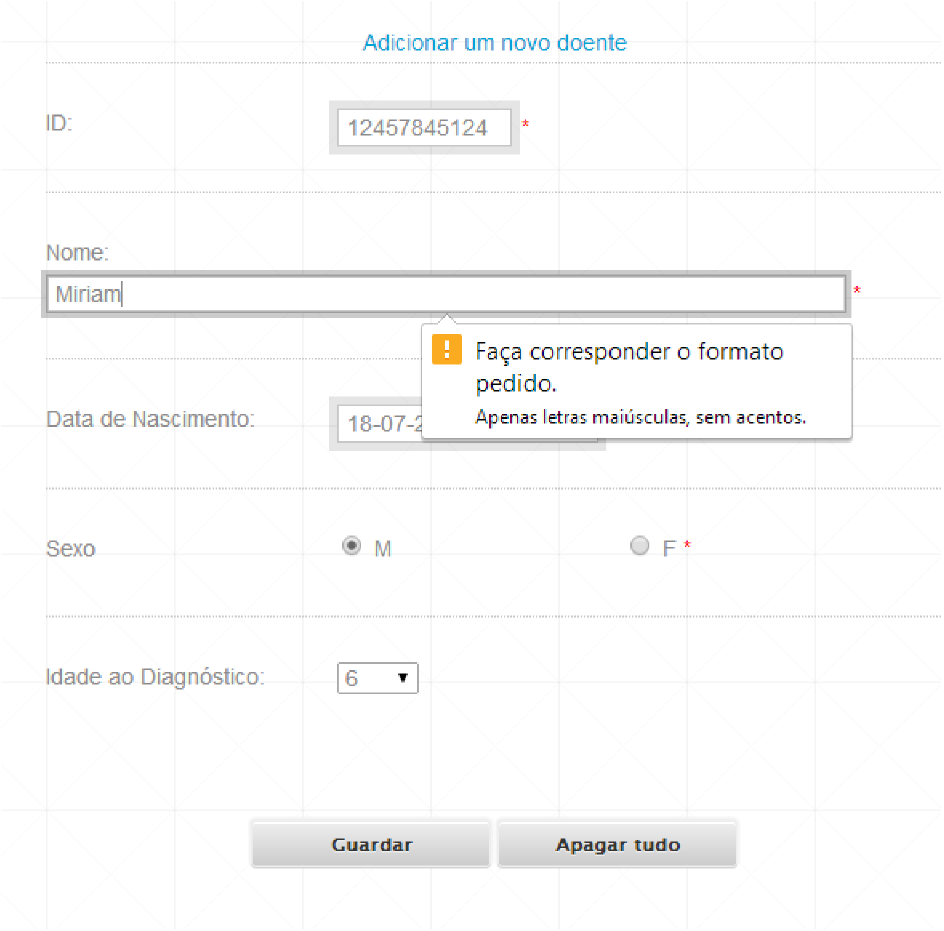
with a team composed by a biomedical and an informatics engineer. The first contact of the user with the application is via the login window (Figure 5.15).

After accessing the application, the user can choose one of the application's views. For instance, the list of patients can be accessed by clicking the "Patients" horizontal tab. Each patient's ID, Name, Date of Birth, Gender and Age at Diagnosis is shown (Figure 5.16). This triggers the presentation of filter boxes (by Name or ID), which provides an easier way of finding a particular patient. A vertical scroll bar was added to the patients' table, so that the user doesn't have to scroll the entire web page, but only the the table. There is also the option of adding new patients, risk factors, evaluations, exams, treatments or biopsies, by clicking the left, vertical menu.

As mentioned, there are several types of information that can be entered into the system. Figure 5.17 illustrates the addition of a medical evaluation. In this case, we can verify the use of the autocomplete feature, as the user starts to write the patient's name.

The user might want to examine a certain patient. After selecting the desired patient, a new page is displayed, revealing the patient's basic information (Name, Gender, Age at diagnosis and the type of diagnosis). Coupled with this, a left vertical menu is shown too, enabling the access of several subcategories of the patient's information, for instance, risk factors or medical evaluations (Figure 5.18).

In this final version, a reporting section is also included. The system allows several other types of data analysis, without the clinicians' need to understand a single line of code in order to retrieve the information they desire. A set of the most relevant questions for clinicians were specified by the CHUC's team, and pre-defined queries were written, so that the system produces the desired results at a touch of a button. The reporting section also allows the filtering



Adicionar um novo doente

ID: 12457845124 *

Nome: Miriam *

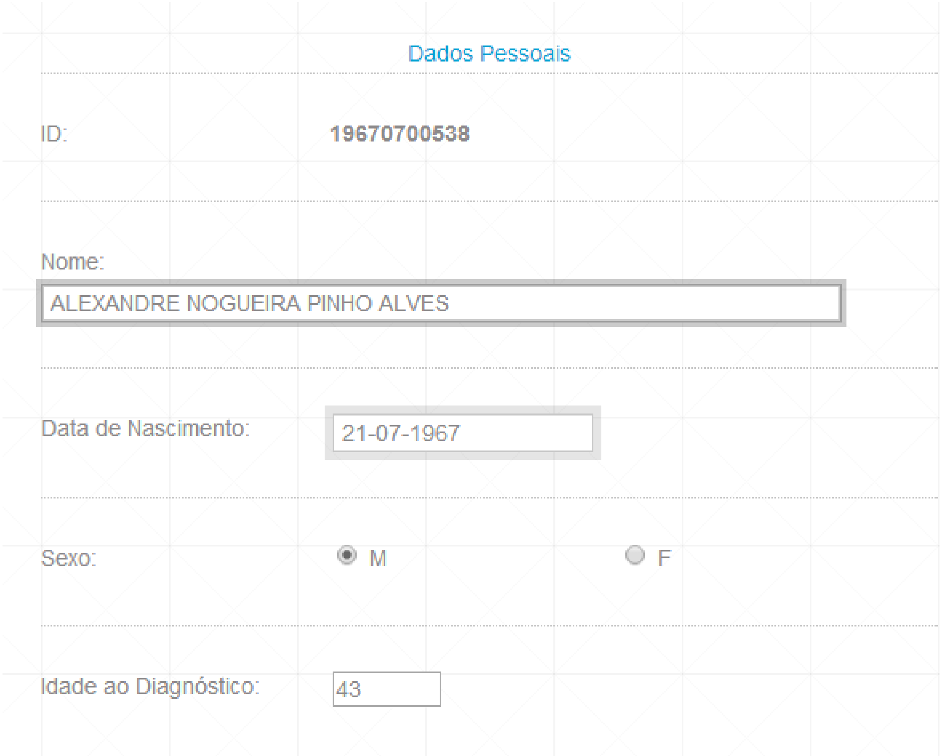
Faça corresponder o formato pedido.
Apenas letras maiúsculas, sem acentos.

Data de Nascimento: 18-07-2

Sexo ● M ○ F *

Idade ao Diagnóstico: 6 ▾

Figure 5.13: Prototype's demographics page showing a validation error for field "Name".



Dados Pessoais

ID: 19670700538

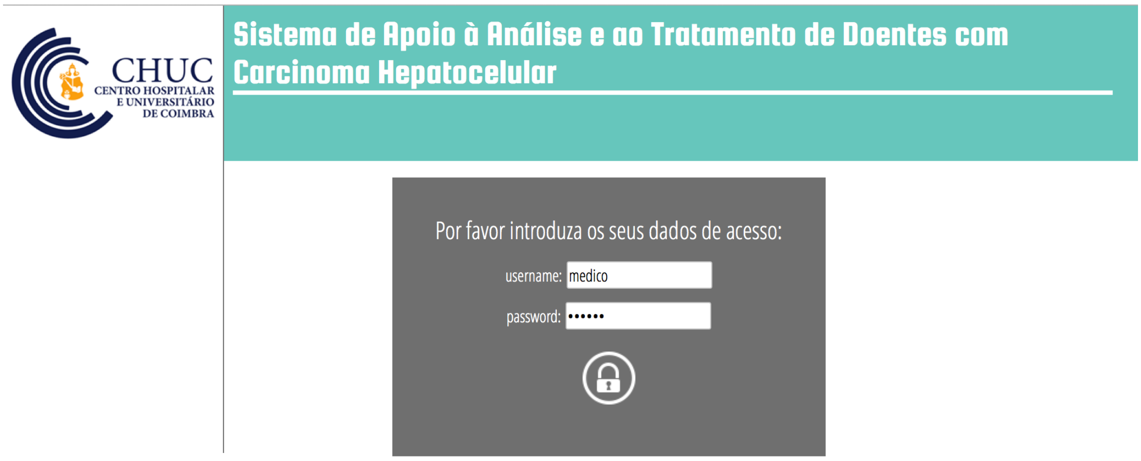
Nome: ALEXANDRE NOGUEIRA PINHO ALVES

Data de Nascimento: 21-07-1967

Sexo: ● M ○ F

Idade ao Diagnóstico: 43

Figure 5.14: Prototype's editing page.



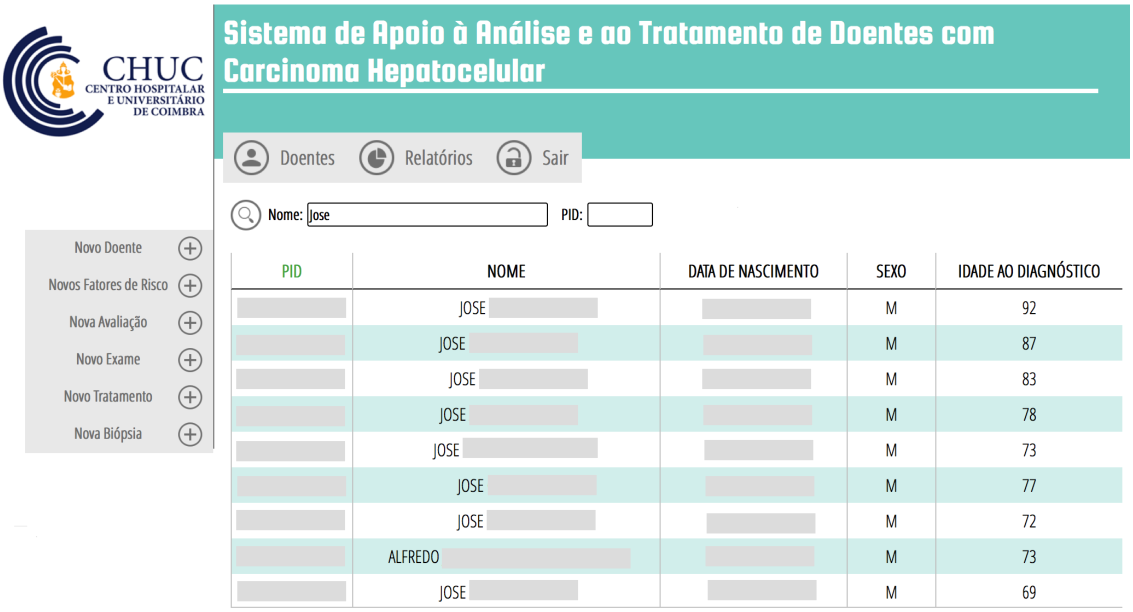
Sistema de Apoio à Análise e ao Tratamento de Doentes com Carcinoma Hepatocelular

Por favor introduza os seus dados de acesso:

username: medico

password: ••••••

Figure 5.15: Interface: login.

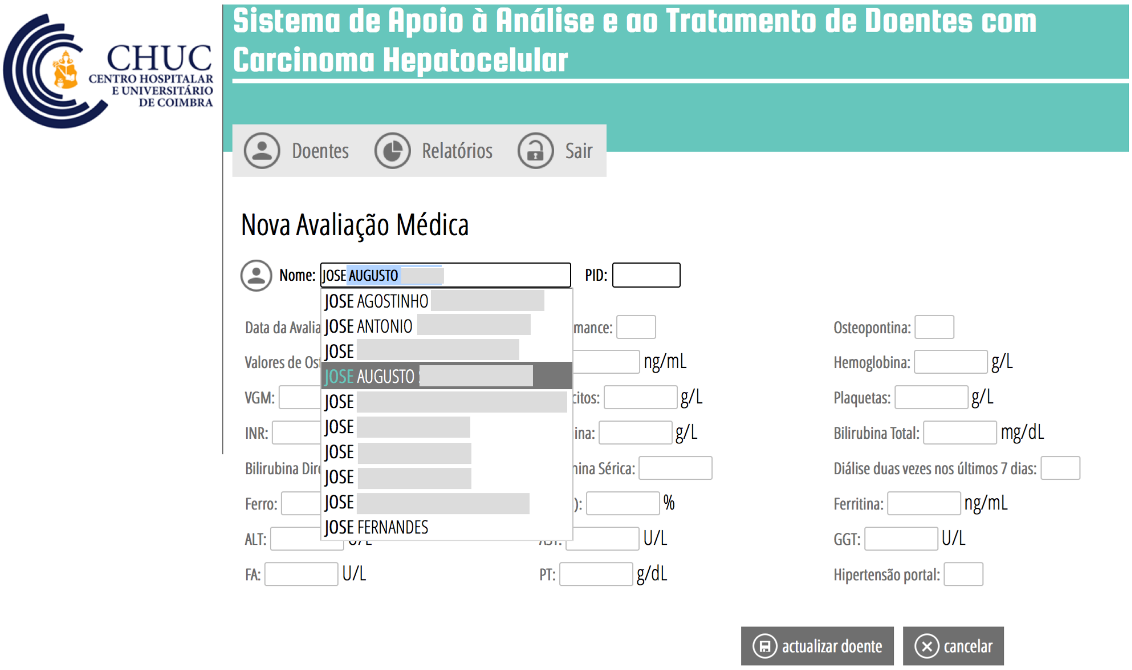


Sistema de Apoio à Análise e ao Tratamento de Doentes com Carcinoma Hepatocelular

Nome: Jose PID:

PID	NOME	DATA DE NASCIMENTO	SEXO	IDADE AO DIAGNÓSTICO
[redacted]	JOSE [redacted]	[redacted]	M	92
[redacted]	JOSE [redacted]	[redacted]	M	87
[redacted]	JOSE [redacted]	[redacted]	M	83
[redacted]	JOSE [redacted]	[redacted]	M	78
[redacted]	JOSE [redacted]	[redacted]	M	73
[redacted]	JOSE [redacted]	[redacted]	M	77
[redacted]	JOSE [redacted]	[redacted]	M	72
[redacted]	ALFREDO [redacted]	[redacted]	M	73
[redacted]	JOSE [redacted]	[redacted]	M	69

Figure 5.16: Interface: list of patients.



Sistema de Apoio à Análise e ao Tratamento de Doentes com Carcinoma Hepatocelular

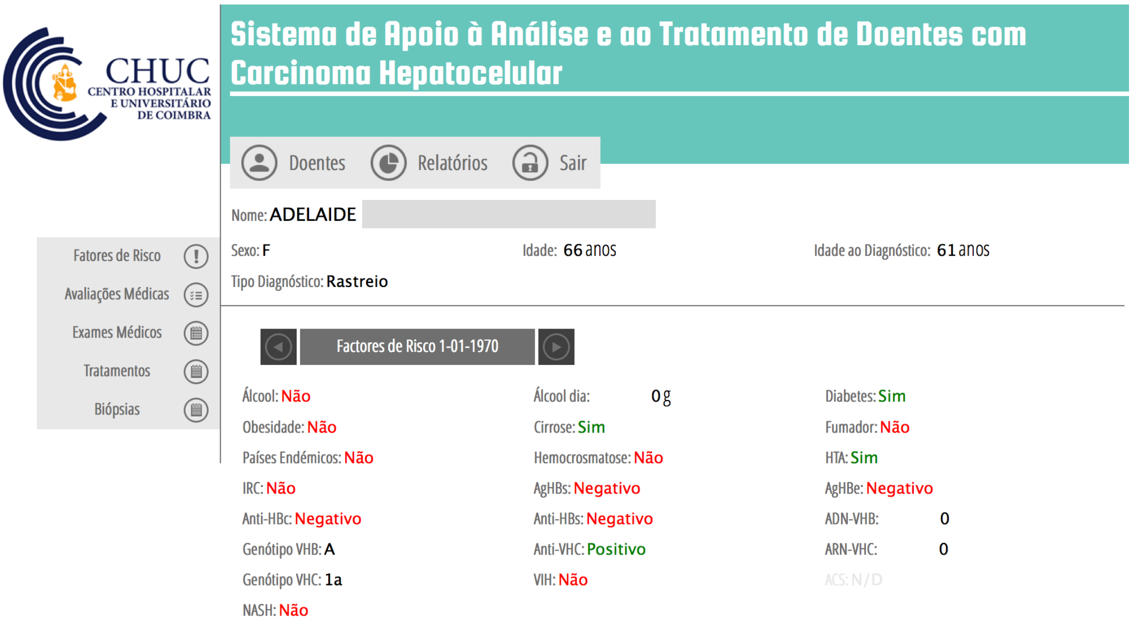
Nova Avaliação Médica

Nome: JOSE AUGUSTO [redacted] PID:

JOSE AGOSTINHO [redacted]
JOSE ANTONIO [redacted]
JOSE [redacted]
JOSE AUGUSTO [redacted]
JOSE [redacted]
JOSE [redacted]
JOSE [redacted]
JOSE [redacted]
JOSE [redacted]
JOSE FERNANDES

Data da Avalia... | ...mance: | Osteopontina:
Valores de Os... | ng/mL | Hemoglobina: g/L
VGM: | ...citos: g/L | Plaquetas: g/L
INR: | ...ina: g/L | Bilirubina Total: mg/dL
Bilirubina Dir... | ...nina Sérica: | Diálise duas vezes nos últimos 7 dias:
Ferro: | ...): % | Ferritina: ng/mL
ALT: U/L | ...: U/L | GGT: U/L
FA: U/L | PT: g/dL | Hipertensão portal:

Figure 5.17: Interface: evaluation insertion.



Sistema de Apoio à Análise e ao Tratamento de Doentes com Carcinoma Hepatocelular

Nome: ADELAIDE [redacted]

Sexo: F | Idade: 66 anos | Idade ao Diagnóstico: 61 anos

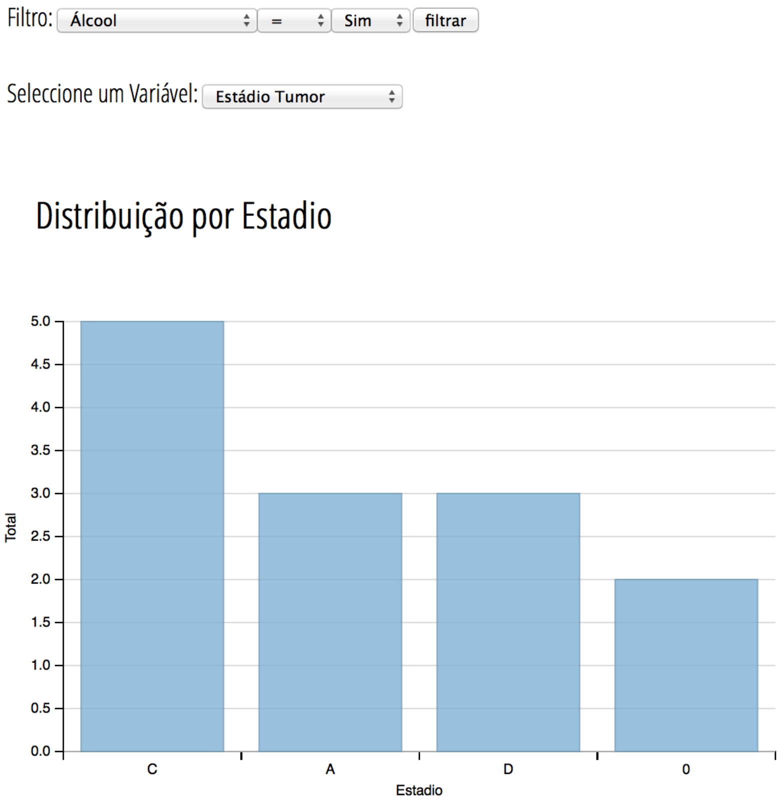
Tipo Diagnóstico: Rastreio

Factores de Risco 1-01-1970

Álcool: Não	Álcool dia: 0 g	Diabetes: Sim
Obesidade: Não	Cirrose: Sim	Fumador: Não
Países Endémicos: Não	Hemocromatose: Não	HTA: Sim
IRC: Não	AgHBs: Negativo	AgHBe: Negativo
Anti-HBc: Negativo	Anti-HBs: Negativo	ADN-VHB: 0
Genótipo VHB: A	Anti-VHC: Positivo	ARN-VHC: 0
Genótipo VHC: 1a	VIH: Não	ACS: N/D
NASH: Não		

Figure 5.18: Interface: patient visualization. Risk factors menu was selected in order to see this patient subcategory information.

of patients to be used in each analysis. For instance, the clinician might want to explore the distribution of alcoholic patients by stage of tumour (Figure 5.19). As another example, Figure 5.20 shows a Kaplan-Meier survival curve with patients' with alcohol consumption and divided by stages of tumour. The clinician may also save the presented graphics in .png or .svg formats by clicking the options "save png" or "save svg", respectively.

