STOCHASTIC AND HEURISTIC MODELLING FOR ANALYSIS OF THE GROWTH OF PRE-INVASIVE LESIONS AND FOR A MULTIDISCIPLINARY APPROACH TO EARLY CANCER DIAGNOSIS

by

Alma Iridia Barranco-Mendoza
B.Sc.Hons., Trent University, 1994
M.Sc., Trent University,1997

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

In the School
of
Computing Science

SIMON FRASER UNIVERSITY

Spring 2005

APPROVAL

Name: Alma Iridia Barranco-Mendoza

Degree: Doctor of Philosophy

Title of Thesis: Stochastic and Heuristic Modelling for Analysis of the Growth of Pre-Invasive Lesions and for a Multidisciplinary Approach to Early Cancer Diagnosis

Examining Committee:

Chair: **Dr. Janice Regan**
Lecturer of Computing Science

Dr. Veronica Dahl
Senior Supervisor
Professor of Computing Science

Dr. Fred Popowich
Supervisor
Professor of Computing Science

Dr. Miriam Rosin
Supervisor
Professor of Kinesiology

Dr. Diana Cukierman
Internal Examiner
Lecturer of Computing Science

Dr. Paul Tarau
External Examiner
Associate Professor Dept. of Computer Science and Engineering, University of North Texas

Date Defended/Approved: April 20th, 2005

SIMON FRASER UNIVERSITY

PARTIAL COPYRIGHT LICENCE

ABSTRACT

To this day, lung cancer remains the leading cause of all cancer deaths for both sexes. Current treatment options lead to a cure in only about ten percent of diagnosed cases of lung cancer. One of the main reasons why this type of cancer has such poor prognosis is that it is very difficult to diagnose at the early stages. It is well known that the survival rates can be improved by the early detection of pre-invasive lesions, which are believed to be the possible precursors to malignant tumours. Although new diagnostic devices are allowing numerous lesions to be detected early, it is becoming clear that only a small percentage of these will actually progress to cancer. Therefore, the critical question is how to determine the factors that will define which of these lesions will become malignant.

In this thesis, two computational models and a novel approach to represent biological knowledge for use in the early diagnosis of cancer are presented. In the first part, a stochastic model representing the early development of pre-invasive neoplastic bronchial epithelial lesions as contact processes is introduced. The results of the simulations run on this model gave us some insight on the probability of growth of specific lesions.

Yet, it also shed light on the fact that for an effective diagnostic tool we would need to consider a lot more information about the patients and their condition beyond the structural behaviour of independent lesions. This led to the development of a new approach to multidisciplinary biological knowledge representation: the Probabilistic Property-Based Model (PPBM). Based on a cognitive model of knowledge construction, PPBM presents a heuristic approach to diagnosis by taking into account multiple-domain elements such as imaging, serum, sputum, cytological and genetic data as well as personal medical history and lifestyle factors.

DEDICATION

To Deryck, my husband, best friend, advisor, and strongest supporter.

To Amir and Daniel, my beloved sons.

To Lupita and Rodolfo, my always loving and supportive parents.

ACKNOWLEDGEMENTS

"I can do anything through Him who gives me strength." Philippians 4:13

This research has been a large collaborative effort involving a lot of people whom I would like to thank and acknowledge in this section. First of all, I want to thank Dr. Deryck Persaud, who not only is a wonderful and supportive husband, best friend, and father to my kids, but also was the main collaborator from the BC Cancer Research Centre (BCCRC) and the main person who supervised this research from the biological side. He kept my research true to the biological facts and did not let it become "just a nice computing exercise".

I also want to express my most sincere appreciation and thanks to my senior supervisor and mentor, Dr. Veronica Dahl, for taking me as her advisee even when this was a new area of application for her and for believing in me and my research when I was ready to give up. I would not have been able to do this without her constant support, advice, and friendship. I also want to thank the rest of my supervisory committee: Dr. Fred Popowich, for all his useful comments and insight from the computational side, and to Dr. Miriam Rosin, also for her advice and insight from the biological side. I look forward to our continuing collaboration in this project. I could not have asked for a better committee.

Thanks also to my examiners, Dr. Diana Cukierman and Dr. Paul Tarau, for all their useful comments to make this a better manuscript. My appreciation also goes to Dr. Janice Regan and Dr. Anne Condon for taking the time from

their busy schedules to attend my defence as chair and external observer, respectively.

I would like to thank and acknowledge Dr. Carole Clem and Dr. Martial Guillaud, from the Imaging Lab. of the BCCRC, who introduced the initial problem that gave birth to this research and for their collaboration in the development of the stochastic model here presented. Also, thanks to Dr. Arvind Gupta for supervising the development of this stochastic model and for all his advice and collaboration at the beginning of my Ph.D. research. I would also like to acknowledge Dr. Rita Aggarwala, from the University of Calgary, Dr. Perry Fizzano, from the University of Puget Sound, and Dr. Fiona Beardwood, from the University of Western Ontario, for their collaboration on the initial analysis of this problem during the Pacific Institute of Mathematical Sciences Industrial Problem Solving Workshop 1997.

From Trinity Western University, I would like to give thanks to my students Gregory Eppel and Bernerd Farrant, for their help in the implementation of the PPBM prototype, and to my colleagues, friends, and the rest of my students for their moral support, prayers, and encouragement while finishing the last version of this manuscript and preparing the defence.

Thank you to the Natural Sciences and Engineering Research Council and the BC Advanced Systems Institute for the scholarships that funded in part this research and to Dr. David Poole, my M.Sc. supervisor, mentor, and friend, for encouraging me to apply for them.

To all my friends and members of the Logic and Functional Programming Lab. and the Algorithms Lab. at SFU: Kimberly Pratt (Thanks for everything; you are the best! The LFP Aquarium truly kept me going), Gabrielle Grün, Manuel Zahariev, Tamara Dakic, Junas Adhikary, Brad Bart, Robert Benkoczi, Peter Gvozdjak, Snezana Mitrovic-Minic, Ann Grbavec, Glendon Holst, Baohua Gu, Dulce Aguilar, Maryam Bavarian, Jiang Ye, and anybody that I may have unknowingly omitted, thanks for your friendship and support over the years.

To Dr. Rizwan Kheraj and the rest of the people at Knowledge Junction Systems, Inc., thanks for your friendship and for enabling me to have a flexible enough schedule to do this research during the years I worked there.

I would like to thank my parents, Lupita Mendoza and Rodolfo Barranco, for their never-failing love and support through out my entire life and specially during my difficult pregnancy and those first months of my sons' lives. I don't know how I would have managed without you. Thanks to my sisters Diana, for those crucial months of babysitting while I was writing this thesis, and Ivette, for your love and moral support.

Last, but not least, I want to thank my sons, Daniel and Amir, for being the light of my life and for making me realize that, after becoming the mother of twins, completing a Ph.D. is not such a big deal after all.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AI	Artificial Intelligence
BCCRC	British Columbia Cancer Research Centre
CAD	Computer-assisted diagnosis
CART	Classification and regression tree
CEA	Carcinogenic antigen
CFR	Concept Formation Rules
CHR	Constraint Handling Rules
CHRG	Constraint Handling Rules Grammar
CT or "CAT"	Computed tomography
CYFRA21-1	Cytokeratin fragment 19
DE	Diagnostic Engine
DKS	Diagnostic Knowledge Store
FNA	Fine needle aspiration
FOB	Conventional white light fiber-optic bronchoscopy
KB	Knowledge Base
LIFE	Lung Imaging Fluorescence Endoscopic Device
MRI	Magnetic resonance imaging
NLM	National Library of Medicine
NSCLC	Non-small cell lung cancer
PCR	Polymerase chain reaction
PDC	Patient Data Concepts
PDE	Partial differential equation
PG	Property Grammars
PPBM	Probabilistic Property-Based Model
PSAC	Patient-Specific Atemporal Concepts
PSTC	Patient-Specific Temporal Concepts
PTH	Parathyroid hormone
SBML	Systems Biology Mark-up Language
SCLC	Small cell lung cancer
SELDI-TOF	Surface-enhanced laser desorption and ionization time-of-flight
UMLS	Unified Medical Language System
PDF	Probability Density Function

CHAPTER 1: INTRODUCTION

Pathologists diagnosing lung cancer in a patient must consider the global architecture of the bronchial tissue as well as the local architecture of cell groups and the appearance of individual cells. In the case of lung cancer, in order to obtain more detailed information on the condition of the bronchial tissue, a bronchoscopy is performed on the patient and tissue samples of any detected lesion are obtained. These samples are extracted from a region of the tissue containing abnormal cells; it is from these sections that the diagnosis must be made. In addition to the architectural analysis, these samples are also analyzed in the laboratory to determine their physiological and molecular characteristics.

Many times an abnormal cell will die naturally without forming a cancer; therefore the pathologists are concerned with identifying only the cases that will eventually develop into cancer. Since currently available cancer treatments are very aggressive and traumatic for the patient, pathologists want to be fairly certain that the abnormal cells present in the sample will in fact lead to cancer before recommending treatment. When the sample contains either large amounts of abnormal cells or none at all the diagnosis is straightforward. In many cases, however, there are just a few abnormal cells in the sample and diagnosis is difficult.