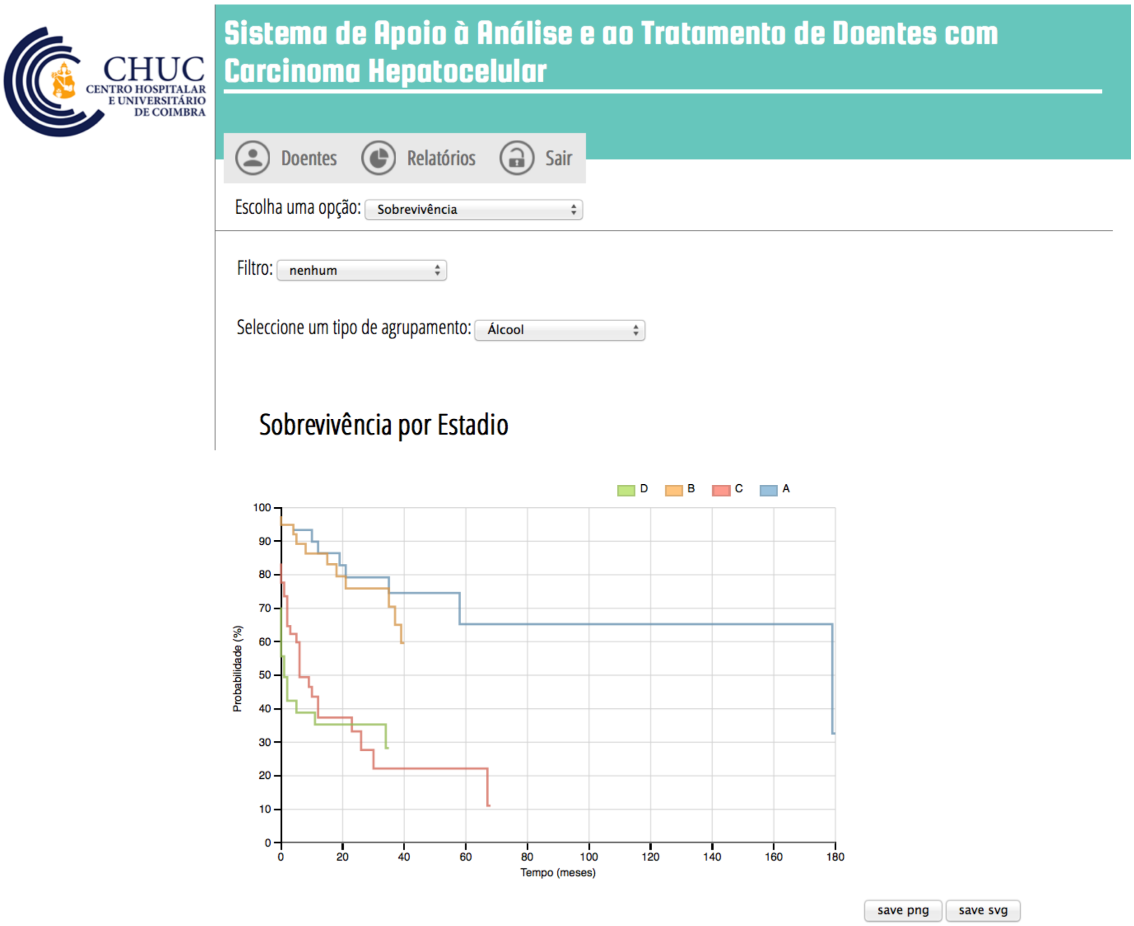


Filtro: Álcool = Sim filtrar

Seleccione um Variável: Estádio Tumor

Figure 5.19: Reporting tab: alcohol intake per tumour stage.

Data exportation is also enabled. The detailed information regarding the analysis performed is shown to the user, and he can select the complete table (clicking in the "Select table" button) and copy it to, for instance, an Excel file (Figure 5.21).



Sistema de Apoio à Análise e ao Tratamento de Doentes com Carcinoma Hepatocelular

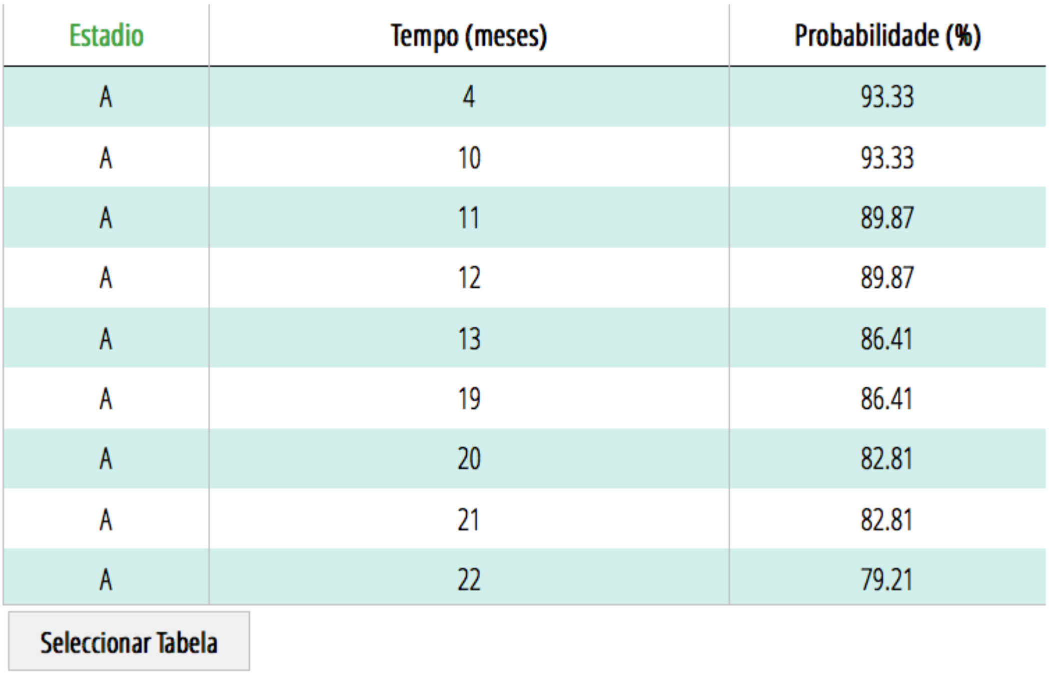
Escolha uma opção: Sobrevivência

Filtro: nenhum

Seleccione um tipo de agrupamento: Álcool

Sobrevivência por Estadio

Figure 5.20: Reporting tab: Kaplan-Meier survival curves.



Estadio	Tempo (meses)	Probabilidade (%)
A	4	93.33
A	10	93.33
A	11	89.87
A	12	89.87
A	13	86.41
A	19	86.41
A	20	82.81
A	21	82.81
A	22	79.21

Figure 5.21: Reporting tab: The Kaplan-Meier data is shown in the table, containing each patient's overall survival in months and survival probability ordered by tumour stage.

Chapter 6

Profiling Hepatocellular Carcinoma Patients

In this chapter we describe several clustering methods to profile a database of HCC patients, with heterogeneous and missing data. We have conducted various analysis (using MATLAB) to find prognostic groups with significantly different survival characteristics. Furthermore, we intended to determine whether the generated prognostic groups comprised heterogeneous populations which could be profiled by the cluster analysis. The following sections report our approaches and findings.

6.1 Risk Factors analysis

We've analysed 23 features related to HCC risk factors. Three of them (age, number of cigar packages smoked per year and alcohol intake per day) were continuous while the remaining were categorical (binary). Four features were complete: Gender, Age, Alcohol intake and Cirrhosis. The remaining all had missing values with 6 features having more than 20% of absent values, namely alcohol intake per day (55%), staying in endemic countries (23%), smoking (25%), cigar packages smoked per year (66%) and esophageal varices (31,52%). Overall, the dataset contained around 14,25% of missing values, with 153 patients having missing observations.

Though some of our dataset's features' missing rates were higher than 20%, we've decided not to discard them for several reasons. First of all, some of them can be coherently imputed according to others related to them. This is the case of cigar packages and alcohol intake per day, that may be filled according to "smoking" and "alcohol intake". That is, if a certain patient doesn't smoke, the number of cigar packages is "0". If he does smoke, the number of packages is filled with the mean of the smokers' number of cigar packages. The same for alcohol intake per day. Given the type of data (mostly categorical features), the size of our sample, and since the remaining missing features rates did not drastically exceed 20% (20%-30%), we've preferred to apply some imputation techniques to our data. Furthermore, clustering binary data is also more complex than clustering numerical data, and thus we avoided deleting features in order to keep as much information as possible.

However, we have studied the influence of the four complete feature vectors in overall survival. In brief, we tried to answer the following question: "Is it possible to model overall survival using only the complete feature vectors?". First, we have studied the correlation between the features and afterwards, applied the Multivariate Adaptive Regression Splines (MARS) as a regression analysis to model the interactions between the considered features and overall survival. Section 6.2 discusses our conclusions.

The correlation between the our dataset's complete feature vectors was analysed for feature selection. If two features are highly correlated, one of the features in the correlated pair may be discarded, since the other contains the same (or related) information. Since these four features are not all of the same type (age at diagnosis is continuous and the remaining are categorical), we had to use appropriate measures to calculate the correlation between features of different types. Table 6.1 resumes the most appropriate correlation indexes for different types of features.

Table 6.1: Appropriate correlation coefficients according to the considered pair's type of features.

	Feature 2			
Feature 1	**Interval/Ratio**	**Ordinal**	**Nominal**	**Dichotomous**
Interval/Ratio	Pearson's r_{xy}	Spearman's r_s		Point Biserial r_{pb}
Ordinal	Spearman's r_s	Spearman's r_s		Rank Biserial r_{rb}
Nominal			Contingency C Cramer's Phi ϕ_c	
Dichotomous	Point Biserial r_{pb}	Rank Biserial r_{rb}		Phi Coefficient r_ϕ

Accordingly, we have chosen the Phi Coefficient to determine the correlation between the categorical features (gender, alcohol and cirrhosis) and the Point Biserial to calculate the correlation between age and the remaining categorical features. Phi Coefficient is given by equation (6.1),

$$r_\phi = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{6.1}$$

where N is the total number of subjects, k is the minimum between the number of rows and columns and χ^2 is the Chi-squared test p-value.

The Point Biserial coefficient is calculated according to the formula 6.2:

$$r_{p.bis} = \frac{M_1 - M_0}{\sigma_t} \times \sqrt{pq} \tag{6.2}$$

M_1 = the mean score of those in one category of the dichotomised feature;
M_0 = the mean score of those scoring in the other category;
p = the proportion scoring in the first category;
q = the proportion scoring in the other category;
σ_t = the standard deviation of all scores on the continuous features;

Table 6.2 shows the respective correlation coefficients. Since the correlation between features is considerably low ($< 0,5$), none was discarded.

Table 6.2: Correlation coefficients between the complete feature vectors.

	Gender	**Alcohol**	**Cirrhosis**
Gender	-	-	-
Alcohol	0,4421	-	-
Cirrhosis	0,2537	0,4587	-
Age	0,1716	0,1624	-0,0015

6.2 Multivariate Adaptive Regression Splines

In univariate regression analysis, the relationship between a certain independent feature and the target feature is evaluated, without considering all others. Multivariate models "choose" the most suitable features for regression, using univariate analysis, and then combine them in a multivariate analysis. That is, multivariate analysis verifies the relationship between a set of features and the target features. MARS is a form of multivariate regression analysis. It can handle both continuous and categorical data, and can be used for classification or regression. In our case, we will use MARS in the regression mode, since our target feature (survival) is continuous.

MARS model pronouncedly failed to fit the data, with a coefficient of determination (R^2) of 0,277. Basically, the R^2 value is a measure of "how well" the independent features describe the target feature. MARS also determines the most appropriate number of basis functions to model the features' relations. The basis functions of our final model are:

$$\begin{aligned}
BF1 &= max(0; Age - 41)\\
BF2 &= max(0; 1 - Cirrhosis)\\
BF3 &= BF1 \times max(0; Cirrhosis)\\
BF4 &= BF2 \times max(0; 74 - Age)\\
BF5 &= max(0; Age - 67) \times max(0; 1 - Cirrhosis)
\end{aligned} \tag{6.3}$$

The final model equation is a combination of all its basis functions:

$$y = 1010,2 - 570,46 \times BF1 + 1622,9 \times BF2 + 556,86 \times BF3 - 347,99 \times BF4 + 411,56 \times BF5 \tag{6.4}$$

According to the model's basis functions, only Cirrhosis and Age are used to build the model. When the MARS model uses only these two features, R^2 rises to 0,42. Figure 6.1 shows the model built considering only Cirrhosis and Age at Diagnosis.

The complete set of features is not sufficient to create a reliable model for overall survival. This results confirmed the need to explore missing data strategies, as we initially expected.

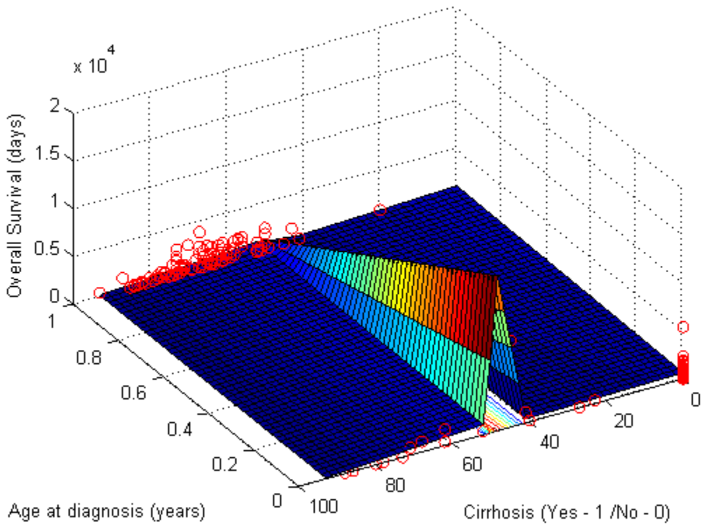


Figure 6.1: MARS model built only with Cirrhosis and Age at Diagnosis.

6.3 Missing Data imputation

For cigar packages and alcohol intake per day we've used mean imputation. Using the dataset with this imputed features, we also explored other two imputation methods: Logistic Regression (SI method) and KNN (ML imputation method).

6.3.1 Logistic Regression Imputation

Regression is mostly used to build models where the target feature is continuous. Thus, the name "Logistic Regression" is somehow misleading. Logistic Regression is used when the response is binary (0/1, Live/Die, Yes/No), and is considered a technique for classification, not regression. Logistic Regression involves a probabilistic view of classification. Overall, Logistic Regression maps a point of a multidimensional feature space to a value in the range 0 to 1, using a logistic function. The logistic model can be interpreted as a probability of class membership by applying a certain threshold to such probability. That is, the logistic models gives the class probability of a certain data point. The class assignment depends on the threshold on chooses to consider.

To impute our absent values, we've built a logistic regression model for each feature with missing values, using only the complete features as predictors. That is, each model was built with Gender, Age at Diagnosis, Alcohol intake and Cirrhosis. For each feature, we've tested several probability thresholds in a 10-fold crossvalidation. The best threshold value was chosen to impute the missing observations. Table 6.3 presents the optimal probability threshold (Optimal t), average F-measure (Avg F-measure) and error (Error F-measure) for each imputed categorical feature.

Table 6.3: Logistic Regression imputation results.

Feature	**Optimal t**	**Avg F-measure**	**Error F-measure**
3	0,6	0,6086	0,0169
6	0,7	0,9677	0
7	0,6	1	0
8	0,7	0,8099	0,0024
9	0,5	0,886	0,0883
11	0,6	0,7913	0,003
12	0,5	0,7842	0,0188
14	0,5	0,8602	0,005
15	0,6	1	0
16	0,9	1	0
17	0,7	0,8047	0,0058
18	0,9	0,9375	0
19	1	1	0
20	0,6	0,9286	$3,85 \times 10^{-17}$
21	0,5	0,7018	0,0191
22	0,5	0,8489	0,0214
23	0,5	0,7276	0,0607

6.3.2 KNN Imputation

KNN imputation requires the distances between samples to be calculated, and k nearest neighbours to decide class membership. This assumptions rise many issues in our dataset. First of all, the choice of a similarity measure that can handle both continuous and categorical features. Secondly, dealing with missing values in different features, per sample. For instance, a certain sample may have missing values in features V1 and V17, while another can have values in both of such features, but have missing observations V6 and V9. Discarding samples with missing data is impractical for us: we would only keep 12 patients. And keeping only the complete feature vectors also didn't seem the best approach. In LR, a model is built according to the feature to impute, and different thresholds can be applied. In KNN imputation, the distances between samples in the four complete feature vectors previously considered are always the same.

Here we describe a different approach. We implemented KNN in order to consider all the samples and features. We have used an distance that handles both continuous and categorical features, Heterogeneous Euclidean-Overlap Metric (HEOM), explained in more detail in the next section. In this metric, unknown values are not ignored in distance calculation. The more missing values a certain sample has, the higher its distance will be regarding all others. Usually, in KNN classification, a crossvalidation (or other sampling technique) is used in order to evaluate the model's performance, and choose k according to the best accuracy of F-measure. However, this cannot be applied to our approach. Different samples have missing data in different features, and thus, a certain k might achieve great results for one particular fold but work terribly in another. Thus, we've opted to fill the absent values according to the closest neighbour (k=1). Moreover, our objective is to keep the dataset's variability, bearing in mind that this is not an ”usual classification approach”. In order to build patients’ personalization models, homogenizing the data may not be the best choice.

6.3.3 Conclusions

The MARS model was again evaluated with LR and KNN imputation. The results are quite interesting. For LR, R^2 did not significantly changed (0,4050). This was expected, since data imputation was only based in the complete set of features, that we had already tested with MARS. However, for KNN, R^2 rose up to 0,4751. This increase in the determination coefficient indicates that our KNN imputation approach resulted in a better fitness in the overall survival method. The final model created for both imputation approaches included the same features: Age at Diagnosis, Symptoms, quantity of alcohol intake per day, HBcAb, Anti-VHC and portal hypertension. The results agree with the main HCC risk factors, presented in BCLC guidelines. KNN imputation has proven to be a better approach than LR, since it maintains as much as possible, the variations in data. Accordingly, we have proceed with a clustering analysis of our data based in KNN imputation, in order to find prognostic profiles for HCC patients.

6.3.4 Agglomerative Clustering with Heterogeneous Data

Computing distances between two examples is a crucial step for many data mining tasks. As mentioned in Section 6.3.2, distance-based algorithms, such as KNN, manage distances as a inner step. Computing the proximity between two instances on the basis of continuous data is a widely common task. A variety of functions are available for such uses, including the Euclidean, Squared-Euclidean, Minkowsky, Mahalanobis and Chebychev. However, none of these functions appropriately handle categorical input attributes. For categorical features, the simplest measure is overlap. Overlap is a similarity measure that increases proportionality according to the number of attributes in the two samples that match. Hamming and Jaccard are other two widely known functions to deal with categorical data. Heterogeneous data contain both continuous and categorical attributes. In these cases, mixed distances are the most appropriate to calculate distances between instances.

Wilson and Martinez [84] performed a detailed study of heterogeneous distance functions. The measure in their study are based upon a supervised approach where each data instance has binary information in addition with a set of continuous features. In our study, we will use their distance function, HEOM, described by equation (6.5).

$$HEOM(x,y) = \sqrt{\sum_{a=1}^{n} d_a(x_a, y_a)^2} \tag{6.5}$$

$$d_a(x,y) = \begin{cases} 1 & \text{, if } x \text{ or } y \text{ is unknown} \\ overlap(x,y) & \text{, if } a \text{ is nominal} \\ m\_diff_a(x,y) & \text{, otherwise} \end{cases}$$

$$overlap(x,y) = \begin{cases} 0 & \text{, if x = y} \\ 1 & \text{, otherwise} \end{cases}$$

$$m\_diff_a(x,y) = \frac{|x-y|}{range_a}$$

$$range_a = max_a - min_a$$

a is the *i-th* feature, in the n-dimensional feature space. x and y are feature vectors.

Overall, we have used 6 different approaches:

HEOM: Heterogeneous Euclidean-Overlap Metric, by Wilson and Martinez (equation (6.5));

HLND: Heterogeneous Linear-Nominal Distance, a heterogeneous distance function similar to HEOM, that reduces the effect of extreme values (equation (6.6));

Discretizing + Hamming distance (DH): We have coded the continuous features into *dummies* and calculated the distances between instances with the Hamming distance;

Discretizing + Jaccard distance (DJ): Discretizing the continuous features and applying the Jaccard distance;

Normalizing + Euclidean distance (NE): The continuous features were normalized in the range 0-1 and the euclidean distance was computed between instances;

Gower distance: Gower's Similarity Coefficient, described by equation (6.7). S_{ijk} is 1 - $m\_diff_{ijk}(i,j,k)$ for ordinal and continuous features, *overlap* for nominal features and Jaccard's for binary features;

$$HLND(x_a, y_a) = \begin{cases} linear(x_a, y_a) & \text{, if } a \text{ is continuous} \\ overlap(x_a, y_a) & \text{, if } a \text{ is nominal} \end{cases} \tag{6.6}$$

$$linear(x_a, y_a) = \frac{|x_a - y_a|}{4\sigma_a}$$

$$overlap(x_a, y_a)) = \begin{cases} 1 & \text{, if } x_a \neq y_a \\ 0 & \text{, if } x_a = y_a \end{cases}$$

$$S_{ij} = \frac{\sum_k^n w_{ijk} S_{ijk}}{\sum_k^n w_{ijk}} \tag{6.7}$$

where S_{ijk} denotes the contribution provided by the k-th feature, and w_{ijk} is usually 1 or 0 depending if the comparison is valid or not for the *k-th* feature.

6.3.5 Prognostic Groups

Each distance metric above was used in a hierarchical clustering in order to find different patient profiles. We are trying to find hidden structures in unlabelled data. In this approach, we have performed agglomerative clustering. Each pattern is considered as a cluster at the start of the process, and pairs of clusters are merged according to their distance. Commonly used linkage metrics include single linkage (SL), complete linkage (CL) and average linkage (AL). We have used those three and others, such as WPGMA (weighted average distance), centroid (unweighted center of mass distance) and ward's (inner squared distance between clusters). Besides an appropriate distance metric, hierarchical clustering algorithms require the desired number of clusters. Finding this optimal clustering solution is not a trivial task. In healthcare contexts, clustering solutions often depends on the clinicians' domain expertise. In our approach we have used the cophenetic correlation coefficient to compare the results of clustering data using different distance calculation methods (Table 6.4). The results were also analysed by CHUC's team, to evaluate the coherence of our conclusions.

Table 6.4: Results of the explored approaches.

Approach	**Clusters**	**Linkage**	**Cophenetic coefficient**
HEOM	2	AL	0,9475
HEOM	2	CL	0,9209
HEOM	2	WPGMA	0,9220
HLND	2	AL	0,9469
HLND	2	CL	0,9276
HLND	2	WPGMA	0,9274
HLND	4	WPGMA	0,9274
HLND	2	SL	0,9129
DH	2	AL	0,8978
DH	3	AL	0,8978
DH	3	CL	0,8630
DH	2	WPGMA	0,8483
DJ	2	AL	0,8281
DJ	2	CL	0,8765
DJ	3	CL	0,8765
DJ	3	WPGMA	0,9003
Gower	2	SL	0,9190
Gower	2	CL	0,7877
NE	4	AL	0,9275
NE	3	CL	0,9209
NE	3	ward	0,7196

Finding the most appropriate distance metric to determine HCC profiles was an iterative task, always having the validation of the CHUC's team. According to our results, 2 main profiles were found. We considered HEOM + AL as the best combination for profiling HCC patients (Figure 6.2). Table 6.5 presents our Prognostic Groups (PG) characterization.

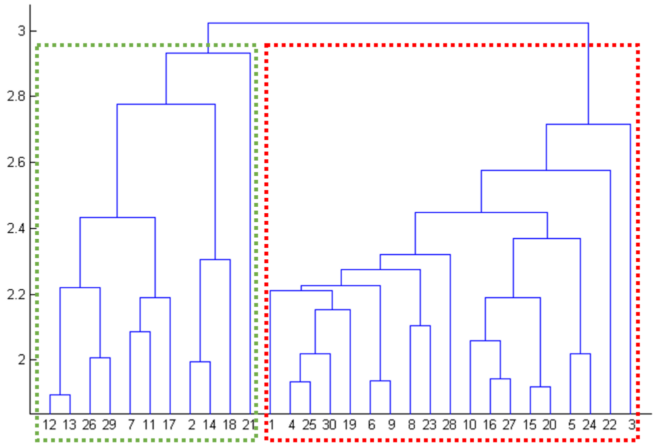
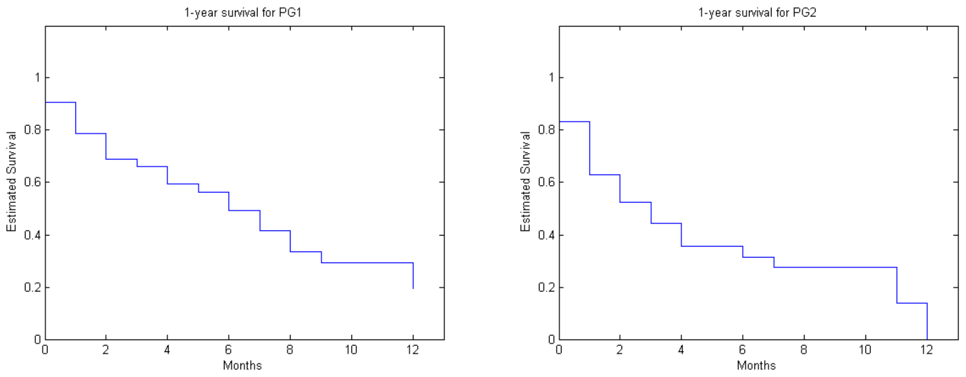


Figure 6.2: HEOM + AL dendogram showing a clear data division in two groups.

Table 6.5: Prognostic groups' characterization.

Prognostic Group	**Characterization**
Prognostic Group 1 (PG1)	PG1 patients are mostly males. Age: $58-76$ years. 60% of them have symptoms of HCC when diagnosed, 80% are alcoholic (they consume about 95 grams of alcohol per day). They are mostly HBsAg and HBcAb negative and all are HBeAg negative. Almost all of them are Anti-VHC positive. 30% have Cirrhosis. Most PG1 are smokers (they smoke about 24 cigar packs per day). Most of them are negative for Diabetes, Obesity, Hemochromatosis, HTA, IRC, HIV and NASH, but usually have esophageal varices, splenomegaly and portal hyperthension.
Prognostic Group 2 (PG2)	PG2 patients are mostly males. Age: $45-81$ years. They usually have symptomatic HCCs, and no not abuse alcohol (about 5 grams per year). They are mostly HBsAg, HBcAb and HBeAg negative. Half of them are Anti-VHC positive and they usually do not have Cirrhosis. They are light smokers (about 7 cigar packs per day). Most of them are negative for Diabetes, Obesity, Hemochromatosis, HTA, IRC, HIV and NASH esophageal varices, splenomegaly and portal hyperthension.

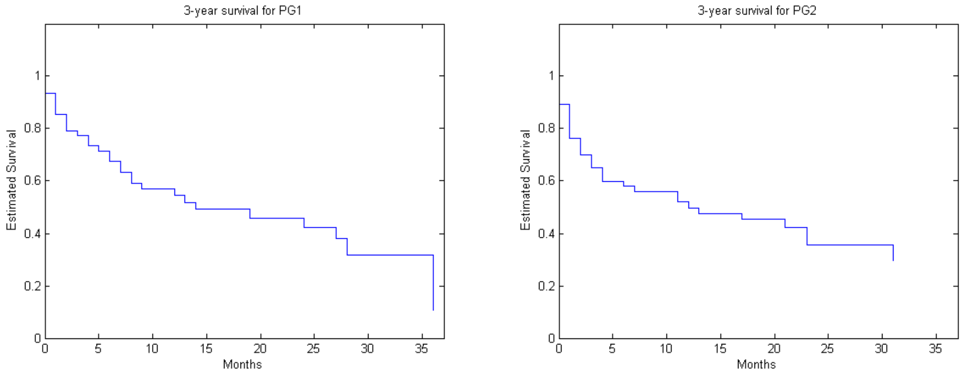
Although the groups present different characteristics, their overall survival is not significantly different, according to Mann-Withney's test ($p-value = 0,6157$). The Kaplan Meier [88] plots for 1-year survival and 3-year survival are presented in Figures 6.3 and 6.4.



(a) (b)

Figure 6.3: Kaplan-Meier survival curves for 1-year survival: prognostic group 1 (a) and prognostic group 2 (b).

According to our analysis, although a patient can be associated with a certain prognostic group, according to his risk factors characterization, this is not sufficient to make any assumptions regarding his overall survival. Heterogeneous data is more complicated to deal with than continuous data, and this unavoidably influences our analysis. The variation between patients cannot be expressed in categorical data, in particular when high percentages of missing data are present in the dataset. This led us to pursue a different approach: Clustering Laboratory Tests. Our findings are presented in Section 6.4.



(a) (b)

Figure 6.4: Kaplan Meier survival curves for 3-year survival: prognostic group 1 (a) and prognostic group 2 (b).

6.4 Laboratory Tests analysis Partitioning Clustering

Our work encompasses several segments of patient's follow-up data. To simplify, we've divided them in "clinical evaluations". Accordingly, each one of these segments contain the pathological information required to make an assessment of patient's conditions, so that an appropriate treatment could be applied. That is, each "clinical evaluation" occurs before the patient engages a new stage of treatment. As they advance in treatment, some patients die. Consequently, the number of available patients to study decreases as clinical evaluations progress, thus reducing our statistical power as regards survival prediction.

The first clinical evaluation is composed of 23 clinical features (heterogenous data). Three features are ordered, while the remaining are numeric. All of these features contained missing values. Particularly, 4 of them contained more than 20% of missing values, causing the dataset to have over 10% of missing values, with 116 patients having missing information in their records. Therefore, these 4 pronouncedly incomplete features were removed from the study. This procedure resulted in a considerable decrease of the dataset's missing data percentage, becoming about 3%, with only 42 patients having absent observations in some features.

The characteristics of this dataset substantially simplify our personalization studies. Ordered features may be converted to numeric, transforming the ordered attributes in numeric while preserving their natural order. The three ordered features correspond to required features for Performace Status (PS) and Child Pugh's (CP) classification, and thus they are codified accordingly to their respective scores for PS and CP calculation.

Based on the final 19 considered features, two clustering algorithms were used - k-means and Partition Around Medoids (PAM). Unlike the previous clustering, there is no need to use a mixed distance to compute similarity between the individuals. We can take advantage of well-known similarity measures such as Euclidean or Cityblock distances.

6.4.1 Data Preprocessing

Good data preparation is the key to produce valid and reliable models. Normalization is one of the steps often performed in data preprocessing, when dealing with numeric features. Normalization allows more robust comparisons of distances between samples or subjects, since the differences in the ranges of the features are minimized. There are several types of normalization approaches. We chose to compare z-score with min-max normalization, given by equations (6.8) and (6.9), respectively.

$$z_i = \frac{x_i - \mu}{\sigma} \tag{6.8}$$

$$y_i = \frac{x_i - min_a}{max_a - min_a} \tag{6.9}$$

In equation (6.8), z is the standard score, μ is the mean of all samples for a certain feature and σ is the standard deviation of such feature as well. Equation (6.9) fits each data point in a specific range: between the maximum (max_a) and minimum (min_a) of a given feature a.

6.4.2 k-means results

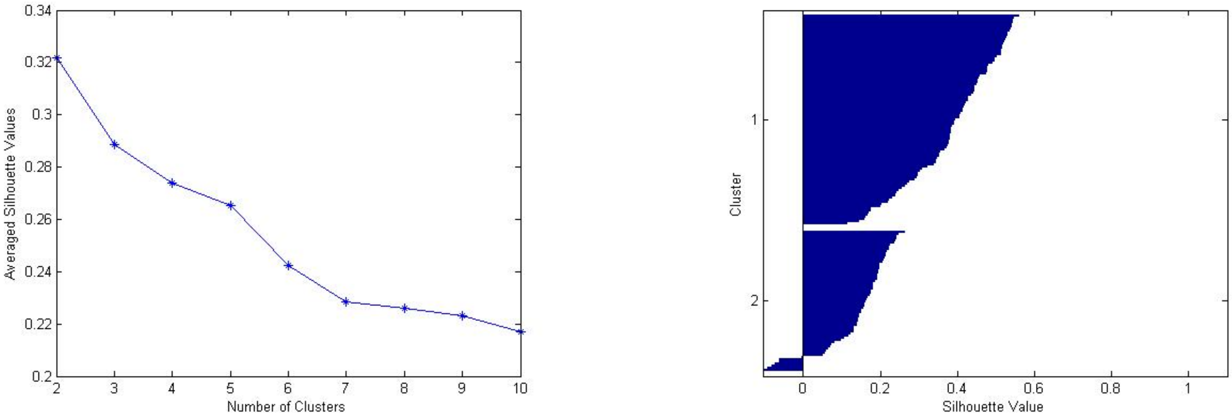
We performed 50 runs of k-means, where the initial centroids were randomly chosen, and considering several distance metrics. The number of clusters is not known prior to the algorithms implementation, as thus a clustering validity index may be used to find the optimal number of clusters for the dataset. The algorithm was run between 2 and 10 clusters to achieve the optimal k, 50 times. After each iteration, Silhouette values were computed in order to assess the group distribution. The best Silhouette values were obtain by min-max normalization, considering the Squared Euclidean distance for both k-means distance and Silhouette's dissimilarity computation (Table 6.6).

Table 6.6: Best average Silhouette results after 50 runs of k-means clustering for each of the considered combinations of clustering metrics and Silhouette's inter-point distances.

k-means	**Silhouette**	**Number**	**Averaged**
distance metric	**inter-point distance**	**of Clusters**	**Silhouette Results**
sqEuclidean	Euclidean	2	0,1927
sqEuclidean	sqEuclidean	2	0,3220
sqEuclidean	cityblock	2	0,2085
cityblock	Euclidean	2	0,1671
cityblock	sqEuclidean	2	0,2691
cityblock	cityblock	2	0,1899

As can be seen, Silhouette gives the best results when 2 clusters are considered. On further inspection, we computed the Silhouette plot to visually evaluate cluster assessment for squared euclidean distance and for 2 clusters in particular (Figure 6.5).

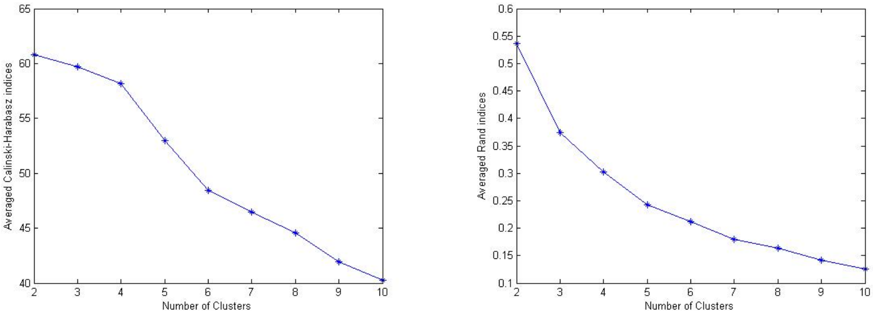
Besides Silhouette values, two other clustering validation indices were explored, namely Calinski and Rand index. Again, 50 k-means iterations were performed, for k ranging from 2 to 10 clusters. The optimum number of clusters estimated by each index strengthens our previous conclusions, as can be seen in Figure 6.6.



(a) (b)

Figure 6.5: Visual evaluation of Silhouette results: (a) Silhouette values ranging 2 to 10 cluster, considering sqEuclidean distance for both k-means and as Silhouette inter-point distance. (b) Silhouette plot for $k = 2$ clusters, considering sqEuclidean distance for both k-means and Silhouette.

Figure 6.6: Validity indices calculated for k-means: (a) Calinski index and (b) Rand index.



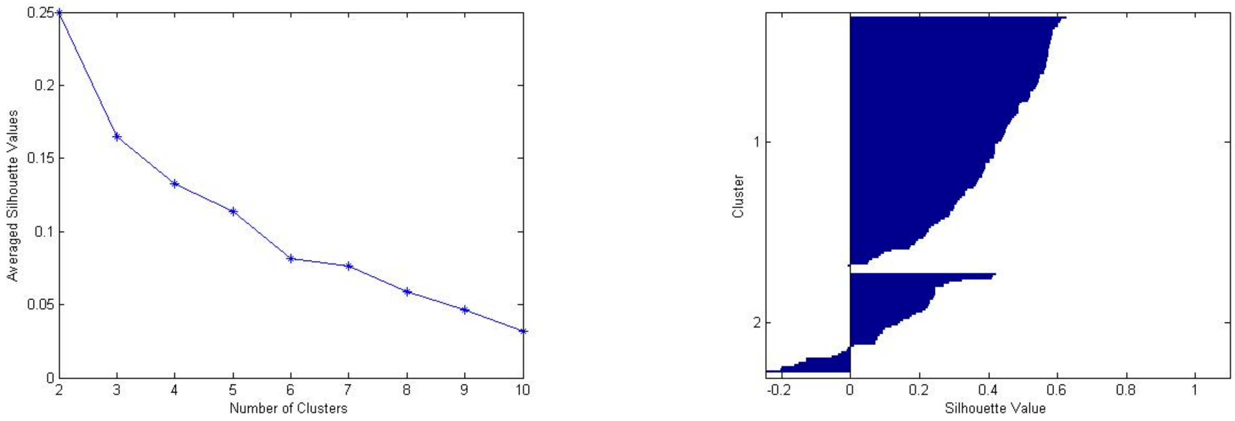
(a) (b)

6.4.3 PAM results

With the same procedure (50 iterations for k ranging between 2 to 10 clusters and using several distance metrics), PAM algorithm was also run. Again, Silhouette values were inspected to evaluate cluster assessment. Table 6.7 and Figures 6.7 and 6.8 resume our conclusions: 2 is the appropriate number of clusters for this dataset.