The motivation for this thesis is based on the following observations:

Observation 1: *Stochastic processes have been proven to be useful modelling tools to represent biological systems involving the behaviour of competing populations.*

Considering the fact that epithelial lesion development depends on the relative growth of normal vs. abnormal cells, we proposed the following theses to explore observation 1:

Thesis 1: *Pre-invasive bronchial epithelial lesions can be represented as a particle system where normal and abnormal cells represent competing populations.*

Thesis 2: *There exists a mathematically tractable particle system that can model characteristic structural behaviour of pre-invasive bronchial epithelial lesions as stochastic processes.*

Thesis 3: *A likelihood probability of growth of pre-invasive bronchial epithelial lesions can be determined from the analysis of a large sample of simulation results.*

From the obtained results after the exploration of Theses 1,2 and 3, and after further observation of the behaviour, progression and actual diagnosis of the disease by physicians, it became clear that for any computational diagnostic system to give reliable diagnostic advice, it would require to consider, not only structural information on individual lesions, but also would need to consider multidisciplinary data that would encompass the patient's medical and family history, lifestyle, imaging (X-rays, etc.), serum (blood samples) and genetic

information. Hence, I recognized the need to develop a formalism to represent and analyse this multidisciplinary data and their interactions. This new question led to the following observation:

Observation 2: *There are constraint-based cognitive formalisms capable to represent knowledge in separate ways — syntactic, semantic, pragmatic and other information of different kinds — while being able to process them simultaneously if needed.*

Based on this observation, we proposed the following theses:

Thesis 4: *There exists a formal characterization for the representation of multidisciplinary biological data concepts that would allow the interaction between concepts from different disciplines.*

Thesis 5: *Biological concepts naturally group into related, although not necessarily independent, partitions, which can decompose the knowledge base and simplify its representation.*

Thesis 6: *The relationships and interactions between multidisciplinary biological data concepts can be represented and analysed in terms of constraint systems.*

Thesis 7: *The relationships and interactions between multidisciplinary biological data concepts will impact the likelihood probability of development of a disease.*

1.1 Document Organization

Chapter 2 surveys the different approaches taken to develop software systems used for the analysis and diagnosis of lung cancer. Section 2.1 gives a brief biological overview of the lung cancer problem. Section 2.2 presents a 3-D computer model of bronchial epithelial lesions. Section 2.3 gives an overview of artificial intelligence research applied to the field of medicine. Section 2.4 briefly discusses current efforts on the development of standards for biological data representation. Section 2.5 presents an analysis of the different approaches to cancer diagnosis based on architectural or other type of analysis.

Chapter 3 presents a stochastic model of the development of pre-invasive neoplastic bronchial epithelial lesions. This model represents the lesion cell development as contact processes and shows how a simple mathematically tractable model can represent some interesting behaviour of lesion growth. It then presents the analysis of the results obtained from the simulations done using this model and discusses why a model only representing structural analysis of independent lesions is insufficient for accurate cancer diagnosis.

Chapter 4 introduces a novel approach for multidisciplinary biological knowledge representation: the Probabilistic Property-Based Model (PPBM). In this chapter, we discuss several knowledge representation methodologies and present the formalization of the PPBM. Also a prototype implementation as proof of concept shows how the PPBM can be used to develop a cancer diagnostic system. Finally, conclusions and future research directions can be found in chapter 5.

CHAPTER 2: COMPUTER-AIDED DIAGNOSIS OF CANCER: A SURVEY

2.1 Understanding Lung Cancer Development

The use of the word "cancer" (from the Latin word for crab) today signifies a generic term for any type of tumour that is malignant. A "tumour" (also called a "neoplasm") is also a common term used to illustrate an abnormal growth that has no useful function to the host. More specifically, the neoplasm or tumour can be defined as a mass that persists in the absence of a stimulus [Willis. 1967]. The growth of this mass is controlled by the cells inherent to the neoplasm and hence is coordinated differently from the surrounding tissues. Tumours are further classified as either benign or malignant.

Benign tumours are slow growing and, depending on the site where they are located, they do not normally cause death. While these cells do not usually show mitosis (process of division of body cells), they are well organized and differentiated, in as much as they resemble cells from normal tissues. According to [Phoenix5. 2002a], in cancer, differentiation is defined as: "How developed the cancer cells are in a tumour. Well differentiated tumour cells resemble normal cells and tend to grow and spread at a slower rate than undifferentiated or poorly differentiated tumour cells, which lack the structure and function of normal cells and grow uncontrollably." Benign tumours produce fewer molecules per cell (i.e., gene-specific products particular to cells of that type of tissue) than normal

tissues. However, once the tumour has increased in size, it can synthesize large amounts of molecules, which can be harmful to the host organism. For example, a benign tumour in the Islets of Langerhans may secrete an excess of insulin that can lead to insulin overdose, which can result in hypoglycaemia and potential death. As the tumour grows in size it pushes the normal tissue ahead of it, causing the thin capillaries of the normal parenchyma (functional cells) to be compressed. This results in insufficient blood to nourish the tissues, ultimately leading to the death of the normal cells and atrophy (loss of size of the tissue or organ). Once the normal cells die, only the connective tissues (stroma) are left, encapsulating the tumour. Nonetheless, benign tumours, in general, cause little or no damage to the host.

Malignant tumours, however, do have the ability to kill the host. Cells of these tumours are pleomorphic — they proliferate quickly and are quite different from benign tumour cells —. Malignant tumours show cells of normal and abnormal mitotic figures; they possess large vesicular nuclei with large nucleoli (these are the regions where ribosomes are made). Malignant cells are normally less differentiated and some cells are anaplastic, in other words, "when cells divide rapidly and bear little or no resemblance to normal cells in appearance or function" [Phoenix5. 2002b]. These anaplastic cells usually invade the surrounding tissues, destroying and substituting them with a mass of disorganized malignant cells.

Small clumps of cells from malignant tumours can then detach from the original mass (primary tumour) and travel to distant organs where they can

implant themselves as a secondary tumour, in a process called metastasis. The secondary tumour then develops stroma and causes the onset of new tumours, which in turn invade and metastasize.

2.1.1 Nomenclature

Tumours have their names derived from the tissue or organs from which they grow. Once this is done, in most cases, a suffix is added to the name to distinguish between benign or malignant. For example, a fibroma is a benign tumour because of the suffix *–oma* at the end of the name. Most malignant tumours on the other hand can be identified by three different suffixes: *-carcinoma* (tumour of epithelial origin), *-sarcoma* (tumour of connective tissues), and *–blastoma* (childhood tumours). As exceptions, some malignant tumours, which do not follow the standard nomenclature, include melanoma (pigmented skin cancer), hepatoma (liver cancer), and leukaemia (their suffixes imply benign though they are not).

2.1.2 Lung Cancer

The lung tissue can be seen as divided into three layers: the basal layer, where stem cells divide; an intermediate layer, which thickens as more abnormal cells are present; and the epithelial layer, which is the top layer where cells flatten and die. In a normal tissue, a stem cell divides and gives birth to two identical daughter cells. One of the new daughter cells stays in the basal layer and becomes a new stem cell, while the other daughter cell differentiates and will slowly move toward the epithelial layer where it will die. A clone is the set of cells

that are descendants of the same stem cell. On occasion, abnormalities may occur in a cell. Most of the times, the body has mechanisms that will simply stop the life cycle of such cell, however, there are a few cases in which the abnormal cell will not die and instead it and its clone can multiply out of control. This leads to a tumour.

2.1.3 Types of Lung Cancer

There are two main groups of lung cancers with several different subtypes, each of which grows and spreads at different rates, responds differently to treatment, and has different survival rates [NCERx Inc. 2004], [BC Cancer Agency. 2005], [Canadian Cancer Society. 2005]. A lung tumour is classified as primary or secondary. Primary disease originates in the lungs, while secondary disease has metastasised (i.e., originated in another organ and spread) to the lungs. Primary cases can be divided into two groups: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC).

2.1.3.1 Non-Small Cell Lung Cancer (NSCLC) Types

NSCLC accounts for approximately 85 percent of all cases of lung cancer. All NSCLC types are spread relatively slow, and patients have higher survival rates than those with the more aggressive SCLC [NCERx Inc. 2004].