Table 6.7: Best average Silhouette results after 50 runs of PAM clustering for each of the considered combinations of clustering metrics and Silhouette's inter-point distances.

PAM distance metric	Silhouette inter-point distance	Number of Clusters	Averaged Silhouette Results
seuclidean	Euclidean	2	0,1493
seuclidean	sqEuclidean	2	0,2499
seuclidean	cityblock	2	0,1845
cityblock	Euclidean	2	0,1566
cityblock	sqEuclidean	2	0,2269
cityblock	cityblock	2	0,1563

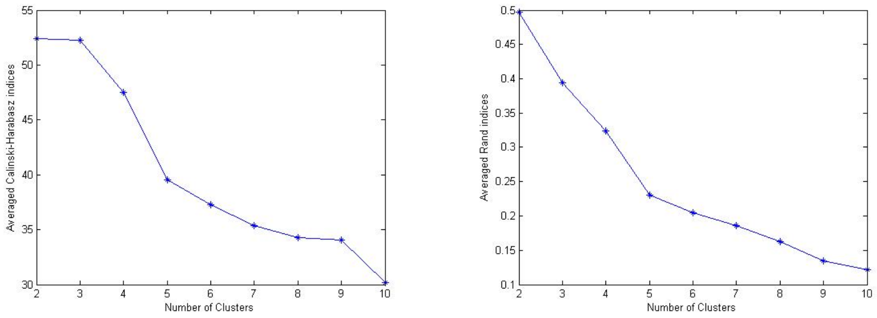


(a) (b)

Figure 6.7: Visual evaluation of Silhouette results: (a) Silhouette values ranging 2 to 10 cluster, considering sqEuclidean distance for both PAM and Silhouette inter-point distance. (b) Silhouette plot for $k = 2$ clusters, considering sqEuclidean distance for both PAM and Silhouette.

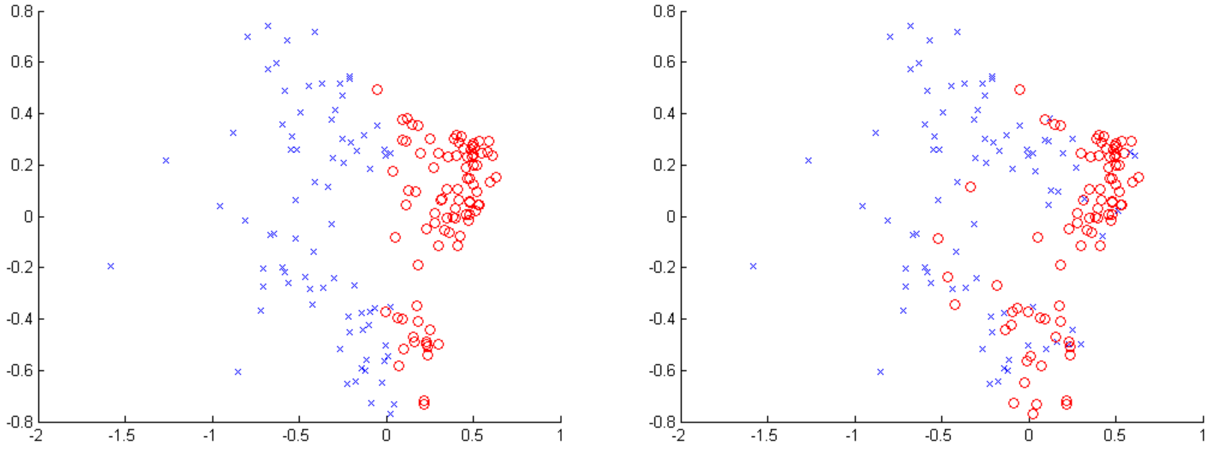
6.4.4 Principal Components Analysis (PCA)

To enable visualization, the original data space (19 features) was transformed by principal component analysis (PCA), and the points were plotted at their projected position against the two (and three) principal components axes (Figures 6.9 and 6.10). Such a plot allows the visualization of the clusters, that are "spread out" as much as possible according to the components considered.



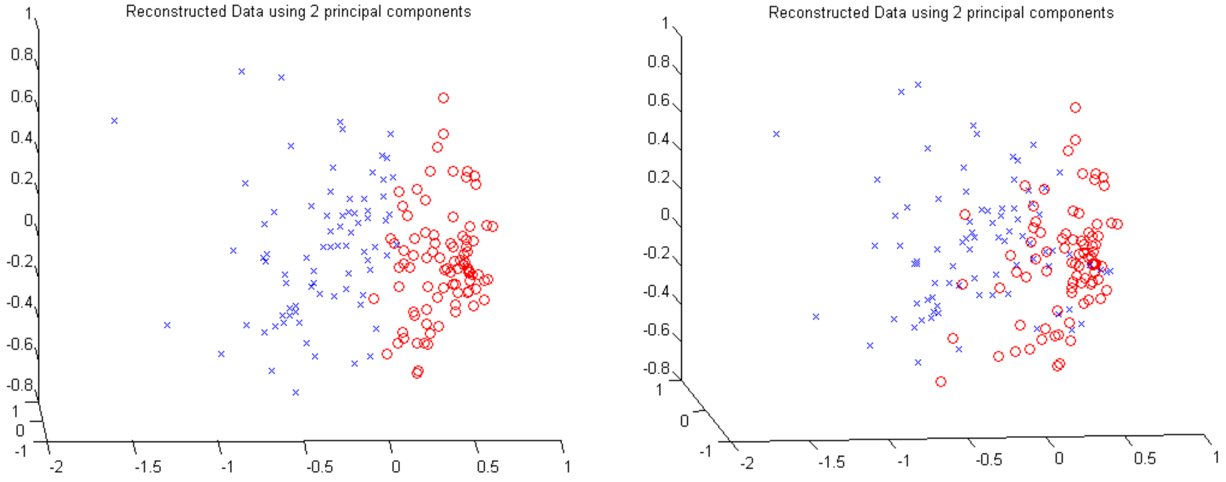
(a) (b)

Figure 6.8: Validity indices calculated for PAM: (a) Calinski index and (b) Rand index.



(a) (b)

Figure 6.9: Biplots of clusters projected on the first and second principal component axes for (a) k-means clustering and (b) PAM clustering.



(a) (b)

Figure 6.10: Plots of clusters projected on the first, second and third principal component axes for (a) k-means clustering and (b) PAM clustering.

From these plots it can be seen that both methods split the clusters similarly: one in the left side of the plot and the other on the right. However, k-means clusters are more compact and better separated than PAM's clusters, which is in agreement with the results of the cluster validation indexes, where k-means indices are higher. It is important to state that we should choose our clustering method based on the validation results rather than the PCA visualization, given that the two/three principal components do not retain enough information about the data, as shown by Table 6.8.

Table 6.8: PCA results. The first two components only retain about 39% of the information, while the first three retain about 50%.

Component	**Eigenvalues**	**Cumulative Variance Percentage** (%)
1	0,1838	22,9234
2	0,1291	39,0319
3	0,0873	49,9259
4	0,0707	58,7431
5	0,0496	64,9344
6	0,0464	70,7198
7	0,0412	75,8625
8	0,0332	80,0089
9	0,0285	93,5599
10	0,0257	86,7605
11	0,0216	89,4535
12	0,0172	91,5945
13	0,0163	93,6245
14	0,0129	95,2309
15	0,012	96,7302
16	0,0093	97,8891
17	0,0084	98,9404
18	0,0049	99,5556
19	0,0036	100

According to the Kaiser criterion [74], the components with eigenvalues above 1 should be kept. This is impracticable in our case, since none of them is above such value. Scree Test [74] suggests discarding the eigenvalues starting where the Scree plot levels off, which in this case, would amount to retain all the eigenvalues (Figure 6.11).

From our experiments we found a slight difference between the this two similar methods, k-means and PAM. Looking at the validity indices values, we found that k-means suggested a clear classification in two groups, although theoretically PAM is a more robust method. Thus, we have chosen k-means as our clustering approach for further work.

6.5 Clusters characterization

k-means clustering was performed 2000 times to assess group assignment for each data point. The resulted in a division into two groups, including 78 patients in Group 1 (G1) and 87 in

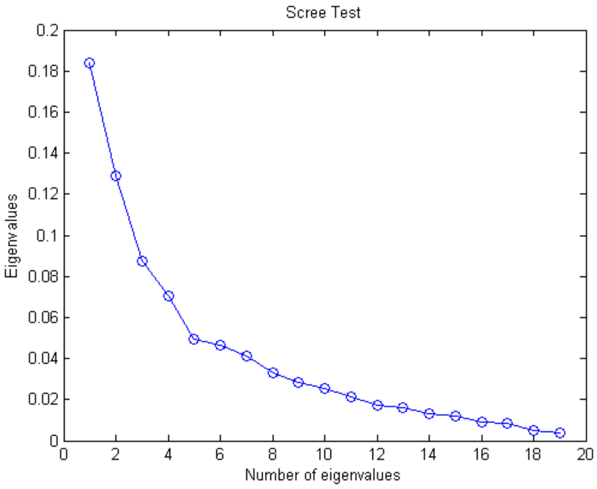


Figure 6.11: Scree Plot: plot of the eigenvalues for our Laboratory Test features.

Group 2 (G2). It is important to examine whether the overall survival in this two groups is statistically significant. In order to do so, the overall survival was subjected to some statistical tests. First of all, it is essential to know if overall survival (our dependent feature) is normally distributed. If so, parametric tests can be applied. On the contrary, if the feature does not meet the normality criterion, it can only be applied non-parametric tests. According to the Kolmogorov-Smirnov test [86], the overall survival is not normally distributed at an $\alpha = 0,05\%$ significance level, with $p\text{-}value = 6,2394 \times 10^{-9}$. For visual assessment, the histogram of overall survival and its empirical cumulative distribution function (ecdf) were plotted, as shown in Figure 6.12.

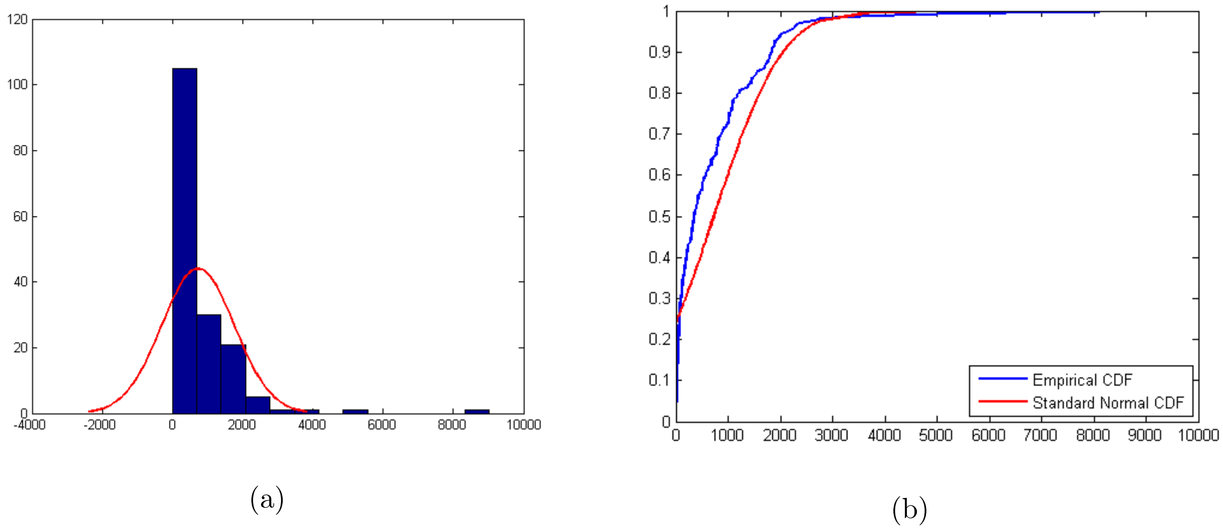


Figure 6.12: Histogram of overall survival, in days (a) and a plot of overall survival ecdf against a normal cumulative distribution function around the same mean and standard deviation.

Since the overall survival does not follow the normal distribution, the most appropriate test to be applied is Wilcoxon-Mann-Whitney's [86], to see if the two clusters shown statistically significant differences in overall survival (Figure 6.13 and Table 6.9). According to this test,

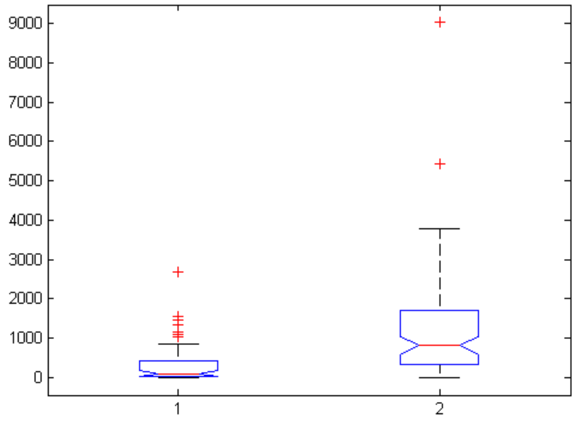


Figure 6.13: Overall survival box-plot for both groups.

there are significant differences in the overall survival of these two groups, with *p-value* $= 8,2050 \times 10^{-11}$.

Table 6.9: Mean and Standard deviation of the both groups.

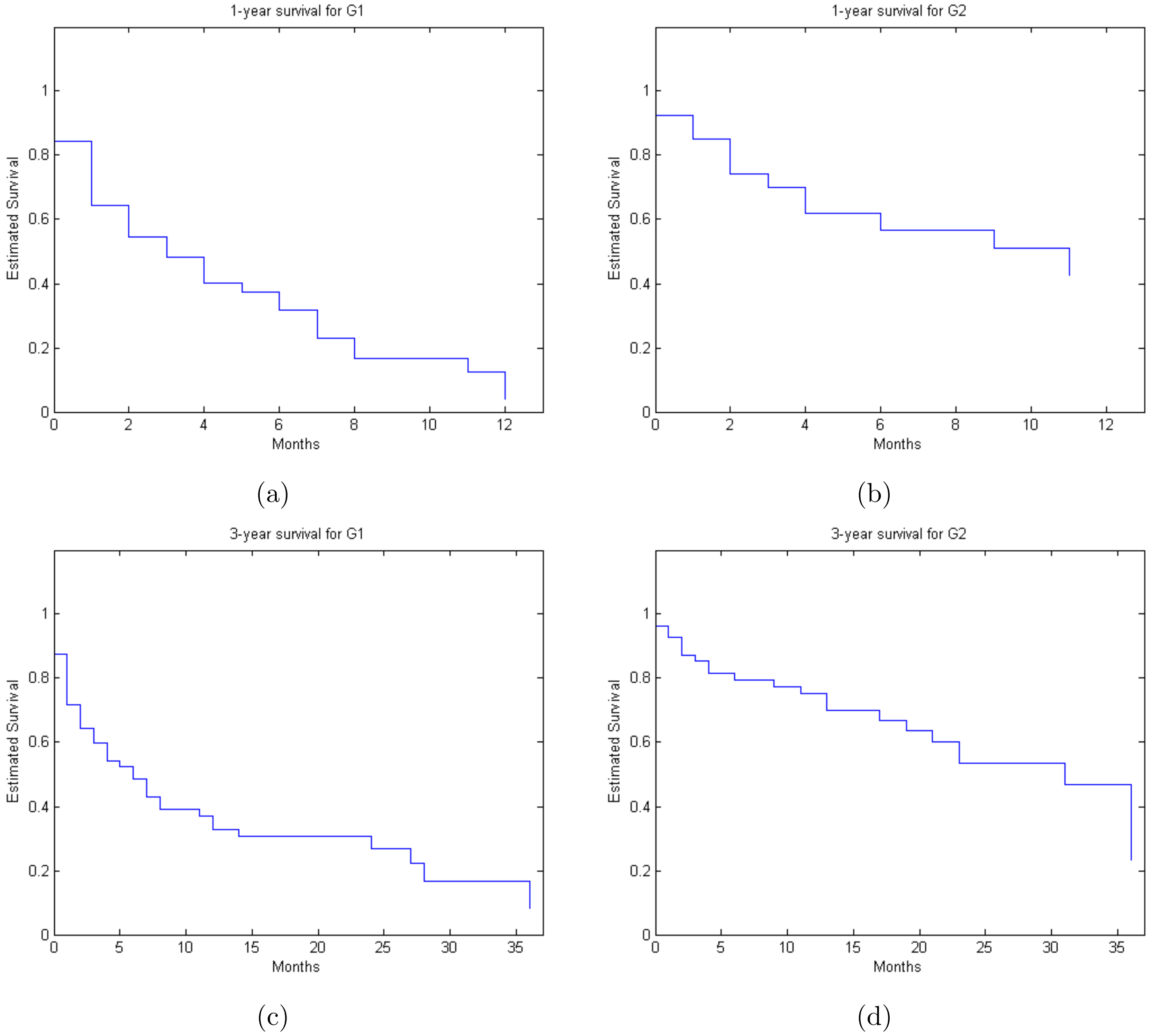
	Mean (days)	**Standard Deviation (days)**
Group 1	312,7	464,8
Group 2	1096,4	1252,3

The Kaplan-Meier curves for 1-year survival and 3-years survival for both groups are shown in Figure 6.14.

It's easily perceived that the groups show a substantial difference at both 1-year and 3-years survival estimates. Group 1 generally has a lower probability of survival than Group 2, when the same intervals are considered. For instance, regarding the 6 month period in the 1-year survival interval. The probability that patients in Group 1 live more than 6 months is about 37% while in patients of Group 2, the same probability rises to 57%. Another example would be to consider the time of survival higher that 30 months (3-year survival curve): patients in Group 1 have less than 20% estimated probability of survival, while patients in Group 2 have an estimated probability of survival over 55%. This can be explained relating the groups to the tumour stages. In fact, as explained by Tables 6.10 and 6.11, G1 includes almost every patient in terminal stage (D), and a good percentage of patients in advanced stage (C). In turn G2 consists mostly in patients in early stage (A) and intermediate stage (B), despite having some cases of stage C. As stated in the BCLC guidelines, stages A and B are expected to have a greater survival, since patients are in early stages of the disease. Thus, the results agree with the expected ones, considering tumour staging.

Table 6.10: Distribution of tumour stages present in G1.

Tumour Stage	**Number of Patients**	**Percentage (%)**
A	1	1,30
B	6	7,79
C	33	42,86
D	36	46,75



(a) (b)

(c) (d)

Figure 6.14: Kaplan-Meier curves for both groups at 1-year survival - Group 1 (a) and Group 2 (b) and at 3-year survival - Group 1 (c) and Group 2 (d).

Table 6.11: Distribution of tumour stages present in G2.

tumour Stage	Number of Patients	Percentage (%)
A	28	33,73
B	33	39,76
C	20	24,10
D	1	1,20

In recent researches, it has been suggested that the BCLC intermediate stage (BCLC-B) should be further divided, since its definition is rather broad and includes a heterogeneous patient population according to tumour extension and liver function [87]. Our results suggest that there is also some heterogeneity in BCLC-C patients.

After concluded that the overall survival is different between the achieved groups, it is fundamental to carry out a detailed examination of how these clusters relate to clinical factors. Thus, we conducted Kolmogorov-Smirnov tests for all the considered 19 features, applying the t-student test to those that followed the normal distribution and the Wilcoxon-Mann-Whitney's test for those which did not. The results are presented in Table 6.12.

Table 6.12: Kolmogorov-Smirnov test for the dataset features.

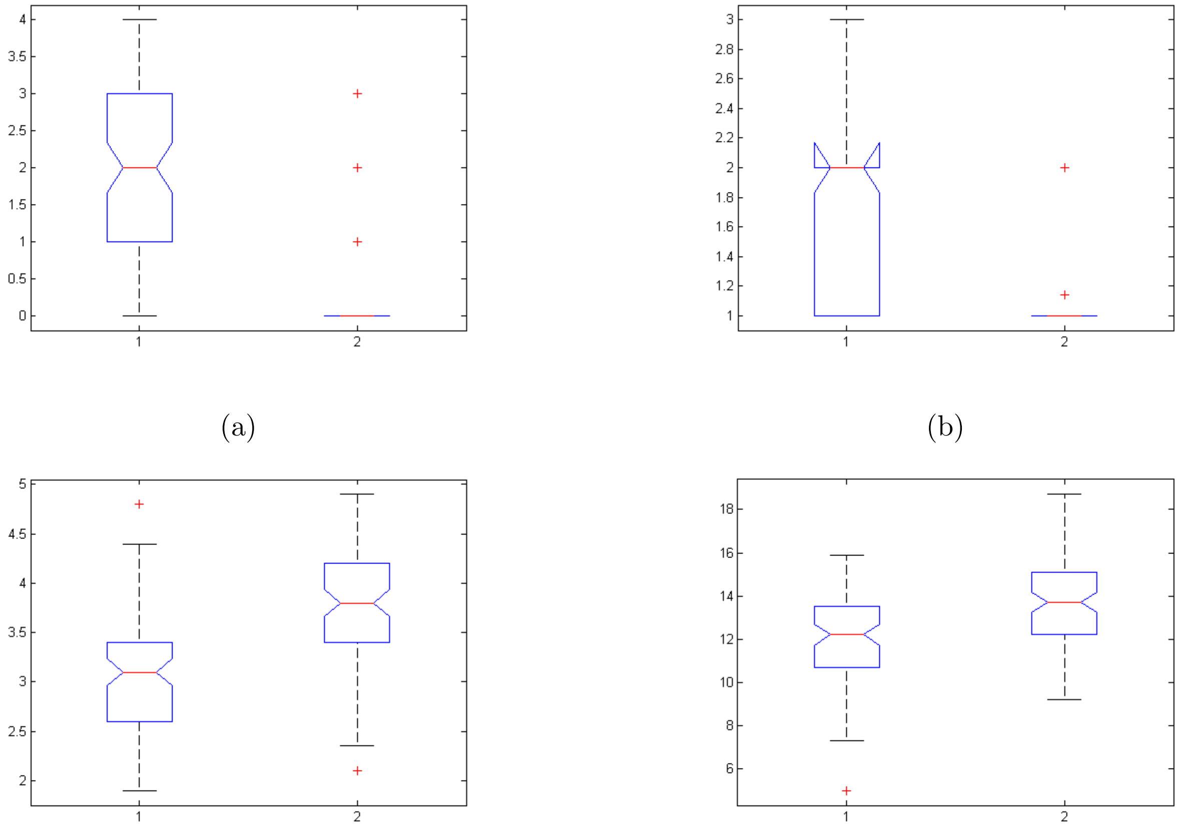
Feature	**Kolmogorov-Smirnov** (*p-value*)
PS (Performance Status)	$8,7669 \times 10^{-13}$
Encephalopathy	$1,2751 \times 10^{-37}$
Ascites	$7,6678 \times 10^{-24}$
INR (Renal Impairement)	$1,1051 \times 10^{-4}$
AFP (Alpha Fetoprotein)	$1,3025 \times 10^{-29}$
Hemoglobin	$6,0014 \times 10^{-1}$
VGM (Average Globular Volume)	$6,6328 \times 10^{-1}$
Leukocytes	$7,2312 \times 10^{-29}$
Platelets	$2,2035 \times 10^{-3}$
Albumin	$2,9314 \times 10^{-1}$
Total Bilirubin	$7,7710 \times 10^{-14}$
ALT (Alanine Amino-Transferase)	$9,4971 \times 10^{-6}$
AST (Aspartate Amino-Transferase)	$4,0045 \times 10^{-7}$
GGT (Gamma Glutamyl-Transferase)	$4,6849 \times 10^{-5}$
FA (Alkaline Phosphatase)	$1,0671 \times 10^{-5}$
PT (Total Proteins)	$2,5283 \times 10^{-27}$
Creatinine	$1,6011 \times 10^{-11}$
Number of Nodules	$1,5794 \times 10^{-9}$
Major Dimension	$2,6241 \times 10^{-3}$

As can be seen, Hemoglobin, VGM and Albumin fail to reject the null hypothesis that the feature comes from a normal distribution. So, for these features, t-student test is the most correct in order to perceive if they are good features to distinguish between the two groups (Table 6.13).

Table 6.13: Mann-Whitney's and t-student's test results for the 19 considered features.

Feature	**Wilcoxon-Mann-Whitney** (*p-value*)	**t-student** (*p-value*)
PS	$1,3332 \times 10^{-20}$	-
Encephalopathy	$1,0000 \times 10^{-3}$	-
Ascites	$2,6340 \times 10^{-15}$	-
INR	$4,3000 \times 10^{-3}$	-
AFP	$7,0000 \times 10^{-4}$	-
Hemoglobin	-	$1,1002 \times 10^{-7}$
VGM	-	$6,0060 \times 10^{-1}$
Leucocytes	$1,1120 \times 10^{-1}$	-
Platelets	$6,5830 \times 10^{-1}$	-
Albumin	-	$1,1143 \times 10^{-12}$
Total Bil.	$3,2602 \times 10^{-5}$	-
ALT	$5,0650 \times 10^{-1}$	-
AST	$2,9000 \times 10^{-3}$	-
GGT	$4,7000 \times 10^{-3}$	-
FA	$6,9192 \times 10^{-6}$	-
PT	$8,9798 \times 10^{-6}$	-
Creatinine	$2,5720 \times 10^{-1}$	-
Number of Nodules	$2,7707 \times 10^{-5}$	-
Major Dimension	$1,4800 \times 10^{-2}$	-

Smaller *p-values* indicate higher discriminative power. According to Mann-Whitney's and t-student's test results, PS, Ascites, Albumin and Hemoglobin are the most significant features to distinguish between G1 and G2 (Figure 6.15).



(a) (b)

(c) (d)

Figure 6.15: Box-plots for the four most discriminative features, namely PS (a), Ascites (b), Albumin (c) and Hemoglobin (d).

These results are in accordance with the BCLC staging system (Section 2.2). Regarding PS, stages A-C are classified as those ranging from 0-2. It is important to notice that stage A and B have PS 0, while C has PS 1 or 2 and D has PS 2 or higher. This is an interesting observation since it may the division of stage C patients in the two groups. Ascites and Albumin are two of the factors considered in Child Pugh's score (Section 2.1.2) calculation, which along with PS defines the patients stage of cancer. A-C stages have CP - A or B (in terms of score), while stage D includes the patients with CP - C. Again, this shows that the staging criteria may not consider the heterogeneity present in patients in the same stage. The "advantage" in dealing with the "raw features", so to speak, is that we are able to study the impact of such features, rather than study only those already used to define the BCLC staging system. Hemoglobin is one of such features. Anemia is a common complication of chronic liver diseases, and a frequent side effect associated with cancer. Normal values range from 12-18 mg/dL, and as we can see from Figure 6.15 (d), G2 has lower ranges, as seems logical.

Our clustering results suggest that stage C patients are somehow heterogeneous. Similarly to the study conducted in the previous section, we've examined the features that might explain the reason why these patients have been placed in different groups. Tables 6.14 and 6.15 show the Kolmogorov-Smirnov's and Mann-Whitney's or t-student's tests according to the criteria applied above.

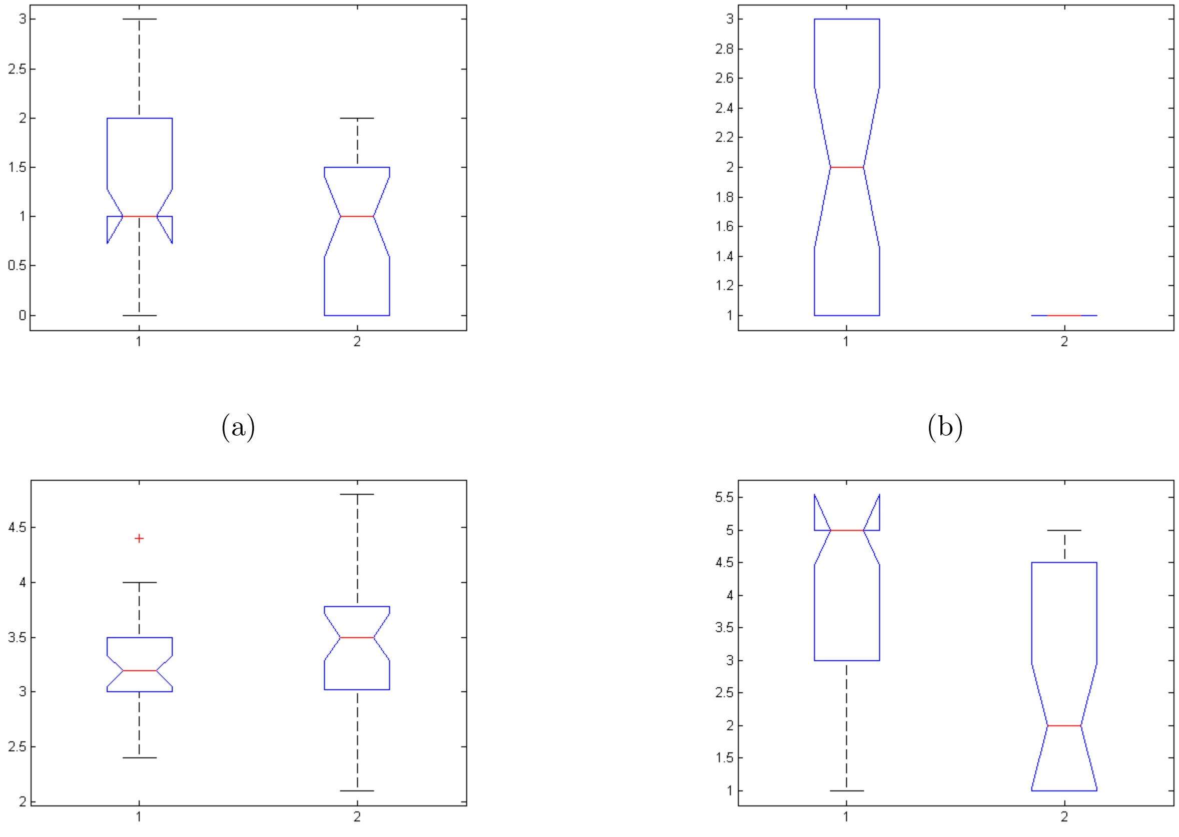
Table 6.14: Kolmogorov-Smirnov test for all features, considering only the stage C patients.

Feature	**Kolmogorov-Smirnov** (*p-value*)
PS	$3,3846 \times 10^{-3}$
Encefalopathy	$3,8381 \times 10^{-14}$
Ascites	$8,3060 \times 10^{-9}$
INR	$2,8494 \times 10^{-3}$
AFP	$6,8028 \times 10^{-9}$
Hemoglobin	$8,3737 \times 10^{-1}$
VGM	$9,3145 \times 10^{-1}$
Leucocytes	$4,2127 \times 10^{-8}$
Platelets	$8,3394 \times 10^{-2}$
Albumin	$8,2605 \times 10^{-1}$
Total Bil	$6,6037 \times 10^{-3}$
ALT	$4,1136 \times 10^{-2}$
AST	$2,5582 \times 10^{-2}$
GGT	$5,3700 \times 10^{-2}$
FA	$2,2391 \times 10^{-2}$
PT	$1,7418 \times 10^{-10}$
Creatinine	$9,6001 \times 10^{-2}$
Number of Nodules	$5,9828 \times 10^{-6}$
Major Dimension	$5,7275 \times 10^{-2}$

Table 6.15: Mann-Whitney's and t-student's test results for all the features considering only the stage C patients.

Feature	**Wilcoxon-Mann-Whitney** (*p-value*)	**t-student** (*p-value*)
PS	$1,4025 \times 10^{-2}$	-
Encefalopathy	$4,5956 \times 10^{-1}$	-
Ascites	$1,5827 \times 10^{-4}$	-
INR	$3,3980 \times 10^{-1}$	-
AFP	$1,2775 \times 10^{-1}$	-
Hemoglobin	-	$1,5816 \times 10^{-1}$
VGM	-	$5,9043 \times 10^{-1}$
Leucocytes	$9,4900 \times 10^{-2}$	-
Platelets	-	$9,7334 \times 10^{-1}$
Albumin	-	$3,5729 \times 10^{-3}$
Total Bil	$2,0164 \times 10^{-1}$	-
ALT	$1,0029 \times 10^{-1}$	-
AST	$1,3572 \times 10^{-2}$	-
GGT	-	$2,4225 \times 10^{-2}$
FA	$1,8952 \times 10^{-1}$	-
PT	$1,2714 \times 10^{-1}$	-
Creatinine	-	$7,2624 \times 10^{-1}$
Number of Nodules	$2,4395 \times 10^{-3}$	-
Major Dimension	-	$6,4405 \times 10^{-1}$

Box-plots for the most interesting features are shown in Figure 6.16. Again, PS, Ascites and Albumin are found between the four most discriminative features. As regards these features related to liver function, BCLC stage C is defined as patients with PS 1 or 2 and CP A or B, which itself encompass heterogeneous patients. Thus, it seems logical that these features are considered discriminative, as, according to our data and results, there can be a set of more specific rules to characterize those patients, creating a new subdivision. Besides PS and CP, stage C consists in patients with multinodular tumours, portal invasion, tumours in regional lymph nodes and metastasis in distant lymph nodes or other organs. This is "the rule" that correctly classifies the majority of patients according to the BCLC system. However, it does not account for every combination: some patients may not verify the rule (or may verify only in part) and furthermore, this staging system does not consider the rest of the features in our study. An interesting results is that the Number of Nodules suggests a good group discrimination. This suggests that some patients in stage C may not have multinodular tumours, and they should be treated accordingly, with a set of personalized "rules". In fact, the mean number of nodules in G2 is 2,5 while the mean in G1 is 4 (multinodular).



(a) (b)

(c) (d)

Figure 6.16: Box-plots for the four most discriminative features, namely PS (a), Ascites (b), Albumin (c) and Number of Nodules (d).

To evaluate the differences between the overall survival in these two groups, we've performed Kolmogorov-Smirnov normality test to stage C patients' survival (*p-value = 0,0025)*, followed by Wilcoxon-Mann-Whitney's, *p-value = 0,1550*. The returned *p-value* indicates that Mann-Whitney's fails to reject the null hypothesis that the two samples come from the same distribution both at a 1 and 5% significance level. Figure 6.17 and Table 6.16 present the summary statistics for each stage C group.

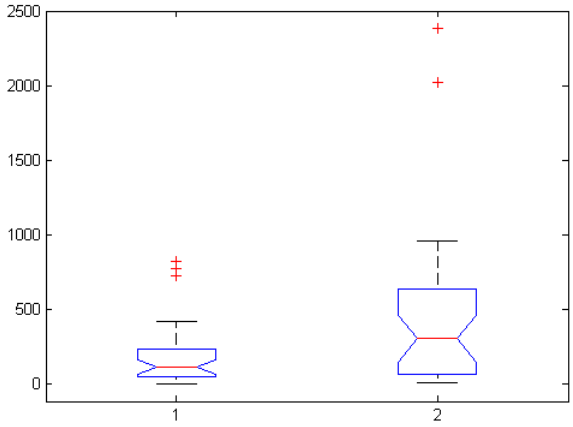


Figure 6.17: Overall survival box-plot for stage C patients in both groups.

Table 6.16: Mean and Standard deviation for stage C patients in both groups.

	Mean (days)	Standard Deviation (days)
Stage C, Group 1	198,2	217,9
Stage C, Group 2	493,5	658,4

Although the Mann-Whitney's test did not return a significant difference between the stage C groups, G2 patients generally have a better prognosis, as shown by the Kaplan-Meier plots in Figure 6.18.

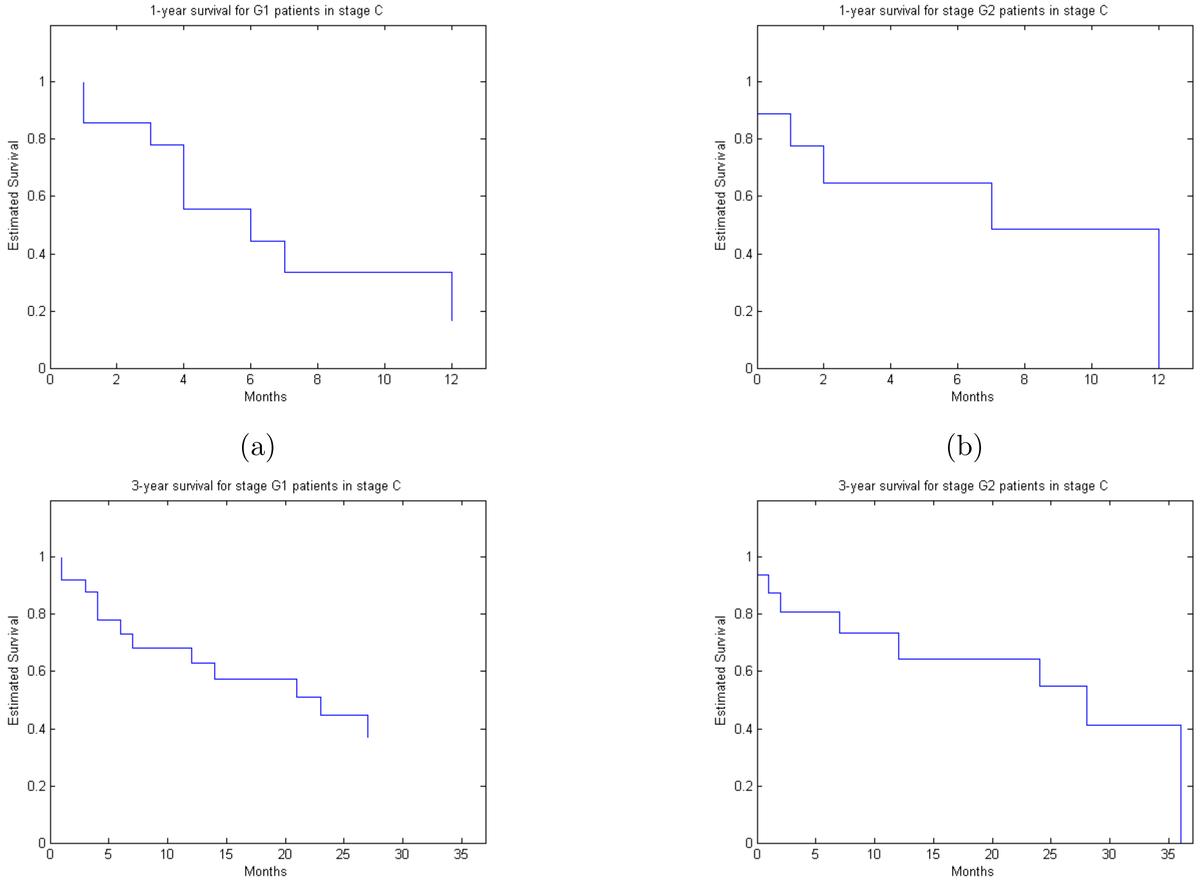
To confirm our findings, we've further inspected the distribution of stage C patients with portal invasion, portal vein tumours and metastasis across both groups. According to the BCLC system, the presence of these three factors are indicative of stage C tumours.