Exposure to tobacco smoke (either directly or second-hand) and radon exposure are the leading causes of NSCLC. Surgical removal of the diseased tissue is the usual treatment option, although radiation therapy and chemotherapy may also be used, depending on tumour growth and degree of

spread. This category includes squamous cell, adeno-carcinoma and large cell undifferentiated cancer. Their characteristics are as follow: [NCERx Inc. 2004], [BC Cancer Agency. 2005], [National Women's Health Resource Center Inc. 2004a]

- Squamous Cell Carcinoma (Epidermoid): Squamous cell carcinoma causes 30 to 35 percent of all cases. Slow-growing, squamous cell carcinoma usually arises in the larger lobar and segmental bronchi of the central part of the lung, and diseased nodules tend to be clumped together. The precancerous phase may last several years during which chest X-rays appear normal yet cancerous cells are found in the sputum. Common treatment is surgery and radiotherapy, as patients with this type of lung cancer tend to respond better than those with other types.
- Adenocarcinoma: Adenocarcinoma cases represent 40 percent of lung cancer cases (most frequently diagnosed type of lung cancer). It usually beginning in the mucous glands. Women are more likely to develop adenocarcinoma than any other lung-cancer type. It is also the type most frequently seen in non-smokers. Tumour cells form recognizable glandular structures and have a higher risk of lymphatic and blood spread. It is the most frequently diagnosed peripheral cancer. Often associated with scarring of the lungs, it may be seen as a subpleural mass that invades the overlying pleura. The prognosis is poorer than for squamous cell carcinoma, except for patients with early

stage tumours. A subtype of adenocarcinoma called bronchioalveolar — or alveolar — cell lung cancer arises from the terminal bronchioles alveoli walls. This subtype is associated less frequently with smoking.

- Large Cell Carcinoma: The least common form of NSCLC, large cell carcinoma occurs in approximately five to fifteen percent of all cases. Tumours may develop in isolation, or in masses. They are characterised with large, abnormal-looking cells that tend to originate along the outer edges of the lungs. Their behaviour generally mimics adenocarcinomas, but with more aggressive spread. The prognosis is worse than for squamous carcinomas, and tumours tend to be large by the time of diagnosis.

2.1.3.2 Small Cell Lung Cancer (SCLC)

SCLC accounts for 20 to 25 percent of all diagnoses, and is most prevalent among smokers. SCLC is also called undifferentiated small cell or "oat cell" cancer, because malignant cells are generally oat-shaped, small and round, or oval. SCLC is the most aggressive of all lung cancers, and spreads quickly. In 60 to 70 percent of cases, the disease has spread to other organs by the time of diagnosis, even if the primary tumour is small. Once metastasized, SCLC is not a candidate for surgery, but does respond to multidrug chemotherapy integrated with radiation therapy. Due to its tendency to spread rapidly, the one-year survival rate for SCLC is only five percent. Yet, if a tumour is localized to the

chest, long-term cure is possible (20 percent). Prophylactic brain radiotherapy is usually employed in curative therapy.

2.1.4 Staging and Prognosis

Treatment options depend on both the type of tumour, and its staging classification. Staging is a way of determining the severity of the carcinoma, whether it has spread to other organs, and how it will respond to treatment.

2.1.4.1 NSCLC TNM Staging

The severity of NSCLC is determined using TNM staging. The disease is classified according to the extent of the primary tumour (T), the status of regional lymph nodes (N), and distant spread or metastases (M)

The extent of the cancer in each of these important areas is then described by means of a simple code in which numbers designate the absence of tumour or increasing levels of disease. The codes are as follows: [BC Cancer Agency. 2005], [National Women's Health Resource Center Inc. 2004a]

Tumour (1-4): T1 is a tumour less than three centimetres; T2 is a tumour larger than three centimetres; T3 is a tumour of any size that passes into the chest cavity, and is operable; T4 is inoperable.

Lymph Node (0-3): N0 is no lymph node involvement; N1 involves the lymph nodes in the lungs; N2 involves the nodes in the chest, oesophagus or windpipe; N3 involves those nodes above the collarbone.

Metastases (0 or 1): M0 is no indication of metastases; in M1, disease has spread to other organs.

NSCLC staging examples might read T2N3M1, or some other combination of the above stages. The TNM data gathered is then used to divide cancers into the following stages: [NCERx Inc. 2004], [BC Cancer Agency. 2005], [National Women's Health Resource Center Inc. 2004b]

Occult stage: Cancer cells are found in sputum, but no tumour can be found on x-rays.

Stage 0: Cancer is only found in a local area and only in a few layers of cells. It has not grown through the top lining of the lung. Another term for this type of lung cancer is *carcinoma in situ.*

Stage 1: Subdivided into 1A and 1B. The tumour is small, contained, and surgically removable with no lymph node involvement. Survival rates range from 57 to 67 percent.

Stage 2: Subdivided into 2A and 2B. The tumour remains operable, but malignant cells have spread to lymph nodes around the lung or to the surrounding chest wall. Survival rates are between 38 and 55 percent.

Stage 3: Subdivided into 3A (occasionally can be operated on) and 3B (usually cannot be operated on). The cancer has spread to the lymph nodes in the area that separates the two lungs (mediastinum); or to the lymph nodes on the other side of the chest or in the neck. The tumour has spread to the lung lining or into the chest cavity. Surgery may remain an option, with radiation

therapy and chemotherapy as other possible alternatives. Survival rates are approximately 23 percent.

Stage 4: Metastasis to other organs has occurred. Treatment options are confined to palliative care, and survival rates drop to 5 percent.

Recurrent: Cancer has come back (recurred) after previous treatment.

2.1.4.2 SCLC Staging

SCLC staging differs from the NSCLC TNM staging. The three stages for SCLC are *limited, extensive* and *recurrent*: [NCERx Inc. 2004], [National Women's Health Resource Center Inc. 2004b]

Limited stage: Cancer is found only in one lung and in nearby lymph nodes. Often limited SCLC is treated with the assumption that metastasis has occurred, just to be safe.

Extensive stage: Cancer has spread outside of the lung where it began to other tissues in the chest or to other parts of the body.

Recurrent stage: Recurrent disease means that the cancer has come back after it has been treated. It may come back in the lungs or in another part of the body.

Lung cancer survival rates for both limited and extensive SCLC are grim. Limited disease averages a life expectancy of fourteen months with treatment. Survival expectancy for extensive disease is usually under a year.

2.1.5 Lung Cancer Diagnosis

Currently the diagnosis of lung cancer involves a number of tests. First, patients undergo a thorough physical examination and the physician may want to perform one or several of the following laboratory or imaging tests:

- Sputum sample test: The sputum sample is matter from the throat and lungs, which is spit out through the mouth. This sample is sent for testing to determine if it contains bacteria, other infectious organisms, or cancer cells; cancer cells may be present in the sputum in certain types of lung cancer.
- Chest radiograph (X-ray): Used to see whether there are enlarged lymph nodes in the chest or a localized mass in the lungs.
- Computed tomography (CT or "CAT") scan: A computer-assisted technique that produces cross-sectional images of the body.
- Magnetic resonance imaging (MRI): A diagnostic method in which hydrogen ions within a patient's body are excited by exposure to a magnetic field. A computer processes the resulting signals to create an image of the chest to define the location and extent of lung involvement.
- Bronchoscopy: A visual examination of the windpipe and lung branches using a flexible scope performed by a pulmonologist. Bronchoscopy may involve washings of the respiratory tissues for cell analysis, brushings (using a small, brush-like device to gather cells

from the tissue lining the respiratory system), or biopsy (removal and examination of small amounts of tissue). If the bronchoscopy is still unrevealing, or "negative," a needle biopsy may be performed.

- Needle biopsy: May be performed, with CT-guidance, on suspicious areas in the lungs or pleura. Fine needle aspiration (FNA) uses a slim, hollow needle that is attached to a syringe. The needle is inserted into the suspicious mass and it is pushed back and forth to free some cells, which are aspirated (drawn up) into the syringe and are smeared on a glass slide for analysis. Large needle, or core biopsy, uses a large-bore needle to obtain a tissue sample for analysis.
- Bone scan: May also be performed to rule out suspicions of metastasis to the bones.

Once the physician diagnoses lung cancer, the next step is to determine if the patient is a candidate for surgery. The imaging studies (X-ray, CT scan, bone scan, etc.) are reviewed to rule out distant metastasis. If there is no evidence of metastasis, the patient may then undergo mediastinoscopy, a surgical inspection of the mediastinum (the tissues and organs of the middle chest, e.g., the heart and large vessels, windpipe, etc.). A small flexible device with a camera, called an endoscope, is inserted into the chest via an incision at the top of the sternum, and the chest cavity is then examined. The mediastinal lymph nodes usually are removed during this procedure. If the mediastinal lymph nodes are "negative" (do not contain any cancer cells), the patient may be a candidate for surgery. However, if mediastinal lymph nodes are "positive" (contain cancer cells) or are

abnormally large on imaging studies (suggesting tumour involvement), the patient is not considered to be a surgical candidate.

2.1.5.1 Cancer Markers

A lot of the recent molecular biology and genetics research in cancer has been focusing in discovery of biological markers (or biomarkers) — that is, molecular elements that are associated with the presence of cancer — for risk prediction and early detection of this disease. For diagnosis, additional tests may be performed to look for lung cancer biomarkers. For example, lung cancer may be indicated by abnormalities in the following:

- PTH (parathyroid hormone): Blood levels of PTH or PTH-related protein may help to distinguish lung cancer from cancer of the pleura or other diseases.
- CEA (carcinogenic antigen): A cancer-specific immune system protein that is present in many adenocarcinomas, including lung adenocarcinoma. Increased preoperative levels of CEA usually suggest a poor prognosis. A CEA level greater than 50 may indicate advanced stage lung cancer and should discourage treatment by resection.
- CYFRA21-1 (cytokeratin fragment 19): A protein marker of lung cancer.