Table 6.17: Comparing the distribution of portal invasion, portal vein tumours and metastases of G1 ad G2.

		Portal Invasion	Portal Vein Tumour	Metastases
G1	**Absent**	51,52%	60,61%	33,33%
	Present	48,48%	39,39%	66,67%
G2	**Absent**	65,00%	63,19%	35,00%
	Present	35,00%	36,84%	65,00%

The distribution is very similar between both groups, which clearly indicates that the heterogeneity between these stage C patients relies on the difference between the patients' general health (Performance Status) and liver function.

According to the BCLC staging systems, the adequate treatment for stage C patients is Sorafenib. We've also examined the combination of treatments performed by stage C patients in G1 and G2. Tables 6.18, 6.19 and 6.20 resume the results.



(a)

(b)

(c)

(d)

Figure 6.18: Kaplan-Meier curves for stage C patients divided in G1 and G2, at 1-year survival - (a) and (b) - and 3-years survival - (c) and (d).

Table 6.18: BCLC treatments codification. RF: radiofrequency ablation, PEI: percitaneous ethanol injection.

Treatment	**Description**
0	No treatment
1	Liver Transplantation
2	Resection
3	PEI
4	RF
5	Microwaves ablation
6	Chemoembolization
7	Sorafenib
8	Supportive Care
9	Clinical Trials
10	Waiting list for transplantation

Table 6.19: Treatments performed by stage C patients in G1.

Stage C, G1	
Treatment Code	**Number of Cases**
467	1
8	13
7	2
6	3
1	3
878	1
61	1
4	1
42	1
67	2
62	1
268	1
47	1
287	1
2	1

Table 6.20: Treatments performed by stage C patients in G2.

Stage C, G2	
Treatment Code	**Number of Cases**
41	1
7	5
2	1
67	1
8	5
47	1
78	1
46478	1
4	1
27	1
48	1
Unknown	1

Considering the patients' follow-up data, we've constructed a set of codes which identify the sequence of treatments performed by each patient. For instance, if a certain patient's treatment code is 67, this means the patients has undergone a Chemoembolization, followed by Sorafenib. Examining Tables 6.19 and 6.20, becomes clear that not every stage C patient is treated only with Sorafenib. Some undergo treatments for earlier stages first, other are never treated with Sorafenib. However, it is noticeable a difference between treatments performed on stage C patients in G1 and G2. Almost half of C-G1 patients are treated in the first place with Supportive Care, not experiencing other earlier stage alternatives. The number of cases in C-G1 that undergo Sorafenib is also considerably lower than C-G2, which explains why these patients have been considered to be closer to stage D cases.

6.6 Classification Task

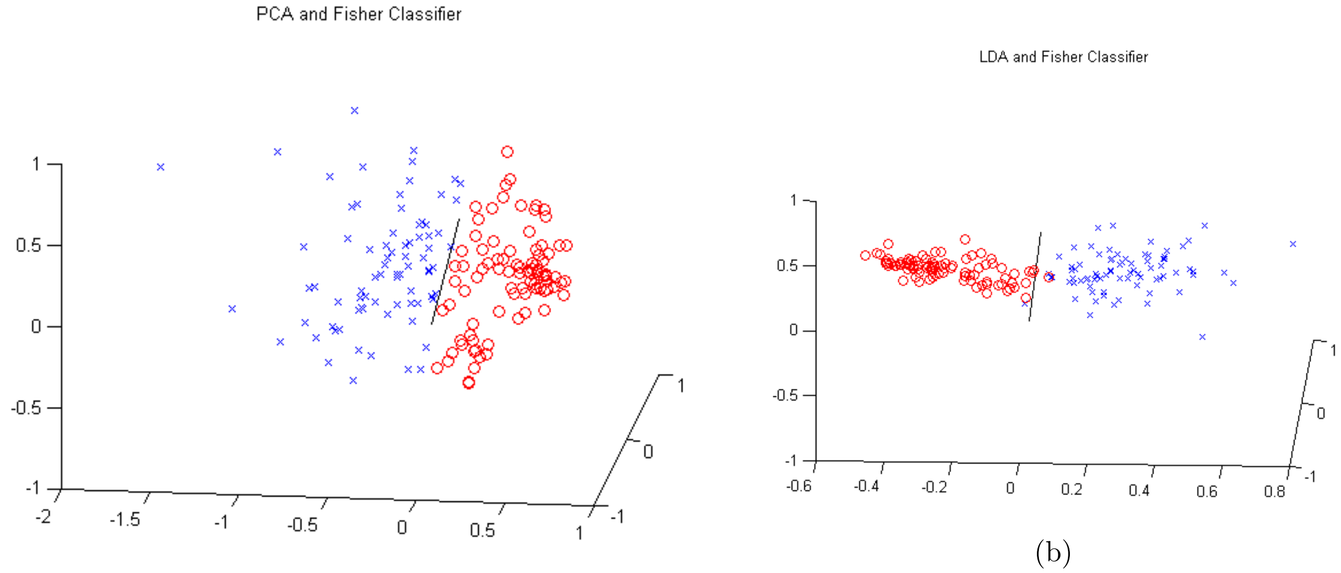
In order to integrate our findings in the system's AI module, we have developed some classification approaches. Every time the system is given a new clinical case, it should generate some recommendations based on the patient's data. This could be achieved by performing k-means clustering with the new complete set of cases, retrieving the best number of clusters and produce recommendation based on the new patient's cluster. However, this would be computationally expensive and time consuming, since the complete set of data had to be analysed each time a new patient was entered into the system. According to our previous conclusions, CHUC's patients can be divide into two main groups: G1 and G2. Thus, our approach consists in studying classification techniques that can accurately predict a new patients group, without the need to evaluate all the data. We have two main objectives: reduce data dimensionality to decrease computation time and finding a model that accurately classifies our data.

In Section 6.4.4, we have studied the dataset's principal components. Our cases suggested to be linearly separable, and thus our first approach was to explore the Fisher Linear Discriminant with both PCA and LDA (Linear Discriminat Analysis). We've performed a 10-fold crossvalidation and bootstrap sampling (20 bootstraps with 100 samples each), using an increasing number of projections (Tables C.1 to C.4). Table 6.21 summarizes the classification results. We have chosen to rely on the 10-fold-crossvalidation experiences, since bootstrap uses resampling, which may not give accurate results in our case, given the dataset's size.

Table 6.21: Classification results for Fisher Classifier, regarding PCA and LDA.

	Accuracy (%)	**F-measure**	**AUC**
Fisher PCA (3D)	98,7868	0,9867	0,8847
Fisher LDA (3D)	98,2353	0,9816	0,8806

The best results are given for 3 projections, considering both PCA and LDA results. Figure 6.19 illustrates Fisher's class assignment for PCA (3D) and LDA (3D), respectively. PCA outperforms LDA in terms of Accuracy, F-measure and AUC, though the results do not pronouncedly differ. Besides Fisher Classifier, we have studied KNN and Bayes Classifier. KNN is an easy concept to grasp for clinicians, and thus our choice. However, KNN is a lazy learner, that is, it does not perform any generalization when creating the predictive model. If a new patient is given to the system, KNN needs to evaluate all the data, in order to classify this new instance. KNN results for different k-neighbours and sampling methods (k-fold and bootstrap) are shown in Tables C.5 and C.6. Table 6.22 resumes the results found for KNN considering all the data, but also considering only 3D feature spaces, given by PCA and LDA, respectively. The best KNN results, in both cases (all data and 3D feature spaces) are given for k=1 and k=2 neighbours. This is not a surprising results, since the dataset's missing values was imputed according to the nearest neighbour for a given instance. Table 6.23 shows the same results for Bayes classifier.



(a)

(b)

Figure 6.19: Fisher's separability criteria for (a) 3D PCA and (b) 3D LDA.

Table 6.22: KNN classification results.

	Accuracy (%)	**F-measure**	**AUC**
KNN (k=1)	90,3162	0,8848	0,8980
KNN (k=2)	90,9559	0,8939	0,9068
KNN PCA (3D, k=1)	95,0319	0,9473	0,9491
KNN LDA (3D, k=1)	98,1569	0,9808	0,9819
KNN PCA (3D, k=2)	95,7255	0,9523	0,9554
KNN LDA (3D, k=2)	98,1619	0,9808	0,9826

Table 6.23: Bayes classification results.

	Accuracy (%)	F-measure	AUC
Bayes	90,2794	0,8945	0,8505
Bayes PCA (3D)	96,3235	0,9640	0,8785
Bayes LDA (3D)	96,9485	0,9725	0,8749

According to our results, a patient's clinical data can be reduced to 3D feature vectors, without the prejudice of decreasing the classification performance. The best results are given for Fisher's classifier considering 3 principal components. A reduced dimensional space with only 3 components requires much less computational effort and allows our system to be faster and more efficient. PCA works great with Fisher Discriminant Analysis, since it allies dimensionality reduction to feature discrimination and data classification. Considering these results, we have chosen the combination between PCA and Fisher Classifier to integrate our AI module and assess a new patient's class (group).

6.7 Conclusions

In this chapter, we have explored several clustering approaches to profile a database of Hepatocellular Carcinoma patients, as a basis to address two questions: first, whether there naturally occurring clusters map onto different prognostic and survival characteristics. Second, whether prognostic groups comprised heterogeneous populations which can be profiled by cluster analysis.

In the first part of our study, we have conducted a clustering approach to the patients' set of risk factors, with heterogeneous and missing data. We have used statistical and machine learning techniques (Mean imputation coupled with Logistic Regression imputation or KNN imputation) to fill absent values in patients' records. MARS algorithm was used to access the need and quality of the chosen imputation techniques: KNN outperformed Logistic Regression imputation. Risk factors data consists in both categorical and continuous features. To perform hierarchical clustering, different similarity measures were tested. HEOM with average linkage distance produced the best results, profiling HCC patients in two distinct groups. However, the groups' overall survival was not statistically different.

This led us to explore a different approach: clustering continuous data. Thus, the second part of our study consisted in partitioning clustering of the patients Laboratory Results. KNN imputation was used to impute missing values. k-means and PAM were used to determine natural clusters in the data. Several clustering solutions were evaluated according to well-known cluster validity indexes, namely Silhouette, Calinski and Rand index. PCA enabled clustering solutions visualization. k-means has proven to be the best clustering solution, with a division in two groups. The prognostic groups, G1 and G2, were found to have statistically different survival curves, as shown by Kaplan-Meier survival analysis. Stage C patients were divided in G1 and G2, which suggested some heterogeneity between these cases. The discriminant features responsible for stage C division were accessed. These features mainly corresponded to features related to liver function status. The treatments performed for both C groups were studied, which confirmed the difference in prognosis in these two types of stage C patients. Finally, a classification task was performed in order to determine a computationally efficient model to predict cluster assignment. Fisher Linear Discriminant, Bayes and KNN classifiers were explored, using two different methods of feature extraction: PCA and LDA. Fisher Discriminant combined with PCA (3D input vectors) outperformed all others, thus being chosen as the AI combination to be integrated in our system's data mining module.

Chapter 7

Conclusions and Future Work

This chapter discusses our work's findings and contributions and outlines directions for future research. Section 7.1 presents a discussion of the conclusions and contributions of the current work, also presenting my personal view regarding this project. Finally, Section 7.2 discusses the future work and brings the thesis to a conclusion.

7.1 Conclusions of the work

This thesis reveals that it is possible to develop a Clinical Decision Support System (CDSS) for HCC patients that integrates clinical data management with AI techniques to support the clinicians' decision-making process. We developed a structured registry system for HCC patients, where the clinicians can systematically register the most influential factors for HCC management. The system allows centralization, multi-user support, real time access to updated information and easy accessibility from any device with access to the internet, without any additional configuration. The structure of the application avoids data inconsistency, since each field has a clear format, and data entry is always validated. The patients' privacy is guaranteed by restricted user access and authentication. An information system for patients' data management avoids the stated problems concerning physical files, since patients' information is available and can be shared at all times. As regards the data mining studies, we've identified 2 main prognostic groups in CHUC database, and the most significant features responsible for this division. The conclusions of our work also suggest that there is some heterogeneity between stage C patients. This is an interesting result which might indicate the need of a subdivision of stage C patients, targeting the treatment of these patients within the paradigm of Personalized Medicine. In conclusion, we have created a framework which allows cancer data management in the HCC context. The framework was intended to allow the clinicians access to patients' information at all times, while supporting them in their daily activities. We have demonstrated that inference models have the potential to assist clinicians in their decisions regarding several therapeutic strategies.

In my opinion, this was truly a challenging work. In real world domains, complex and unexpected problems often arise. Scheduling plans are not always as they were set out to be, and pressure is a constant. Working with a multidisciplinary team led me to develop my knowledge in different areas of expertise. At the end of this work, I came to master technologies and concepts that I had no contact with before. Mastering the required medical terminology was my first obstacle. Regarding the system development, I came across unknown programming and markup languages. Finally, I had not foreseen the need of dealing with missing data. Nevertheless, if the project's goals had not have been this bold, I wouldn't have the opportunity to experience real life situations, with all the problematic issues associated, and learn from them.

7.2 Future Work

This work could be further developed in two main scopes: refining the developed system in the HCC context and extending the system to other medical contexts.

As regards the HCC context, the main approaches would be to improve the data quality. This could be achieved by revising incomplete cases and trying to fill in the absent values or using hot-deck to replace cases with missing values. The first approach requires an extensive review of cases, thus subjected to scheduling issues, errors in data entry, and others mentioned in Section 1.4. The second approach consists in retrieving new cases from another hospital service or institution and substituting patients with incomplete records with patients with complete sets of data from that institution's database.

Extending the developed system to other medical contexts is a more challenging idea. Extending our approach to other areas of Oncology is perhaps the most direct extension of this work. This would require an extensive study of other disease's patterns, in order to identify the fundamental features to include in the system. The system's structure would also have to be adapted to another reality, where the information flow might differ. Finally, in terms of imputation strategies and AI techniques, there are various techniques which could be applied, depending on the type, quality of data and objective function defined.

Bibliography

[1] International Agency for Research on Cancer and World Health Organization. ”Globocan 2008 : Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012” [Online]. Available at: http://globocan.iarc.fr/. [Accessed on: 21 Jan 2014].

[2] World Health Organization. ”Cancer Fact Sheets” [Online]. Available at: http://www.who.int/mediacentre/factsheets/fs297/en/index.html. [Accessed on: 21 Jan 2014].

[3] European Association for the Study of the Liver, European Organisation for Research and Treatment of Cancer, ”EASLEORTC Clinical Practice Guidelines: Management of hepatocellular carcinoma.”, Journal of Hepatology, Vol.56, No.4, pp.908943, 2012.

[4] Tvi24 - Sociedade. ”Cancro do fígado pode aumentar 70 por cento até 2015”. [Online]. Available at: http://www.tvi24.iol.pt/sociedade/tvi24-cancro-do-figado-doenca-hepatica-alcool-sociedade-portuguesa-hepatologia/1162496-4071.html. [Accessed on: 22 Jan 2014]

[5] Rui Tato Marinho, José Giria e Miguel Carneiro Moura. ”Rising costs and hospital admissions for hepatocellular carcinoma in Portugal (1993-2005)”. World Journal of Gastroenterology, Vol.13, No.10, pp.1522-1527, 2007.

[6] Ângelo Alves de Mattos, Fernanda Branco, Luciana dos Santos Schraiber, Andrea Benevides Leite, Livia Caprara Liono e Ane Micheli Costabeber. ”Perfil dos pacientes com diagnóstico de carcinoma hepatocelular acompanhados no Ambulatório de Nódulos Hepáticos da Irmandade Santa Casa de Misericórdia de Porto Alegre”. Revista da AMRIGS, Porto Alegre, Vol. 55, No.3, pp. 250-254, Jul-Set 2011.

[7] Daniel Basílio Leitão. ”Caracterização Clínico-Patológica do Carcinoma Hepatocelular em doentes diagnosticados e tratados no IPO-Porto e avaliação de sobrevivência dos doentes registados no Registo Oncológico da Região Norte (RORENO)”. Dissertação de Mestrado, Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Jun 2010.

[8] David L. Sackett, William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes, W. Scott Richardson, ”Evidence based medicine: what it is and what it isn’t.”, BMJ 1996, 312:71-72.

[9] Andreia da Silva Almeida. ”Os Sistemas de Gestão da Informacão nos Hospitais Públicos Portugueses: uma perspectiva actual”. Master’s thesis, Faculdade de Letras da Universidade de Lisboa, 2012.

[10] World Health Organization. ”Hepatitis C”, Fact Sheet No. 164, April 2014. Available at: http://www.who.int/mediacentre/factsheets/fs164/en/]. [Accessed: 10 Feb 2014]

[11] Zeuzem Stefan, ”Hepatitis B - Risks, prevention and treatment”, ELPA (European Liver Patients Association), 2007.

[12] Zeuzem Stefan, "Hepatitis C - Risks, prevention and treatment", ELPA (European Liver Patients Association), 2009.

[13] Josep M. Llovet *et al.*, "Sorafenib in Advanced Hepatocellular Carcinoma". The New England Journal of Medicine, Vol.359, pp. 378-390, July 24, 2008.

[14] Joshua E. Richardson, Joan S. Ash, Dean F. Sittig, Arwen Bunce, James Carpenter, Richard H. Dykstra, Ken Guappone, James McCormack, Carmit K. McMullen, Michael Shapiro, Adam Wright, Blackford Middleton. "Multiple Perspectives on the Meaning of Clinical Decision Support". AMIA 2010 Symposium, pp. 1427.

[15] Guilan Kong, Dong-Ling Xu, Jian-Bo Yang. "Clinical Decision Support Systems: A review on knowledge representation and inference under uncertainties.", International Journal of Computational Intelligence Systems, Vol.1, No.2, pp.159-167, May 2008.

[16] Punam S. Pawar, D. R. Patil. "Review on Clinical Decision Support System for Electronic Health Record System for Major Diseases.", Proceeding of the Internacional Conference on Advances in Computer, Electronics and Electrical Engineering 2012, pp. 46-50.

[17] Ida Sim, Paul Gorman, Robert A. Greenes, R. Brian Haynes, Bonnie Kaplan, Harold Lehmann, Paul C. Tang. "Clinical Decision Support Systems for the Practice of Evidence-based Medicine.", Journal of the American Medical Informatics Association, Vol.8, No.6, pp. 527-533, Dec 2001.

[18] K. Rajalakshmi, Dr. S. Chandra Mohan, Dr.S.Dhinesh Babu. "Decision Support System in Healthcare Industry", International Journal of Computer Applications, Vol.26, No.9, pp.42-44, Jul de 2011.

[19] Berner ES, Tonya J. La Lande. "Clinical Decision Support Systems: Theory and Practice", Health Informatics 2nd ed. Cap.1, "Overview of Clinical Decision Support Systems", pp.3-9, 2007.

[20] Matthias Samwaldemail, Karsten Fehre, Jeroen de Bruin, Klaus-Peter Adlassnig. "The Arden Syntax standard for clinical decision support: Experiences and directions", Journal of Biomedical Informatics, Vol. 45, Issue 4, pp. 711-718, August, 2012.

[21] Dejan Dinevski, Uros Bele, Tomislav Sarenac, Uros Rajkovicand, Olga Sustersic. "Telemedicine Techniques and Applications.", Cap.8, "Clinical Decision Support Systems.", pp.185-207, InTech, Jun 2011.

[22] M. M. Abbasi, S. Kashinyarndi. "Clinical Decision Support Systems: A discussion on different methodologies used in Heath Care.", Marlaedalen University Sweden. Available at: http://www.idt.mdh.se/kurser/ct3340/ht10/FinalPapers/15-Abbasi_Kashiyarndi.pdf. [Accessed on: 25 Fev 2014]

[23] Liljana Aleksovska. "Review of Reasoning Methods in Clinical Decision Support Systems.", 18th Telecommunications forum TELFOR 2010, pp. 1105-1108. Nov 2010.

[24] Muhamad Adnan, Wahidah Husain, Abdul Rashid. "Data Mining for Medical Systems: A Review". Proceeding of the International Conference on Advances in Computer and Information Technology - ACIT 2012.

[25] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Sori. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction". International Journal of Computer Applications. Vol. 17, No. 8, pp. 43-48. Mar 2011.

[26] Duda, R. O., Hart, P.E., and Stork, D.G. (2001). Pattern Classification, 2nd ed. Wiley Interscience, ISBN: 0-471-05669-3

[27] I. T. Pisa, A. Galina, P.R.L. Lopes, C.N. Barsottini, A.C. Roque. ”Lepidus R3: implementação de sistema de apoio à decisão médica em arquitetura distribuída usando serviços web”. IX Congresso Brasileiro de Informática em Saúde - CBIS’2004, 2004, Ribeirão Preto-SP. Anais do IX Congresso Brasileiro de Informática em Saúde - CBIS’2004. Ribeirão Preto-SP: Sociedade Brasileira de Informática em Saúde, 2004. pp. 224-229.

[28] J. Vasconcelos, A. Rocha e R. Gomes. ”Sistemas de Informação de Apoio à Decisão Clínica: Estudo de um Caso de uma Instituição de Saúde”: Atas da 5ªConferência da Associação Portuguesa de Sistemas de Informação, Lisboa, Portugal, Nov 2004.

[29] Rudolf Wechsler, Meide S. Anção, Carlos José Reis de Campos e Daniel Sigulem. ”A informática no consultório médico”. Jornal de Pediatria, Sociedade Brasileira de Pediatria, 2003, 0021-7557/03/79-Supl.1/S3.

[30] Caisis, BioDigital. Available at: http://www.caisis.org/. [Accessed on: 9 Jan 2014]

[31] DOCgastro, Sistema Integrado em Gastroenterologia. Mobileware, Tecnologias de Informação S.A. Available at: http://www.mobilwave.pt/files/DOCgastro4.pdf. [Accessed on: 9 Jan 2014]

[32] André Narciso, Ângela Oliveira, Pedro Silva. ”MyRisk, Support System for Cancer Diagnosis”, 6th Iberian Conference on Information Systems and Technologies (CISTI), pp.1-5, 15-18 June 2011.

[33] Fox Chase Cancer Center. Available at: http://labs.fccc.edu/nomograms/. [Accessed on: 12 Jan 2014]

[34] Cancer Prognostics and Health Outcomes Unit, University of Montreal. ”Take The Nonogram Challenge” [Online]. Available at: http://www.nomogram.org/. [Accessed on: 16 Jan 2014]

[35] J.C Horrocks, F.T de Dombal, D.J Leaper, J.R Staniland, A.P McCann. ”Computer-aided diagnosis of acute abdominal pain”. British Medical Journal, Vol.2, No.5804, pp.9-13, 1972.

[36] Josceli Maria Tenório, Anderson Diniz Hummel, Vera Lucia Sdepanian, Ivan Torres Pisa and Heimar de Fátima Marin. Ëxperiências internacionais da aplicaç ão de sistemas de apoio à decisão clínica em gastroenterologia.”, Journal of Health Informatics, Vol.3 , No.1, May 2011.

[37] A. Das, T. Ben-Menachem, F.T. Farooq, G.S Cooper, A. Chak, M.V. Sivak, R.C. Wong. ”Artificial neural network as a predictive instrument in patients with acute nonvariceal upper gastrointestinal hemorrhage.”, Gastroenterology, Vol. 134, No. 1, pp. 65-74, January 2008.

[38] A. Chu, H. Ahn, B. Halwan, B. Kalmin, E.L. Artifon, A. Barkun, M.G. Lagoudakis, A. Kumar. ”A decision support system to facilitate management of patients with acute gastrointestinal bleeding.”, Artificial Intelligence in Medicine, Vol. 42, No. 3, pp. 247-259, 2008.

[39] E.S. Berner ES, T.K. Houston TK, M.N. Ray MN, J.J. Allison JJ, G.R. Heudebert, W.W. Chatham et al. ”Improving ambulatory prescribing safety with a handheld decision support system: a randomized controlled trial.” Journal of the American Medical Informatics Association, Vol. 13, No. 2, pp. 171-179, 2006.

[40] K. Farion, W. Michalowski , S. Wilk , D. OSullivan , S. Rubin S, D. Weiss. "Clinical decision support system for point of care useontology-driven design and software implementation.", Methods of Information in Medicine. Vol. 48, No. 4, pp. 381-390, 2009.

[41] S. Sadeghi, A. Barzi, N. Sadeghi, B. King. "A Bayesian model for triage decision support.", International Journal of Medical Informatics, Vol. 75, No. 5, pp. 403-411, 2006.

[42] R.H. Lin. "An intelligent model for liver disease diagnosis.", Artificial Intelligence in Medicine, Vol. 47, No. 1, pp. 53-62, 2009.

[43] P. Aruna, N. Puviarasan, B. Palaniappan. "Diagnosis of gastrointestinal disorders using DIAGNET.", Expert Systems Applications, Vol. 32, No. 2, pp. 329-335, 2007.

[44] Leonard Berliner, Heinz U. Lemke, Eric van Sonnenberg, Hani Ashamalla, Malcolm D. Mattes, David Dosik, Hesham Hazin, Syed Shah, Smruti Mohanty, Sid Verma, Giuseppe Esposito, Irene Bargellini. "Information and communication technology in personalized medicine: a clinical use-case for hepatocellular cancer". EPMA Journal, Vol.5, No.59, Fev 2014.

[45] Hanna A. Wasyluk, Janusz Cianciara, Leon Bobrowski, Alicja Drapato. "Founding of database for cirrhotic patients for early detection of hepatocellular carcinoma". Hepatology, Vol.6, No.3, pp. 13-16, 2010.

[46] W. H. Ho, Lee K. T., H. Y. Chen, T. W. Ho, H. C. Chiu. "Disease-Free Survival after Hepatic Resection in Hepatocellular Carcinoma Patients: A Prediction Approach Using Artificial Neural Network". Plos One, Vol.7, No.1, Article No. 29179, 2012.

[47] W. H. Ho, Lee K. T., H. Y. Chen, T. W. Ho, H. C. Chiu. "Mortality Predicted Accuracy for Hepatocellular Carcinoma Patients with Hepatic Resection Using Artificial Neural Network". The Scientific World Journal.

[48] Martin Dugas, Rolf Schauer, Andreas Volk and Horst Rau. "Interacive decision support in hepatic surgery.", BMC Medical Informatics and Decision Making, Vol. 5, No. 2, May 2002.

[49] Robert S. Ledley and Lee B. Lusted. "Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason". Science, 130(3366):921, 1959.

[50] H. R. Warner, A. F. Toronto, L. G. Veasey, and R. Stephenson. A mathematical approach to medical diagnosis. application to congenital heart disease. JAMA : the journal of the American Medical Association, 177:177183, July 1961.

[51] Adam Wright, Dean F. Sittig. "A four-phase model of the evolution of clinical decision support architectures.", International Journal of Medical Informatics. Vol.77, No.10, pp. 641-649, Mar de 2008.

[52] HELP Health Evaluation Through Logical Processing. Open Clinical. AI Systems in Clinical Practice. Available at: http://www.openclinical.org/aisp_help.html [Accessed on: 28 Fev 2014]

[53] Howard L. Bleich. "Computer Evaluation of Acid-Base Disorders.", The Journal of Clinical Investigation. Vol. 48, No.9, pp. 1689-1696, 1969.

[54] Edward H. Shortliffe. ”MYCIN: A knowledge computer program applied to infectious diseases.”, Proceeding of the Annual Symposium on Computer Application in Medical Care, pp. 66-69, Oct 1977

[55] E. Lahner, M. Intraligi, M. Buscema, M. Centanni, L. Vannella, E. Grossi, B. Annibale. ”Artificial neural networks in the recognition of the presence of thyroid disease in patients with atrophic body gastritis.”, World Journal of Gastroenterology, Vol. 14, No. 4, pp. 563-568, 2008.

[56] J. Yang, A.S. Nugroho, K. Yamauchi, K. Yoshioka, J. Zheng, K. Wang, et al. ”Efficacy of interferon treatment for chronic hepatitis C predicted by feature subset selection and support vector machine.” Journal of Medical Systems, Vol. 31, No. 2, pp. 117-123, 2007.

[57] David Howell. ”The treatment of missing data.”, The Sage handbook of social science methodology, pp. 208-224, 2007.

[58] Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken, New Jersey: John Wiley & Sons, Inc.; 1987.

[59] Little RJA, Rubin DB. Statistical analysis with missing data, 2nd ed., Hoboken, New Jersey: John Wiley & Sons, Inc.; 2002.

[60] Federico Cismondi, Andr S. Fialho, Susana M. Vieira, Shane R. Reti, Joo M.C. Sousa, Stan N. Finkelstein. ”Missing data in medical databases: Impute, delete or classify?”, Artificial Intelligence in Medicine, Vol.58, No.1, pp. 63-72, May 2013.

[61] J.W. Graham, ”Missing Data: Analysis and design.”, Springer (about 323 pages), 2012.

[62] Craig K. Enders. ”Applied Missing Data Analysis (Methodology in the Social Sciences)”, Guilford Press (about 377 pages), 2010.

[63] Loris Nanni, Alessandra Lumini, Sheryl Brahnam. ”A classifier ensemble approach for the missing feature problem.”, Artificial Intelligence in Medicine, Vol. 55, No. 1, pp. 37-50, May 2012.

[64] M. Mostafizur Rahman, Darryl N. Davis. ”Fuzzy Unordered Rules Induction Algorithm Used as Missing Value Imputation Methods for k-mean Clustering on Real Cardiovascular Data”, Proceedings of the World Congress on Engineering, Vol. 1, pp. 391-395, 2012.

[65] Pedro J. Garcia-Laencina, Jos-Luis Sancho-Gómez, Anibal R. Figueiras-Vidal. ”Classifying patterns with missing values using Multi-Task Learning perceptrons”, Expert Systems with Applications, Vol. 40, Issue 4, pp. 1333-1341, March 2013.

[66] José M. Jerez, Ignacio Molina, Pedro J. Garcia-Laencina, Emilio Alba, Nuria Ribelles. ”Missing data imputation using statistical and machine learning methods in a real breast cancer problem.”, Artificial Intelligence in Medicine, Vol. 50, No. 2, pp. 105-115, May 2010.

[67] Joyce C. Ho, Cheng H. Lee, Joydeep Ghosh. ”Spectic Shock prediction for patients with missing data”, Information Systems, Vol. 5, No. 1 (about 15 pages), April 2014.

[68] P.D. Allison. ”Missing data”, Thousand Oaks, Sage Publications, 2001.

[69] Nazziwa Aisha, Mohd Bakri Adam, Shamarina Shohaimi. ”Effect of Missing Value Methods on Bayesian Network Classification of Hepatitis Data”, International Journal of Computer Science and Telecommunications, Vol. 4, Issue 6, pp. 8-12, June 2013.

[70] T.R. Sivapriya, A.R. Nadira Banu Kamal, V. Thavavel. "Imputation And Classification Of Missing Data Using Least Square Support Vector Machines - A New Approach In Dementia Diagnosis", International Journal of Advanced Research in Artificial Intelligence, Vol. 1, No. 4, pp. 29-33, 2012.

[71] Pedro Henriques Abreu, Hugo Amaro, Daniel Castro Silva, Penousal Machado, Miguel Henriques Abreu, Nomia Afonso, Antnio Dourado. "Overall Survival Prediction for Women Breast Cancer Using Ensemble Mehtods and Incomplete Data.", IFMBE Proceedings, Vol. 41, Springer International Publishing, 2014.

[72] Lior Rokach. "Data Mining with Decision Trees: Theory and Applications.", pp. 73-76, World Scientific, 2008.

[73] B. Scholkopf, K. Tsuda, J.P. Vert. "Kernel Methods in Computational Biology.", MIT Press series on Computational Molecular Biology, 2004.

[74] J.P Marques de Sá. "Pattern Recognition: Concepts, Methods and Applications", Springer Science & Business Media, 2001.

[75] Bernardete Ribeiro. "Pattern Recognition Techniques", course slides 2013/1014.

[76] J. Shawe-Taylor, N. Cristianini. "Kernel Methods for Pattern Analysis", Cambridge University Press, 2004.

[77] B. Scholkopf, A. Smola. "Learning with Kernels.", MIT Press, Cambridge MA, 2002.

[78] Xiaofei He, Deng Cai, Shuicheng Yan, Hong-Jiang Zhang. "Neighborhood Preserving Embedding". Computer Vision, Tenth IEEE International Conference, Vol. 2, pp. 1208-1213, Oct, 2005.

[79] Ian Sommerville. "Software Engineering", Addison-Wesly, 9 edition, March 2013.

[80] IIBA International Institute of Business Analysis. "A Guide to the Business Analysis Body of Knowledge", BABOK Guide, 2009.

[81] J.G. Ibrahim, M.H Chen, S.R Lipsitz, A.H. Herring. "Missing data methods for generalized linear models: a comparative review.", Journal of the American Statistical Association, Vol. 100, No. 469, pp. 332-346, 2005.

[82] C.F Manski. "Partial identification with missing data: concepts and findings.", International Journal of Approximate Reasoning, Vol. 39, No. 2, pp. 151-165, 2005.

[83] B. E. Boser, I. M. Guyon, V. N. Vapnik. "A training algorithm for optimal margin classifiers.", 5th Annual ACM Workshop on COLT, pp. 144 - 152, ACM Press, 1992.

[84] D.R. Wilson, T.R. Martinez. "Improved heterogeneous distance functions", Journal of Artificial Intelligence, Vol.6, No.1, Jan 1997.

[85] D.R. Wilson, T.R. Martinez. "Instance Pruning Techniques", Machine Learning: Proceedings of the Fourteenth International Conference, Morgan Kaufmann Publishers, San Francisco, CA, pp. 404-411, 1997.

[86] João Marôco. "Análise Estatística com Utilização do SPSS", ReportNumber, Lda, ISBN: 9899676322

[87] Fabio Piscaglia, Luigi Bolondi. "The intermediate hepatocellular carcinoma stage: Should treatment be expanded?", Digestive and Liver Disease, Vol. 42, No. 3, Jul 2010.

[88] Jason T. Rich, J. Gail Neely, Randal C. Paniello, Courtney C. J. Voelker, DPhil, Brian Nussenbaum, Eric W. Wang. "A practical guide to understanding Kaplan-Meier curves". OtolaryngologyHead and Neck Surgery, Vol. 143, pp.331-336, 2010.

Appendices

Appendix A

Comparative Analysis of CDSSs

Table A.1: Resume of selected applications for sharing and managing clinical data

	Caisis [30]	**DOCgastro** [31]	**MyRisk** [32]	**CancerNanograms.com** [33]	**nanogram.org** [34]
Open Source Tool	x	-	x	x	x
Data Management/Decision Support	Data Management	Data Management	Decision Support	Decision Support	Decision Support
Disease	Cancer	Gastroenterology	Cancer	Cancer	Cancer
Web-based	x	-	x	x	x
Typical Users	Researchers + Health Professionals	Health Professionals	Health Professionals	Health Professionals + Patients	Health Professionals
Data Exportation	x	-	-	-	-
Research/Clinical Context	Research	Clinical Context	Clinical Context	Clinical Context	Clinical Context
Query Information	x	x	-	-	-
System's Specifications	x	-	x	-	-
Stand-alone/Integrated	Stand-alone	Integrated	Stand - alone	x	x
Prototype	x	-	x	x	x

Table A.2: Resume of selected publications in [36]. KNN: k-nearest neighbours; CART: classification and regression tree; CBR: case-based reasoning; ANN: artificial neural networks; SVM: support vector machines; LDA: linear discriminant analysis; SC: shrunken centroid; RF: random forest; LP: Logistic Regression; BP: backpropagation; RBF: radio basis function; Trn: training set; Tst: test set; V: validation set.

Author	**Clinical Issue**	**Studied disease**	**IA techniques**	**Sample size**	**Sensitivity**	**Comparison with expert reviews?**	**User feedback?**	**Improvement in clinical practice**	**Critical issues?**
Lin	Diagnosis	Liver disease	CART CBR	510 clinical cases	CART:92,94% CBR:91,09%	No	No	No	No
Farion *et al.*	Diadnosis	Abdominal pain	-	-	-	No	No	No	No
Das *et al.*	Clinical Approach	Accute gastrointestinal bledding	ANN	Trn:194 Tst:193 V:200	Te:81% VE: 61%	No	No	No	No
Lahner *et al.*	Diagnosis	Thyroid disorders in gastritis patients	RNA	253 clinical cases	75,8%	No	No	No	No
Chu *et al.*	Clinical Approach	Accute gastrointestinal bleeding	SVM,ANN,SC KNN,LDA,RF LR, Boosting	189 clinical cases	80%	No	No	No	No
Aruna *et al.*	Diagnosis	Gastrointestinal disorders	ANN (BP e RBF)	1125 clinical cases	-	No	No	No	No
Yang *et al.*	Drug do effectiveness	Hepatitis C	SVM KNN	112 clinical cases	Effective: 85% Uneffective: 83%	No	No	No	No
Berner *et al.*	Safe medication prescribing	Gastrointestinal bleeding	-	68 physicians	-	Yes	No	Yes	No
Sadeghi *et al.*	Clinical Approach	Non-traumatic abdominal pain	Bayesian network	90 clinical cases	56%	Yes	No	No	Yes

Table A.3: Resume of selected applications for management of Hepatocellular Carcinoma. MEBNs: MultiEntity Bayesian Networks; ANN: artificial neural networks.

Study	**Objective**	**System/Algorithm**	**Methods**	**Comments**
Model-Based Medical Evidence [44]	Diagnosis, Prognosis and personalized treatment of HCC patients	System	MEBNs	So far, the system is only a proposal
e-Hepar III [45]	Diagnosis of Liver Disorders	System	Diagnostic Maps, Case-based reasoning and regression models	The methods are not detailed
Ho *et al.* [46]	Prediction of free survival disease after hepatic ressection	Algorithm	ANNs, Logistic Regression, Decision Trees	Of limited potential: only applied to patients subjected to hepatic ressection
Ho *et al.* [47]	Mortality Prediction after hepatic ressection	Algorithm	ANN, Logistic Regression	Of limited potential: only applied to patients subjected to hepatic ressection
HCC risk assessment tool [48]	Recommendation of the most appropriate surgery procedure	System	Similar Cases are considered	Uses only 5 parameters when finding similar cases. Plots Kaplan-Meier curves, considering overall survival

Appendix B

Function Requirements Full Description

Table B.1: U-1 description.