2.1.5.2 Early Diagnosis

Until recently, the only diagnostic tool available to localize pre-malignant cellular alterations and early bronchial cancer was conventional white light fiber-optic bronchoscopy (FOB). Since only the relatively thick or polyploid lesions are visualized by FOB, only 29 percent of the lesions were actually visible to an experienced endoscopist [Nagamoto *et al.* 1993], [Woolner *et al.* 1984]. In an effort to overcome these problems, the British Columbia Cancer Agency and Xillix Technologies Corporation (Richmond, British Columbia, Canada) developed the Lung Imaging Fluorescence Endoscopic Device (LIFE) which utilizes differences in tissue autofluorescence to detect precancerous and carcinoma *in situ* lesions at a much higher rate than FOB [Hung *et al.* 1991], [Lam *et al.* 1993], [Lam *et al.* 2000].

An epithelium biopsy obtained during a bronchoscopy contains a vertical cross-section of the lung tissue including cells from all three layers of the tissue. From this sample, biopsies of the lesion are obtained from which the pathologists must predict whether the lesion will evolve into a malignant tumour or if it will regress or, at least, not evolve towards cancer.

LIFE has allowed easier detection of pre-invasive neoplastic bronchial lesions, which are believed to be the possible precursors of malignant tumours. The natural history of lung cancer development, from the initial genetic event through other multiple genetic changes, cell kinetics, cell-cell, and cell-host interactions, is not completely understood. New techniques (microdissection and polymerase chain reaction (PCR) amplification) and tools (quantitative cytology

and quantitative histology) are elucidating the neoplastic development process. These techniques are generally dealing with snapshots (biopsies, bronchial fragments) of a continuously evolving epithelium. Current understanding suggests that as pre-invasive neoplastic epithelial tissue becomes more likely to develop into an invasive neoplasia, quantitative genetic changes and genetic heterogeneity in the tissue occur. It is possible to measure selected changes in the genetic makeup of individual cells in a biopsy or tissue section. However, at present it remains impossible to determine the genetic relationship of all the cells in a pre-invasive neoplastic lesion during the development into invasive cancer. This knowledge would be required to completely understand the evolution of normal epithelium into invasive neoplasia. In an attempt to uncover such understanding, models have been developed, which try to simulate the initial stages of the neoplastic process and, most importantly, to try to simulate the development pathway from normal tissue to abnormal lesion. This simulated development of an abnormal lesion requires a model that takes into account not only the individual cell, but also the whole architecture of the tissue (the interaction of all the cells that together conform the tissue).

2.2 Modelling Bronchial Epithelial Lesions

A graphical computer model of the 3-D architecture of bronchial epithelial lesions was developed by Dr. Carole Clem *et al.*, [Clem *et al.* 1997a], [Clem *et al.* 1997b], [Clem and Rigaut. 1995] in order to refine hypotheses concerning the progressive spatial disorganization of the bronchial epithelium during the pre-invasive neoplastic process.

There are two main parts in this model. First, there is a static model that simulates the physical arrangement of cells in normal and pre-invasive neoplastic tissue of the bronchial epithelium. Secondly, there is a dynamic component, which simulates the continuously interacting nature of living tissue using the 3-D representation obtained from the static model as a starting point.

In the static part, the positions, sizes, shapes and orientations of the nuclei are used as a basis for the 3-D modelling of the architecture. The representation also takes into account the spatial arrangement of the nuclei, modelling several cell layers. The nuclei are modelled by tri-axial spheroids. The sizes of the major and minor axes of each nucleus are deduced from cytomorphometric analysis. A homogeneous 3-D Poisson point process is used to simulate the candidate-positions of nuclei. This point process is layered to take into account the different intensities on the different layers (basal, intermediate and epithelial). In addition, the model generates a random angle of orientation for each nuclear axis. Each newly-generated nucleus is then inscribed in a suitably oriented rectangular parallelepiped with faces parallel to the planes defined by the spheroid axes. If this parallelepiped has an intersection with a parallelepiped of any earlier generated nucleus, the newly generated candidate-position with its nucleus is deleted.

In order to determine whether the model's behaviour has an acceptable range of accuracy for its intended purpose of simulating the physical arrangement of cells in the tissue, the system computes the values of 2-D parameters from several computer "sections" through the simulated 3-D image.

An iterative process is used, based on statistical comparison between the 2-D parameters computed and those used from real (2-D) histological sections. If the t-test shows a statistically significant difference between the obtained values and the expected ones, the corresponding values are modified and the process is repeated until no statistically significant differences are found.

The dynamic part of the model can be seen as a tissue growth process applied to the 3-D representations obtained from the static model. Before applying this growth process, an initialization procedure is used in order to define the different cell types that can be found in the tissue (i.e., stem or differentiated cells). The simulated tissue can be considered as a closed volume where no cell, even if it is submitted to a force that pushes it out of the box, can leave except by passing into the epithelial layer. Each cell is defined by some internal states, which include its capacity of division, its position in the tissue, its age, its displacement capacity, its lifetime and its cell type. Under normal conditions, only the stem cells are able to divide and only the differentiated cells can migrate from basal to epithelial layer. At each time step, several events may occur: a stem cell can divide and a new cell can appear; the volume of a stem cell can increase; a differentiated cell can move towards the lumen; a collision between two cells can occur; a cell can die; or a nucleus can enter into pyknosis. All these events induce local and global modifications of the tissue architecture and require the model to check the structural stability of the tissue at each time step. Furthermore, all these processes, in order to occur, require an analysis of the local environment of the cell that is involved in one of these events.

Simulations of different diffusion patterns of abnormal cells within the bronchial epithelium during the pre-invasive neoplastic process have been obtained as well. This model has proven useful in providing insight into the development and architecture of bronchial epithelial lesions, however it does not provide any further extensions that could be used for diagnostic purposes.

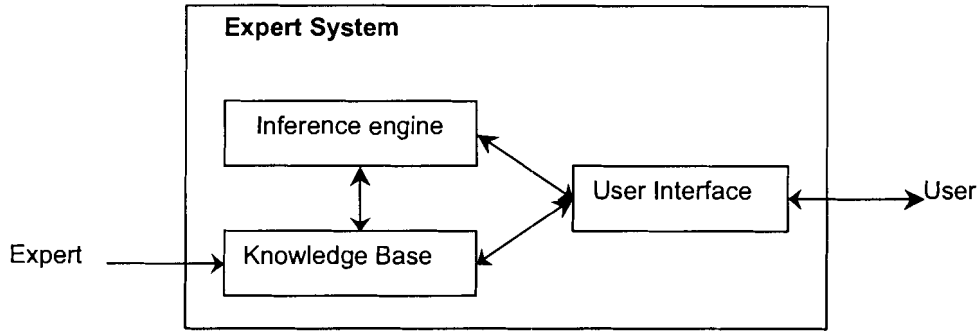
2.3 Artificial Intelligence in Medicine

Since the late 1950's, researchers in the field of artificial intelligence (AI) have been addressing problems in the field of medicine [Altman. 1999]. One of the first areas that drew the attention of AI researchers was that of medical diagnosis, as it contains many common reasoning tasks. The seminal paper of Ledley and Lusted [Ledley and Lusted. 1959] explained that medical reasoning contains well-recognized inference strategies, such as Boolean logic, Bayesian probability and symbolic logic, and that diagnostic reasoning could be formulated using these techniques. These concepts have influenced a lot of research over the last 45 years. Many computer systems have been developed that address important medical diagnosis issues. Here are some characteristic examples: PROMIS [Tufo *et al.* 1977] is amongst the first systems to implement a truly electronic medical record, which supported a problem-oriented medical information methodology; CASNET [Kulikowski and Weiss. 1982] used causal (*i.e.* physiological) models to explain symptoms and describe diagnostic possibilities; MYCIN [Buchanan and Shortliffe. 1984] used production rules to make expert-level diagnosis of infectious diseases; the PIP system (Present Illness Program) [Szolovits and Pauker. 1976] modeled the cognitive processes

of short- and long-term memory to develop programs that considered multiple diagnosis but quickly focused on the few most likely solutions. The INTERNIST/QMR [Miller *et al.* 1982], [Miller *et al.* 1986] is a knowledge base and inference program to diagnose any problem within internal medicine. Its knowledge base associated diseases with findings using a frequency of association and an evoking strength, which an algorithm then collected and computed the more likely diagnoses. A similar approached is followed by DXPLAIN [Barnett *et al.* 1987] and ILIAD [Bouhaddou *et al.* 1995].

From the early days of expert systems, rules have been the prime formalism for expressing knowledge in a symbolic way. They offer simplicity, transparency, uniformity and ease of inference, which make them very attractive to represent medical knowledge obtained from a human expert (Figure 2.1) However, it quickly became evident that the knowledge acquisition is the most complex part of the development of expert systems. Rules obtained from a human expert risk capturing the biases of one single person, as rules will be the formulation of that particular expert's "rules of thumb" on the subject. Even though they may appear as a coherent and modular set of knowledge, they may reveal inconsistencies, gaps and other problems.