Use Case ID	U-1
Use Case Name	Patient Quick Filter
Actors	User
Description	The user is provided with two input boxes, one for the patient name and the other for Institution Patient ID (PID), that he can use to filter the patients by name or PID in order to quickly find someone in specific.
Trigger	This functionality is available as soon as the Patient List View is loaded.
Normal Flow	The user selects one of the two input boxes and starts typing either the name or the PID. Whenever the user releases a key, any previous Ajax [1] Requests are cancelled. A new Ajax Request is sent, with the content the user has typed, and returns a filtered list of patients. When the request finishes the previous patient list is replaced with a new one, displaying the filtered results.
Alternative Flows	The user clears the content of the Quick Filter Input boxes; the current patient list is replaced with a new one displaying the unfiltered patient list.
Notes and Issues	If the Ajax request returns a empty patient list it means that no patient that matched the filter was found in the database. Thus, an empty list is displayed to the user.

[1] Ajax is a group of Web development techniques to exchange data with a server. An Ajax Request requests data from the server, while an Ajax Post sends data to the server.

Table B.2: U-2 description.

Use Case ID	U-2
Use Case Name	Enter Patient View
Actors	User
Description	This use case allows the user to access the patient's information and medical data.
Trigger	This functionality is available as soon as the Patient List View is loaded.
Preconditions	The User is in Patient List View.
Postconditions	The User is in Patient View.
Normal Flow	The User clicks with the left mouse button over the desired patient's row from the patient's list table.
Assumptions	The patients list is not empty.

Table B.3: U-3 description.

Use Case ID	U-3
Use Case Name	Insert Patient
Actors	User
Description	The user is provided with a form that allows for the insertion of a new patient.
Trigger	User clicks with the left mouse button over the button "Insert Patient".
Preconditions	The User is in the Patient List View.
Postconditions	The User is in the Patient View with the inserted patient selected.
Normal Flow	The user fills the patient information regarding each of the different fields. The user clicks with the left mouse button over the button "Insert". The form is submitted via Ajax Post Request and the new patient is inserted. If the request is successful, the user is taken to the Patient View of the inserted patient.
Alternative Flows	The user fills in the patient's information regarding each of the different fields. The user clicks with the left mouse button over the button "Insert". The form is submitted via Ajax Post Request and the new patient is inserted. If an error occurs during the patient insertion, the user is informed and he is taken to the Patient View List.

Table B.4: U-4 description.

Use Case ID	U-4
Use Case Name	Edit Patient General Information
Actors	User
Description	The user has the ability to quickly and easily edit any of the patient's information.
Trigger	The User clicks with the left mouse button over the Text of any pair (Label: Text) regarding any of the patient's information (attributes) displayed in the Patient View.
Preconditions	The User is in the Patient View
Normal Flow	The Text in the (Label: Text) pair where the user clicked is replaced with an input field tailored for the respective attribute's type. The User clicks with the left mouse button outside of the input field (the input field must loose its focus). The patient information edited by the User is sent by Ajax Post Request to the server and is updated in the database.

Table B.5: U-5 description.

Use Case ID	U-5
Use Case Name	Remove Patient
Actors	User
Description	The user has the ability to quickly and easily remove any patient and all of his associated data from the database.
Trigger	This functionality is available as soon as the Patient View is loaded.
Preconditions	The User is in the Patient View.
Postconditions	The User is in the Patient View List.
Normal Flow	The User clicks with the left mouse button over the button "Remove Patient". A confirmation box is displayed to the User. If the User confirms his intent to remove the patient, an Ajax Post Request is sent to the server. The User is redirected to the Patient View List.
Alternative Flows	The User clicks with the left mouse button over the button "Remove Patient". A confirmation box is displayed to the User. The User chooses "Cancel" option and is redirected to the initial condition.

Table B.6: U-6 description.

Use Case ID	U-6
Use Case Name	Insert New Patient Evaluation
Actors	User
Description	The user is provided with a form that allows for the insertion of a patient's Medical Evaluation.
Trigger	User clicks with the left mouse button over the button "Insert New Patient Evaluation".
Preconditions	The User is either in the Patient List View or the Patient View.
Postconditions	The User is in the Patient View.
Normal Flow	The user fills the information regarding each of the different fields. The user clicks with the left mouse button over the button "Insert". The form is submitted via Ajax Post Request and the new patient data is inserted. If the request is successful the user is redirected to the respective Patient View.
Alternative Flows	The user fills in the patient information regarding each of the different fields. The user clicks with the left mouse button over the button "Insert". The form is submitted via Ajax Post Request and the new patient is inserted. If an error occurs during the patient's insertion, the user is informed of the error and is redirected to the respective Patient View.
Special Requirements	The User is provided with two fields regarding the patient's identification, a name field and a PID field. If the user entered the Use case from the Patient View, these fields are already filled in. Otherwise, the User will have to type part of patient's the name or PID in order to gain access to a list of patients, filtered by the user inserted text.

Table B.7: U-7 description.

Use Case ID	U-7
Use Case Name	Insert New Patient Biopsy
Actors	User
Description	The user is provided with a form that allows for the insertion of a patient's Biopsy information.
Trigger	The User clicks with the left mouse button over the button "Insert New Patient Biopsy".
Preconditions	The User is either in the Patient List View or in the Patient View.
Postconditions	The User is in the Patient View.
Normal Flow	The user fills the information regarding each of the different fields. The user clicks with the left mouse button over the button "Insert". The form is submitted via Ajax Post Request and the new patient data is inserted. If the request is successful, the user is redirected to the respective Patient View.
Alternative Flows	The user fills the patient information regarding each of the different fields. The user clicks with the left mouse button over the button "Insert". The form is submitted via Ajax Post Request and the new patient is inserted. If an error occurs during the patient's insertion, the user is informed and he is redirected to the respective Patient View.
Special Requirements	The User is provided with two fields regarding the patient identification, a name field and a PID field. If the user entered the Use case from the Patient View, these fields are already filled. Otherwise, the User will have to type part of the patient's name or PID in order to gain access to a list of patients, filtered by the user's inserted text.

Table B.8: U-8 description.

Use Case ID	U-8
Use Case Name	Insert New Patient Exam
Actors	User
Description	The user is provided with a form that allows for the insertion of a patient's Medical Exam.
Trigger	User clicks with the left mouse button over the button ”Insert New Patient Exam”.
Preconditions	The User is either in the Patient List View or the Patient View.
Postconditions	The User is in the Patient View.
Normal Flow	The user fills the information regarding each of the different fields. The user clicks with the left mouse button over the button ”Insert”. The form is submitted via Ajax Post Request and the new patient data is inserted. If the request is successful, the user is redirected to the respective Patient View.
Alternative Flows	The user fills in the patient information regarding each of the different fields. The user clicks with the left mouse button over the button ”Insert”. The form is submitted via Ajax Post Request and the new patient is inserted. If an error occurs during the patient's insertion, the user is informed and he is redirected to the respective Patient View.
Special Requirements	The User is provided with two fields regarding the patient identification, a name field and a PID field. If the user entered the Use case from the Patient View, these fields are already filled in. Otherwise, the user will have to type part of the patient's name or PID in order to gain access to a list of patients, filtered by the User inserted text.

Table B.9: U-9 description.

Use Case ID	U-9
Use Case Name	Insert New Patient Treatment
Actors	User
Description	The user is provided with a form that allows for the insertion of a patient's Medical Treatment.
Trigger	User clicks with the left mouse button over the button "Insert New Patient Treatment".
Preconditions	The User is either in the Patient List View or in the Patient View.
Postconditions	The User is in the Patient View.
Normal Flow	The user fills the information regarding each of the different fields. The user clicks with the left mouse button over the button "Insert". The form is submitted via Ajax Post Request and the new patient's data is inserted. If the request is successful, the user is redirected to the respective Patient View.
Alternative Flows	The user fills in the patient information regarding each of the different fields. The user clicks with the left mouse button over the button "Insert". The form is submitted via Ajax Post Request and the new patient is inserted. If an error occurs during the patient's insertion, the user is informed and he is redirected to the respective Patient View.
Special Requirements	The User is provided with two fields regarding the patient's identification, a name field and a PID field. If the user entered the Use case from the Patient View, these fields are already filled in. Otherwise, the User will have to type part of the patient's name or PID in order to gain access to a list of patients, filtered by the User inserted text.

Table B.10: U-10 description.

Use Case ID	U-10
Use Case Name	Insert Patient Risk Factors
Actors	User
Description	The user is provided with a form that allows for the insertion of a patient's Risk Factors.
Trigger	User clicks with the left mouse button over the button "Insert New Patient Risk Factors".
Preconditions	The User is either in the Patient List View or the Patient View.
Postconditions	The User is in the Patient View.
Normal Flow	The user fills in the information regarding each of the different fields. The user clicks with the left mouse button over the button "Insert". The form is submitted via Ajax Post Request and the new patient data is inserted. If the request is successful, the user is redirected to the respective Patient View.
Alternative Flows	The user fills in the patient information regarding each of the different fields. The user clicks with the left mouse button over the button "Insert". The form is submitted via Ajax Post Request and the new patient is inserted. If an error occurs during the patient's insertion, the user is informed and he is redirected to the respective Patient View.
Special Requirements	The User is provided with two fields regarding the patient's identification, a name field and a PID field. If the user entered the Use case from the Patient View, these fields are already filled in. Otherwise, the User will have to type part of the patient's name or PID in order to gain access to a list of patients, filtered by the User inserted text.

Table B.11: U-11 description.

Use Case ID	U-11
Use Case Name	Edit Patient Data
Actors	User
Description	The User is allowed to edit any of the Patients Risk Factors or any other of its Medical Data on-the-fly.
Trigger	The User clicks with the left mouse button over a Text part of any pair (Label: Text) regarding any of the patient's information. A confirmation box os shown. The user acknowledges the existence and dangers of the on-the-fly. He edits the desired functionality and clicks "Yes".
Normal Flow	The User clicks with the left mouse button over a Text part of any pair (Label: Text) inside any of the Patients View: Evaluations, Biopsies, Exams, Treatments and Risk Factors. The Text where the user clicked is replaced with an input field tailored for the respective attribute type. After editing the information, the User clicks with the left mouse button outside of the input field (the input field must loose its focus). The patient's information edited by the User is sent by Ajax Post Request to the server and updated in the database. The User is informed of the success of the operation.

Table B.12: U-12 description.

Use Case ID	U-12
Use Case Name	Remove Patient Data
Actors	User
Description	The User is able to remove any of the inserted patient medical data: Medical Evaluations, Biopsies, Exams, Treatments and Risk Factors.
Trigger	The User clicks the button labelled ”Delete”.
Preconditions	The User is in the Patient View.
Postconditions	The User is redirected to the closest patient's record of the same type, if available (Evaluation, Biopsy, Exam, Treatment or Risk Factors) or Risk Factors by default in case there is no more information of the same type for this patient.
Normal Flow	The User clicks with the left mouse button over the button labelled ”Delete”. A confirmation box is displayed confirming the elimination of the current selected Evaluation, Biopsy, Exam, Treatment or Risk Factors. The User confirms his intent to delete the selected data. An Ajax Request is sent to the server and the data is eliminated from the database. The User is informed of the completion of the operation.

Table B.13: U-13 description.

Use Case ID	U-13
Use Case Name	Authentication
Actors	User
Description	When the application is loaded for the first time, or any time the user session becomes void or invalid, a authentication form is presented to the user so he can enter his login information.
Trigger	The authentication form is available as soon as the page loads.
Preconditions	The User's browser loaded the page for the first time or the user session became void or invalid.
Postconditions	The User is authenticated in case of a successful authentication.
Normal Flow	The user fills the information regarding the username and corresponding password. The user clicks with the left mouse button over the submit button. The provided login information is sent and validated by the server. The page is reloaded with access to the application in case of a successful login.
Alternative Flows	The user fills the information regarding the username and corresponding password. The user clicks with the left mouse button over the submit button. The provided login information is sent and validated by the server. The login fails and the User is redirected to the page's initial condition.

Table B.14: U-14 description.

Use Case ID	U-14
Use Case Name	View Distribution Report
Actors	User
Description	The User has access to a report that includes a Bar Chart and a Data Table regarding the patient's distributions. The target feature for which the User wants to see the patient's distributions can be chosen from several of the patient's inserted medical data and the User has the ability to filter the patients prior to their distribution.
Trigger	The User selects "See Patients Distribution" from the Select Input in the Reports View.
Normal Flow	The User may select a Filter and fill in the corresponding options to filter the patients in the database prior to the distribution calculation. The User may select a different feature as the target of the Distribution. The View or Selected distribution is updated automatically every time the User changes one of the selected options.

Table B.15: U-15 description.

Use Case ID	U-15
Use Case Name	View Kaplan-Meier Survival Function Estimation
Actors	User
Description	The User has access to a report that includes a Step Graph and a Data Table regarding the Kaplan-Meier Survival Function Estimation for the selected conditions. The target feature for which the User wants to see the Survival Estimation can be chosen from several of the patient's inserted medical data and the User has the ability to filter the patients prior to the calculation.
Trigger	The User selects "See Patients Survival" from the Select Input in the Reports View.
Normal Flow	The User may select a Filter and fill the corresponding options to filter the patients in the database prior to the Kaplan-Meier calculation. The User may choose a different feature for grouping the patients and calculate the Survival Estimation for each of the groups. The View or Selected Survival Estimation is updated automatically every time the User changes one of the selected options.

Table B.16: A-1 description.

Use Case ID	A-1
Use Case Name	Import Data
Actors	Admin
Description	The Admin is able to import patient data from an Excel data file that follows a specific template determined in conjunction with the Institution during the development of the application.
Preconditions	A file named ”mainxls.xlsx” must be present in the root folder of the web server and must follow the established template of the original Excel Data file provided by the Institution.
Postconditions	The database is update with the information of the Patients included in the Excel file.
Normal Flow	The Admin opens the file import functionality URL. The script opens and parses the information in the Excel file, inserting any Patient found and any Medical Data regarding the Patient.
Frequency of Use	This Use Case should only be used once, to setup the initial database, or in case of a new patient database, carefully formatted to the correct template, that will be appended to the current Patient’s database.
Notes and Issues	This use case is quite destructive and should be used with care.

Appendix C

AI Module Classification Studies

Table C.1: Fisher Classification results for PCA, with increasing number of principal components kept. Validation was made using a 10-fold crossvalidation sampling.

	1D	**2D**	**3D**	**4D**	**5D**	**6D**	**7D**	**8D**	**9D**	**10D**	**11D**	**12D**	**13D**	**14D**	**15D**	**16D**	**17D**	**18D**	**19D**
Accuracy (%)	94,4387	94,5956	98,7868	98,1618	97,6103	96,3603	95,8088	96,3235	95,1838	95,1471	95,7721	94,4755	96,4706	95,7721	96,9118	96,2868	96,4706	96,9485	96,3603
F-measure	0,93643	0,93888	0,98667	0,97905	0,97238	0,95897	0,95578	0,95824	0,94728	0,94812	0,95722	0,94221	0,96523	0,95652	0,96907	0,96167	0,96379	0,96907	0,9624
AUC	0,88472	0,8776	0,88472	0,88472	0,88333	0,87778	0,87865	0,88316	0,87897	0,87579	0,87205	0,88194	0,86667	0,87743	0,8776	0,87051	0,88056	0,87569	0,87552

Table C.2: Fisher Classification results for PCA, with increasing number of principal components kept. Validation was made using a bootstrap sampling.

	1D	**2D**	**3D**	**4D**	**5D**	**6D**	**7D**	**8D**	**9D**	**10D**	**11D**	**12D**	**13D**	**14D**	**15D**	**16D**	**17D**	**18D**	**19D**
Accuracy (%)	95,2	96	98	97,45	97,6	97,7	98,2	97,95	98,2	95,8088	98,55	97,85	97,9	98,55	98,9	98,8	98,56	98,7	98,95
F-measure	0,94616	0,957	0,97837	0,97167	0,97431	0,97458	0,98106	0,97848	0,98055	0,95574	0,98459	0,97731	0,97774	0,98458	0,98856	0,98655	0,98555	0,98603	0,98922
AUC	0,9787	0,97795	0,97984	0,9799	0,97832	0,97937	0,97948	0,97951	0,97982	0,87743	0,97964	0,97941	0,97844	0,97992	0,97927	0,98031	0,98073	0,98043	0,97988

Table C.3: Fisher Classification results for LDA, with increasing number of principal components kept. Validation was made using a 10-fold crossvalidation sampling.

	1D	**2D**	**3D**	**4D**	**5D**	**6D**	**7D**	**8D**	**9D**	**10D**	**11D**	**12D**	**13D**	**14D**	**15D**	**16D**	**17D**	**18D**	**19D**
Accuracy (%)	97,5319	97,6103	98,2353	97,0588	97,6471	98,1985	96,9853	97,5735	97,4951	97,0221	97,5686	98,1618	96,3186	96,9485	97,6471	97,5319	97,6103	96,4338	96,3554
F-measure	0,97412	0,97634	0,98162	0,97059	0,97569	0,98157	0,97046	0,97634	0,97531	0,96967	0,97466	0,98054	0,96157	0,96828	0,975	0,97412	0,97495	0,96384	0,96277
AUC	0,88194	0,88472	0,88056	0,87917	0,88333	0,88194	0,88333	0,88316	0,88294	0,88333	0,88333	0,88157	0,88056	0,88016	0,88194	0,87718	0,88194	0,87222	0,88056

Table C.4: Fisher Classification results for LDA, with increasing number of principal components kept. Validation was made using a bootstrap sampling.

	1D	**2D**	**3D**	**4D**	**5D**	**6D**	**7D**	**8D**	**9D**	**10D**	**11D**	**12D**	**13D**	**14D**	**15D**	**16D**	**17D**	**18D**	**19D**
Accuracy (%)	98,5	98,25	98,25	98,95	98,2	98,6	98,65	98,65	98,6	98,75	98,85	99,05	98	98,6	98,6	99,05	98,8	98,6	98,9
F-measure	0,98375	0,98127	0,98046	0,98872	0,9817	0,98491	0,9854	0,98578	0,98559	0,9865	0,98817	0,98963	0,97971	0,98538	0,98565	0,98967	0,98737	0,98505	0,98825
AUC	0,97965	0,97994	0,98049	0,9802	0,9791	0,98089	0,98023	0,98004	0,97948	0,98095	0,97993	0,98076	0,97958	0,97998	0,97968	0,98093	0,98005	0,9808	0,98009

Table C.5: KNN Classification results for increasing number of nearest neighbours. Validation was made using a 10-fold crossvalidation sampling.

	k=1	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**	**k=8**	**k=9**	**k=10**	**k=11**	**k=12**	**k=13**
Accuracy (%)	90,3162	90,9559	87,9412	89,0392	90,2941	89,5588	91,4706	90,9191	89,7426	90,1838	89,0074	89,1176	88,4191
F-measure	0,88478	0,8939	0,85823	0,86212	0,88092	0,86936	0,89985	0,89004	0,86229	0,87381	0,86005	0,8644	0,85602
AUC	0,89802	0,90675	0,8753	0,88552	0,89732	0,89018	0,91071	0,90625	0,89107	0,89464	0,88393	0,88571	0,87768

Table C.6: KNN Classification results for increasing number of nearest neighbours. Validation was made using a bootstrap sampling.

	k=1	**k=2**	**k=3**	**k=4**	**k=5**	**k=6**	**k=7**	**k=8**	**k=9**	**k=10**	**k=11**	**k=12**	**k=13**
Accuracy (%)	100	100	96	97,95	93,35	96,1	93,25	94,7	90,65	93,55	91,75	94,9	92,15
F-measure	1	1	0,95696	0,97831	0,91744	0,95647	0,92189	0,9383	0,88471	0,92277	0,90099	0,9409	0,90311
AUC	1	1	0,95829	0,97892	0,92502	0,95862	0,92835	0,94373	0,897785	0,92973	0,9057	0,94547	0,91339