Figure 2.1: Schema of an expert system of the early 80's



Source: Based on Figure 1 from [Lavrac *et al.* 2000]

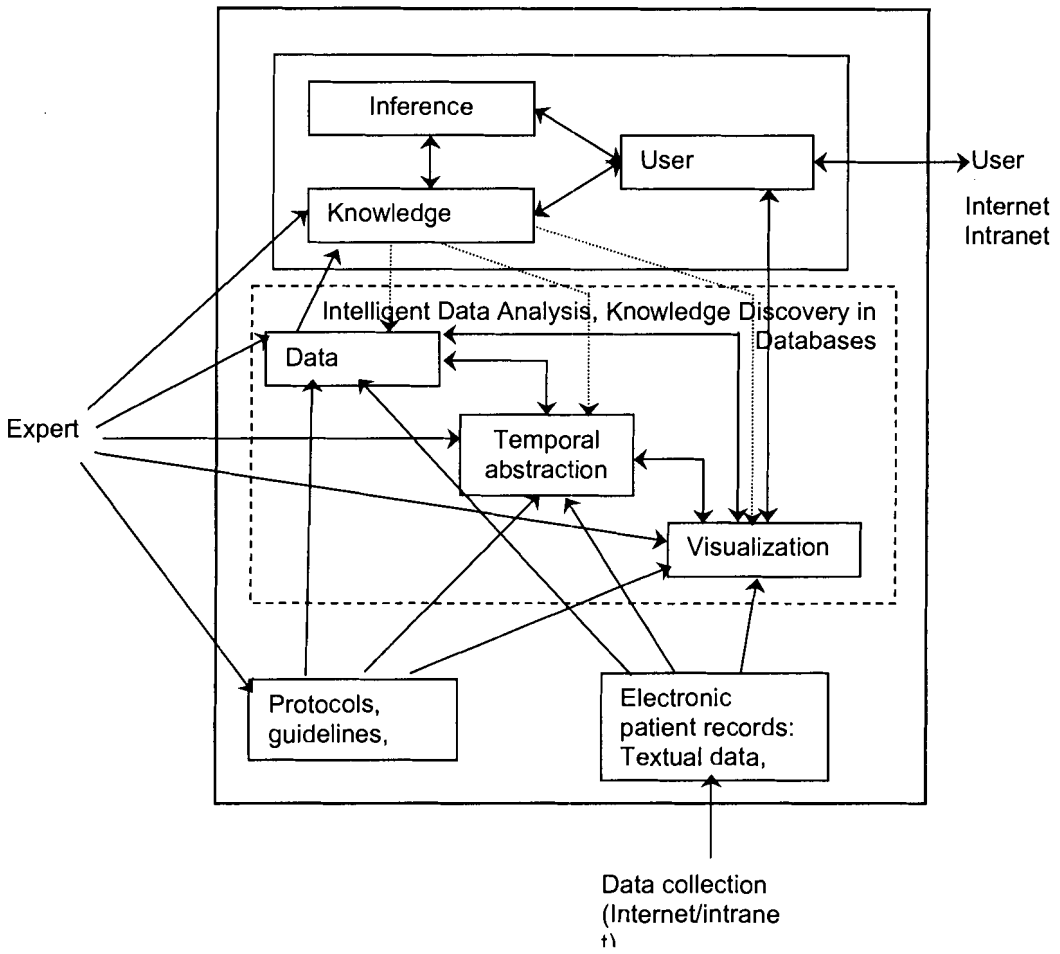
These limitations along with the high costs of acquiring the knowledge directly from the experts and the increasing availability of databases of sample cases shifted the focus to the consideration of the learning of rules from such databases. Appearing less biased, more efficient and more cost-effective this option was attractive. This led to the development of machine-learning algorithms that would extract rules automatically from existing data. Since real-life data may contain errors or be incomplete, the algorithms created need to deal with such imperfections. [Michalski. 1983] and [Quinlan. 1983] are examples of the early works on extraction of decision trees from pre-existing data. CART [Breiman *et al.* 1984], ASSISTANT [Bratko and Kononenko. 1987], [Cestnik *et al.* 1987], AQ [Michalski *et al.* 1986], ID3 [Quinlan. 1986], CN2 [Clark and Niblett. 1989], [Clark and Boswell. 1991], and C4.5 [Quinlan. 1993] are cases of machine-learning systems that deal with noisy, real-life data. Human experts are still actively involved in the development of machine-learning systems providing the sample cases and verifying the resulting rules. The advantage of the learning approach is that it ensures the resulting rules are hierarchically organized and consistent

as a consequence of decision trees. The difficulty is that the sample data set must be complete enough for the domain in question (without significant gaps in the knowledge), otherwise the resulting rules will not provide adequate coverage and/or sufficient accuracy.

Research in the last decade has been characterized by the efforts to bridge the gap between the large amounts of un-interpreted data and the understanding of such data. Thus, the research emphasis is now on data analysis. Data mining, knowledge discovery in databases [Frawley *et al.* 1991], and intelligent data analysis, along with machine learning techniques, are the latest focus areas of medical computing research.

The need for intelligent data analysis in medicine is evident in the following: (i) for instance, to support the analysis of individual patients' raw data of specific knowledge-based problem solving activities such diagnosis, prognosis, monitoring, treatment planning, etc. and (ii) the use of data mining in the discovery of new medical knowledge that can be extracted from collections of example cases. Figure 2.2 represents a possible schema of a recent decision support system. In this schema, large volumes of data have to be processed, e.g., patient records comprising images and textual data (possibly transferred through the Internet or an intranet), protocols, guidelines, etc. The solid arrows denote the normal flow of information and the dotted arrows show the flow of information in processes that involve loops and iterations between the different steps of the intelligent data analysis process.

Figure 2.2: Schema of a decision support system of the late '90s



Source: Based on Figure 2 from [Lavrac *et al.* 2000]

2.4 International Standards for Biological and Medical Data Representation

According to E.H. Shortliffe in [Shortliffe. 1993], the successful integration of AI systems into patient care settings may be influenced by the following three factors: international standards, enhancement of training, and information infrastructure. Of these factors, the establishment of international standards is the one with the most impact to researchers working in the development of said

systems. For example, during the panel discussion of the Artificial Intelligence in Medicine Europe conference (AIME 97) the following important issues that arise from the emerging globality of information and data were identified [Lavrac *et al.* 2000]:

- *The provision of standards in terminology, vocabularies and formats to support multilinguality and sharing of data,*
- *standards for the abstraction and visualization of data,*
- *standards for interfaces between different sources of data,*
- *integration of heterogeneous types of data, including images and signals,*
- *standards for electronic patient records, and*
- *reusability of data, knowledge and tools.*

Over the last few years, numerous efforts have been taken place to establish international standards for medical and biological data representation.

2.4.1 UMLS

The Unified Medical Language System (UMLS) project began in 1986 by the National Library of Medicine (NLM) [National Library of Medicine. 2004]. The main objective of this long-term research and development project is to develop "knowledge sources" to make it easy for users to link disparate information systems, including computer-based patient records, bibliographic databases, factual databases, and expert systems, and to overcome the retrieval problems caused by differences in representation and the scattering of relevant information across many databases.

UMLS comprises three knowledge sources:

- UMLS Metathesaurus: Provides a uniform, integrated distribution format from over 100 biomedical vocabularies and classifications (the majority in English and some in multiple languages), linking many different names for the same concepts. The Metathesaurus has been distributed since 1990.
- SPECIALIST Lexicon: Contains syntactic information for many terms, component words, and English words, including verbs, which do not appear in the Metathesaurus.
- UMLS Semantic Network: Contains information about the types or categories to which all Metathesaurus concepts have been assigned and the relationships allowed among these types.

The NLM has also developed the "MetamorphoSys" software, which is useful in producing customized versions of the Metathesaurus. This software facilitates the exclusion of any vocabulary considered inappropriate or irrelevant by the UMLS user. The NLM and many other institutions have developed several applications using the UMLS Knowledge Sources, such as patient data creation, curriculum analysis, natural language processing, and information retrieval.

2.4.2 SBML

Systems Biology Markup Language (SBML) [Hucka *et al.* 2003] was developed by the Caltech unit of the ERATO Kitano Symbiotic Systems Project, with frequent input from the open-source community. SBML is a description

language for simulations in systems biology. SBML is oriented towards representing biochemical networks common in research on a number of topics, including cell signalling pathways, metabolic pathways, biochemical reactions, gene regulation, and many others. It is mostly useful for exchange of models between different software. This representation, based on Extensible Markup Language (XML), consists of formalized statements about the different components of the model of a system of biochemical reactions. It can represent: (i) *species* of chemical substances taking part in a reaction, (ii) *compartments* in which the species are located, (iii) chemical *reactions* affecting the species, (iv) *parameters* representing numerical variables, (v) definitions of *units* on numerical quantities, (vi) definitions of mathematical *functions* used in formulas, (vi) discrete *events* presenting changes in the system's state, and (vii) additional mathematical *constraints* on the system.

2.5 Software Tools to Assist Diagnosis

2.5.1 Image analysis research

Most of the work done on computer-assisted diagnosis (CAD) of lung cancer has mainly focused on image analysis technologies. Many computer-aided image analysis systems use artificial neural networks to identify nodules in chest radiographs, CT scans, etc. Artificial neural networks may be used for both supervised and unsupervised learning. For modelling medical data and CAD, the most frequently used neural network supervised learning paradigm is the feed-forward, multilayered neural network [Rumelhart and McClelland. 1986], [Fausett. 1994]. These computational structures consist of interconnected processing

elements or nodes organized in a multilayered hierarchical architecture. Each node calculates the weighted sum of its inputs and produces its output by filtering this data through a sigmoid function. The outputs of the nodes from one layer serve as inputs for the nodes of the next layer and the weights that are associated with each node are determined from training instances. There are several learning algorithms but the most popular is backpropagation. This algorithm first sets the weights of the nodes to an arbitrary value, then considering one or more training instances at a time, adjusts the weights so that the difference between the expected values and those actually obtained at the output level is minimized. This is repeated until the overall classification error falls below some specified threshold.

As examples of CAD systems, [Lin *et al.* 1995] have developed a system based on a parameterized two-level, convolution artificial neural network, on a special multi-label output encoding procedure, which was used in the diagnosis of lung cancer nodules found on digitized chest radiographs. [Yoshida *et al.* 1997] developed a system that used snake wavelet transforms to isolate the nodules embedded in the background structures also on digitized chest radiographs. A measure that represents the goodness of this nodule identification process is then combined with some morphological features, in order to train an artificial neural network for effective distinction between nodules and false positives. [Hayashibe *et al.* 1996] proposed an automatic method based on the subtraction between two serial mass chest radiographs, which was used in the detection of new lung nodules. [Penedo *et al.* 1998] developed a system that

employed two artificial neural networks in the detection of lung nodules found on digitized chest radiographs. The first artificial neural network was utilized to detect suspicious regions in a low-resolution image and then another one used to deal with the curvature peaks of the suspicious regions. [Chiou *et al.* 1993] designed an artificial neural network based hybrid lung cancer nodule detection system, which was used to improve the accuracy and the speed of diagnosis of lung cancer from pulmonary radiology. [Zhou *et al.* 2002] proposed an automatic pathological diagnosis procedure that utilizes an artificial neural network ensemble to identify lung cancer cells in the images of the specimens of needle biopsies. The ensemble is built on a two-level ensemble architecture. The first-level ensemble is used to judge whether a cell is normal with high confidence, where each individual network has only two outputs respectively *normal cell* or *cancer cell.* The predictions of those individual networks are combined using a method that judges a cell to be normal only when all of the individual networks judge it is normal. The second-level ensemble analyses the cancer cell outputs from the first-level ensemble, where each individual network has five outputs respectively *adenocarcinoma*, *squamous cell carcinoma*, *small cell carcinoma*, *large cell carcinoma*, and *normal.* The predictions of those individual networks are combined by a prevailing method. [Kanazawa *et al.* 1996] presented a system that extracted and analyzed features of the lung and pulmonary blood vessel regions from helical CT images and then utilized defined rules to perform diagnosis, which was used in the detection of tumour candidates.

These systems provide valuable information about the patient's condition based on the structure of the lesion however, as we discussed previously, at the very early stages, the lesions are so small that structurally it is difficult to discern which ones are malignant and which are not. Therefore, it is important to consider other symptoms in addition to the lesion's morphological features to be able to provide a more accurate early diagnosis.

2.5.2 Other approaches

In [Fretz and Peterson. 1996] a multidisciplinary database to assess the risk of malignancy in a solitary pulmonary nodule is presented. The authors, based on the research findings in [Cummings *et al.* 1986] and [Gurney. 1993], determined likelihood ratios for the incidence of various clinical and radiographic features of a lung nodule, where

Likelihood Ratio = Probability in patients with disease / Prob. in subjects without disease

= Sensitivity/(1 – specificity).

They then used Bayes' theorem, where

Current Odds = Prior Odds x Likelihood Ratios;

to determine the probability of malignancy in solitary pulmonary nodules. The main difference of this approach from others is that they considered the population incidence of malignancy, patient age, smoking history, haemoptysis (coughing up of blood from the respiratory tract), and previous history of malignancy; in addition to the structural information, such as: nodule size, nodule

location, edge characteristics on chest x-ray, cavity wall thickness, and calcification pattern on CT scan.

This approach is interesting with respect to the incorporation of other factors in addition the structural information, however, it is oversimplified and the data analysis is a very simple statistical analysis, which provides little additional insight to the physician making the diagnostic. As well, the system requires that all information requested be provided in order to make a diagnosis, which in some cases is not possible, as some tests may have not yet been performed.

2.5.2.1 Use of Biomarkers in CAD

The Early Detection Research Network [Early Detection Research Network. 2002], an international scientific consortium funded in 1999 and coordinated by the National Cancer Institute's Division of Cancer Prevention in the US, has developed the following detailed guidelines to ensure good practice in the design and analysis of nested, case-control studies of early detection biomarkers: [Baker *et al.* 2002]

1. *"For the clearest interpretation, statistics should be based on false and true positive rates – not odds ratios or relative risks.*

2. *To avoid overdiagnosis bias, cases should be diagnosed as a result of symptoms rather than on screening.*

3. *To minimize selection bias, the spectrum of control conditions should be the same in study and target screening populations.*

4. *To extract additional information, criteria for a positive test should be based on combinations of individual markers and changes in marker levels over time.*

5. *To avoid overfitting, the criteria for a positive marker combination developed in a training sample should be evaluated in a random test sample from the same study and, if possible, a validation sample from another study.*

6. *To identify biomarkers with true and false positive rates similar to mammography, the training, test, and validation samples should each include at least 110 randomly selected subjects without cancer and 70 subjects with cancer."*

Only in very recent years has any work been done using computers to identify biomarkers. As an example of this, In [Markey *et al.* 2003] a decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer is presented. They use a classification and regression tree (CART) model that was trained to classify 41 clinical specimens as disease/non-disease based on 26 variables computed from the mass-to-charge ratio *(m/z)* and peak heights of proteins identified by mass spectroscopy. The CART model built on all of the specimens (no cross-validation) had an error rate of 10%. This model suggested that mass spectra peaks in the 8000–10000, 20000–30000, 45000–60000, and >125000 *m/z* ranges may be valuable in distinguishing between the disease and non-disease specimens.

2.6 Summary

Lung cancer is a very aggressive disease that is very hard to diagnose at its early stages even with all the new advances in molecular biology and imaging technology. Almost all the current research and development done on CAD systems has focused exclusively in one aspect of the diagnostic process, namely

on image analysis. It has become clear to us, however, that a multidisciplinary approach is required in order to obtain a more accurate diagnosis at the early stages of the disease as image analysis results provide little information prior to the presence of tumours and, once tumours are present, the prognosis is not favourable.

Another critical area that must be considered is that of data interactivity and the use of international standards. Large amounts of biological and medical data are becoming available and consequently it is important to adhere to standards so that this wealth of information can be usable, not only by one single application, but by many.

In the upcoming chapters, we will show how first we developed a stochastic model to analyse the likelihood probability of growth of independent lesions. From the insights obtained from those results we will explain how we determined that it is necessary to consider multidisciplinary patient data to perform a more accurate diagnosis. Therefore, we will show how we incorporated the available patient information — health history records, imaging, genetic, serum and sputum cytology data — into a knowledge representation model, the Probabilistic Property-based Model (PPBM), that would consider all this information to estimate the likelihood of the patient to develop cancer. PPBM is able to calculate the probability even if only partial information is given and will recommend follow-up tests to be performed to improve the accuracy of the diagnosis. Also, PPBM is sufficiently flexible to allow the addition or modification of the knowledge base as new scientific and medical information becomes

available, in particular in the area of cancer biomarkers, where a lot of the current biological research is focusing in.

CHAPTER 3: STOCHASTIC MODEL OF THE DEVELOPMENT OF PRE-INVASIVE NEOPLASTIC BRONCHIAL EPITHELIAL LESIONS

3.1 Introduction

As we mentioned earlier, lung cancer accounts for 28% of all cancer deaths in North America. However current treatment options lead to a cure in only 10% of diagnosed cases. It is well known that the survival rates can be improved by the early detection of pre-invasive lesions, which are believed to be the possible precursors of malignant tumours. Although, as we showed, new technology is allowing numerous early lesions to be detected, it is becoming clear that only a small percentage of these will progress to cancer. As explained in section 2.1.5, one of the ways to obtain an early diagnosis of lung cancer is to perform a bronchoscopy and obtain a needle biopsy from identified lesions. A fundamental problem in the analysis of biopsies — longitudinal two-dimensional (2-D) sections of the central area of the lesions— is the quantification of tissue heterogeneity. One can distinguish abnormal cells from normal cells and analyse their spatial arrangement, but it is currently impossible in the case of pre-invasive lesions in the early stages, that is, when just a few abnormal cells are present in the biopsy, to tell if one observed pattern is more aggressive than another one.

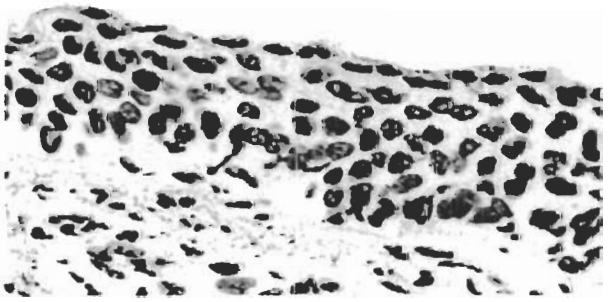
In this chapter a stochastic model for the growth of pre-invasive neoplastic bronchial epithelial cells is presented. The results are analysed to differentiate

progressive lesions from regressive ones given a particular biopsy. The problem is initially simplified to taking a one-dimensional (1-D) cross-section from a 2-D process and estimating the maximum likelihood rate of growth based on this limited information. We propose a number of extensions to eventually extend this approach to the higher dimension model by analysing 2-D sections and estimating their growth rates in a three-dimensional (3-D) process.

3.2 The Biological Problem

Pathologists diagnosing lung cancer in a patient must consider the global architecture of the bronchial tissue, as well as the local architecture of cells and the appearance of individual cells. In order to obtain more detailed information on the condition of the bronchial tissue, a bronchoscopy is performed on the patient and tissue samples of any detected lesion are obtained. The pathologist obtains 2-D sections of the samples extracted from a region of the tissue containing abnormal cells (Figure 3.1) and from these sections the diagnosis must be made.

Figure 3.1: 2-D section of a pre-invasive neoplastic epithelial lesion



Source: Dr. Martial Guillaud, Cancer Imaging Laboratory, BCCRC, with permission

Many times an abnormal cell will die naturally without forming a cancer so the pathologists are concerned only with identifying the cells that will eventually develop into cancer. Since currently available cancer treatments are very aggressive and traumatic for the patient, pathologists want to be fairly certain that the abnormal cells present in the sample will lead to cancer before recommending treatment. When the sample contains large amounts of abnormal cells or none at all the diagnosis is simple; however, in many cases there are just a few abnormal cells in the sample and diagnosis is difficult.

The lung tissue can be seen as divided into three layers: the basal layer, where stem cells divide; an intermediate layer, which thickens as more abnormal cells are present; and the epithelial layer, which is the top layer where cells flatten and die. The tissue is about 10 cells in thickness. In a normal tissue, a stem cell divides and gives birth to two identical daughter cells. One of the new daughter cells stays in the basal layer and will become a new stem cell, while the other daughter cell differentiates and will slowly move toward the epithelial layer where it will die. A clone is the set of cells that are descendants of the same stem cell. On occasion, abnormalities may occur in a cell. Most of the times, the body has mechanisms that will simply stop the life cycle of such cell, however, there are a few cases that the abnormal cell will not die and instead it and its clone will multiply out of control. This will lead to a tumour being produced. The epithelium biopsy obtained during the bronchoscopy will contain a vertical cross-section of the lung tissue including cells from all three layers of the tissue. From this

sample, 2-D biopsies of the lesion are obtained from which the pathologists must predict whether the lesion would evolve into a malignant tumour or it would, if not regress, at least not evolve towards cancer.

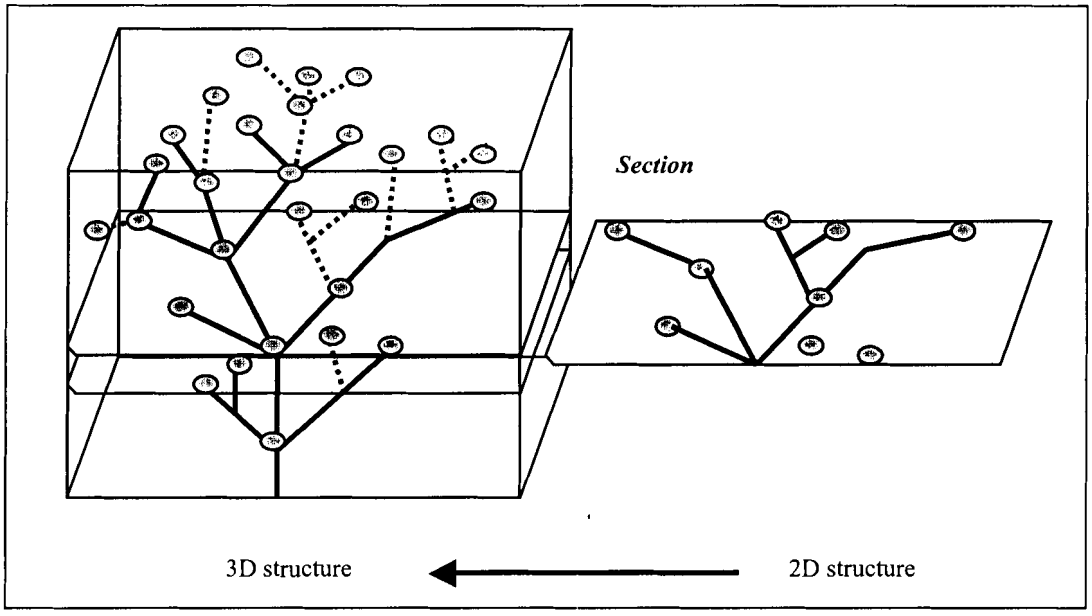
As part of the Pacific Institute of Mathematical Sciences Industrial Problem Solving (PIMSIPS) Workshop (Vancouver, BC, September 1997), Drs. Carole Clem and Martial Guillaud, two members of the Cancer Imaging Laboratory of the British Columbia Cancer Research Centre (BCCRC), presented the graphical computational model of pre-invasive bronchial epithelial lesions discussed in section 2.2. This model presupposes a large number of parameters designed to accurately reflect the biological process. [Clem *et al.* 1997a], [Clem *et al.* 1997b], [Clem and Rigaut. 1995] [Clem *et al.* 1992] Unfortunately, the complexity of the model renders it mathematical intractable and, therefore, does not allow for the required flexibility to be modified and used to develop a system that could aid in the early diagnosis of lung cancer. In the PIMSIPS Workshop, Drs. Guillaud and Clem asked the following question:

"Suppose a 2-D cross-section of tissue form the model is presented. What information can be determined about the original lesion? In particular, is it possible to predict the structure of the three-dimensional (3-D) lesion accurately enough to determine whether the lesion will progress towards cancer?" (Figure 3.2) [Barranco-Mendoza *et al.* 1997]

The objective of our research was to develop a new model that maintains the principal characteristics of the biological process but that is sufficiently simple

to allow the analysis of 2-D biopsies to determine the initial 3-D structure from which they were taken and try to predict the rate of growth of the lesion.

Figure 3.2: Determine the initial 3-D structure from the 2-D section



Source: Dr. Martial Guillaud, Cancer Imaging Laboratory, BCCRC, with permission

3.3 Assumptions

As with any tractable mathematical model, there were a number of assumptions that had to be made about the biopsy procedure and cell behaviour [Barranco-Mendoza *et al.* 1997], [Barranco-Mendoza and Gupta. 1998], [Barranco-Mendoza *et al.* 1999]:

1. The biopsy section consists of a vertical plane of the bronchial tissue comprising cells from the three layers: basal, intermediate and epithelial.

2. Any given lesion starts with only one abnormal cell; it is reasonable to assume that the probability of more than one abnormal cell evading all natural check points at the same time and multiplying to form part of the same lesion is negligible.
3. The biopsy section will contain the site where the lesion began with high probability.
4. Normal cells are formed at the basal layer through cell division of stem cells. They then have a tendency to "drift" towards the epithelial layer. As a simplification to the problem, we will assume that the stem cells will always remain static at the basal layer and the new cells will be the ones moving upwards. This assumption is valid since mother and daughter cells are identical.
5. When a cell reaches the epithelial layer it dies.

3.4 The General Approach

As explained in [Barranco-Mendoza *et al.* 1997], there are a large number of parameters to consider in solving this problem. However, it is necessary to postulate a simpler model than Clem's if we wish to implement a mathematically tractable system. To gain a deeper understanding of the underlying mathematical model, we chose to begin by modelling a 2-D process. That is, we consider a 2-D lung in which lesions are formed. We can then study the problem of taking a 1-D cross-section (a biopsy) and determining the expected structure of the 2-D

process. As a first step, our initial model assumes that cells divide with a fixed probability and that cells cannot move once they are formed (that is, there is no lateral or upward movement).

3.5 Static Two-Dimensional Model

As shown in [Barranco-Mendoza *et al.* 1997], [Barranco-Mendoza and Gupta. 1998], [Barranco-Mendoza *et al.* 1999], we worked on a square lattice with non-negative coordinate points. Assume an initial abnormal cell at position (0,0). Given an abnormal cell at position (i,j), an abnormal cell will occur at position $(i,j+1)$ with fixed probability p $(0<p<1)$, and independently at position $(i+1,j)$ with the same fixed probability p $(0<p<1)$.

Since the total height of a cross-section in the 3-D case is at most ten, we will only allow cells to occupy lattice positions (i,j) such that

$$0 \le i+j < 20,\ 0 \le i,j < 10.$$

That is, we restrict to a height of 10 cells along the diagonal, *i.e.*, the points (0,0) , (1,1) , … , (9,9).. We will assume that a lesion with height of more than 10 cells can be considered aggressive enough to require medical treatment without further analysis. Notice that since there are only a finite number of configurations of abnormal cells, it would be possible to enumerate all configurations and assign each a probability (as a function of p).

By the definition of the model, it can be determined that the conditional probabilities of an abnormal cell occurring in location (m,n) in a configuration when only $(m,n\text{-}1)$ or only $(m\text{-}1,n)$ or both occur in the configuration are:

Equation 3.1:

$$P[(m,n)|(m,n\text{-}1) \wedge \neg(m\text{-}1,n)] = P[(m,n)|(m\text{-}1,n) \wedge \neg(m,n\text{-}1)]=p,$$

$$\forall m,n\ ,\ 0 < m,n \leq r$$

$$P[(m,n)|(m,n\text{-}1) \wedge (m\text{-}1,n)] = p^2,\ \forall m,n\ ,\ 0 < m,n \leq r,$$

where r is the maximum number of cells allowed in the cross-section.

Also by the definition of the model, it can be determined that:

Equation 3.2:

$$P[(m,n)] = P[(n,m)],\ \ \forall m,n\ ,\ 0 < m,n \leq r$$

where $P[(m,n)]$ is the probability of a cell occurring in location (m,n) in a configuration. Hence, from the above, $\forall m,\ 0 \leq m \leq r$, and $\forall n,\ 0 \leq n \leq r$:

Equation 3.3:

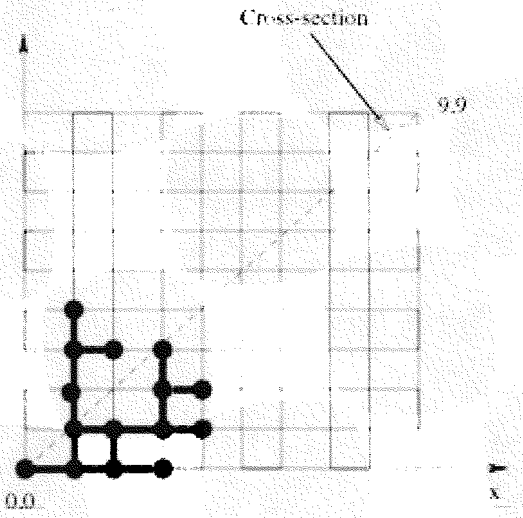
$$P[(m,n)] = \begin{cases} p^n, \quad if\ m = 0\ (or\ p^m, \quad if\ n = 0); \\ 2pP[(n,n-1)] - p^2P[(n,n-1)]^2, if\ m = n\ and\ n \neq 0; \\ p \times \{P[(m,n-1)] + P[(m-1,n)] - p \times P[(m,n-1)] \times P[(m-1,n)]\}, \\ if\ n \neq m\ and\ 0 < m,n \leq r \end{cases}$$

A cross-section of a configuration (the only information we assume is available) will be the line passing through the points (0,0) and (r,r). (Figure 3.3).

Based on the assumption that a lesion with height of more than 10 cells can be considered aggressive enough by the pathologists to recommend treatment without further analysis, the total height of a cross-section specific to

our research is at most 10 cells. Hence, we will only allow cells to occupy lattice positions (i,j) such that $0 \leq (i + j) < 20$, $0 \leq i,j < 10$. That is, we restrict to a height of 10 along the diagonal, i.e., the points (0,0), (1,1) ... (9,9).

Figure 3.3: An example growth pattern on the lattice with $p = 0.6$



Source: © Alma Barranco-Mendoza 2005

The specific problem that we addressed in [Barranco-Mendoza *et al.* 1999] is the following:

Given a cross-section of a configuration generated using some probability p, find an interval $[p_L,p_U]$ such that 90% of the time, in repeated experiments, similar intervals will contain the true value p, that is, find an approximate 90% confidence interval for the parameter p.

Notice that we can view a cross-section as a sequence of 0's and 1's of length r where the first element is always 1 and represents the abnormal cell at (0,0). The sequence for the cross-section in Figure 3.3 would be 1101000000 since there are abnormal cells at positions (0,0), (1,1) and (3,3). We denoted a

cross section by a *r*-tuple ($X_0, X_1, \ldots, X_r$) where $X_i=0$ if a normal cell is in position i on the cross section and $X_i=1$ if an abnormal cell is in position i.

3.6 Texture Models

The problem in question can be viewed as a texture model. Texture can be defined in stochastic terms as the spatial distribution of intensities with a two-dimensional random field. This stochastic approach is described in [Faugeras and Pratt. 1980]:

"The stochastic formulation is based on a model in which a texture region is viewed as a sample of a two-dimensional stochastic process describable by its statistical parameters."

and is also described by [Cross and Jain. 1983]:

"We consider a texture to be a stochastic, possibly periodic, two-dimensional image field."

The literature distinguishes between stochastic and structural models of textures. In [Smith. 1998], the author divides stochastic texture models into three major groups: Probability Density Function (PDF) models, Gross Shape models and Partial models.

The PDF methods model a texture as a random field. They fit a statistical PDF model to the spatial distribution of intensities in the texture. Typically, these methods measure the interactions of small numbers of pixels. PDF models are divided into two groups: those that use a parametric PDF model, and those that use a non-parametric PDF model. Examples of PDF models are the Gauss-

Markov Random Field [Hao Chen and Chen. 2002], [Chellappa *et al.* 1985], [Manjunath *et al.* 1990] and Grey Level Co-occurrence methods [Clausi and Zhao. 2002], [Haralick *et al.* 1973], [Gotlieb and Kreyszig. 1990].

Gross shape methods model a texture as a surface. They measure features, which a viewer would consciously perceive, such as the presence of edges, lines, intensity extrema, waveforms and orientation. These methods measure the interactions of larger numbers of pixels over a larger area than is typical in PDF methods. Gross shape methods are divided into three groups: Harmonic methods, Primitive methods, and Blob and Mosaic methods. Harmonic methods measure periodicities in the texture. They look for perceptual features that recur at regular intervals, such as a waveform. Harmonic methods measure spatially dispersed features of the texture, e.g., auto-correlation methods [Faugeras and Pratt. 1980]. Primitive and Blob and Mosaic methods measure spatially compact features of the texture. Primitive methods detect a set of spatially compact perceptual features, such as lines, edges and intensity extrema. The output of the feature extraction stage, the feature vector, is composed of the density of these perceptual features, in the texture, e.g., mathematical morphology methods [Haralick. 1979].

Mathematical morphology methods generate transformed images by erosion and dilation with structural elements. These structural elements correspond to texture primitives. For example, consider a binary image with pixel values on and off. A new image can be formed, in which a pixel is on if the corresponding pixel, and all pixels within a certain radius, in the original image

are on. This transformation will erode regions in which the pixels are on; in the terminology of mathematical morphology, this transformation is an erosion with a circular structural element. The number of on pixels in the transformed image will be a function of the texture of the original image and of the structural element.

In contrast, Blob and Mosaic methods detect a single perceptual feature. The feature vector is composed of the properties measured from instances of this feature, such as the average elongation and orientation of blobs, e.g., [Voorhees and Poggio. 1988] and [Chen *et al.* 1995] describe methods in which features are extracted from non-contiguous blobs.

Partial methods focus on some specific aspect of texture properties at the expense of other aspects. This group includes Fractal methods and Line methods. Fractal methods explicitly measure the how a texture varies with the scale it is measured at, but do not measure the structure of a texture at any given scale, e.g. [Wu *et al.* 1992]. Line methods measure properties of a texture along one-dimensional contour in a texture, and do not fully capture the two-dimensional structure of the texture.

Primitive methods are also related to structural texture methods; both methods model textures as being composed of primitives. However, structural models differ in two significant ways. Firstly, structural models tend to have one arbitrarily complex primitive, whereas primitive methods model texture as composed of many, simple primitives. Secondly, the relative placement of primitives is important in structural models, whereas it plays no role in primitive methods. The model here presented can be categorized as a primitive method.

3.7 Experimental results

The static model described above was developed and run through computer simulations for various values of p. Experimental results showed that at $p = .64$ there is a threshold effect - any smaller p results in mainly small configurations whereas any larger p results in most configurations having some cells that reach the last level.

Using Monte Carlo simulation methods, generating a million 2-D lattices for each value of p between .1 and .9, incrementing by .01, we were able to determine the probability of an abnormal cell being at a particular lattice point in the cross-section, for all points in the diagonal slice under consideration, that is, we were able to find the marginal distributions of the random variables

$$X_{i,i}=0,1, \ldots ,9.$$

Our strategy for estimating p was to choose as a test statistic the last grid point along the cross-section that was occupied (i.e., the position of the right-most 1 in a cross-sectional sequence, or $max(i : X_i = 1)$). The rationale for choosing this value as a test statistic was that it was evident from running simulations that the last grid point was highly dependent on the true value of p and therefore, given this information, estimation of the parameter p could be relatively precise. Furthermore, this is a single (univariate) random variable, which can more easily be studied thoroughly using simulation.

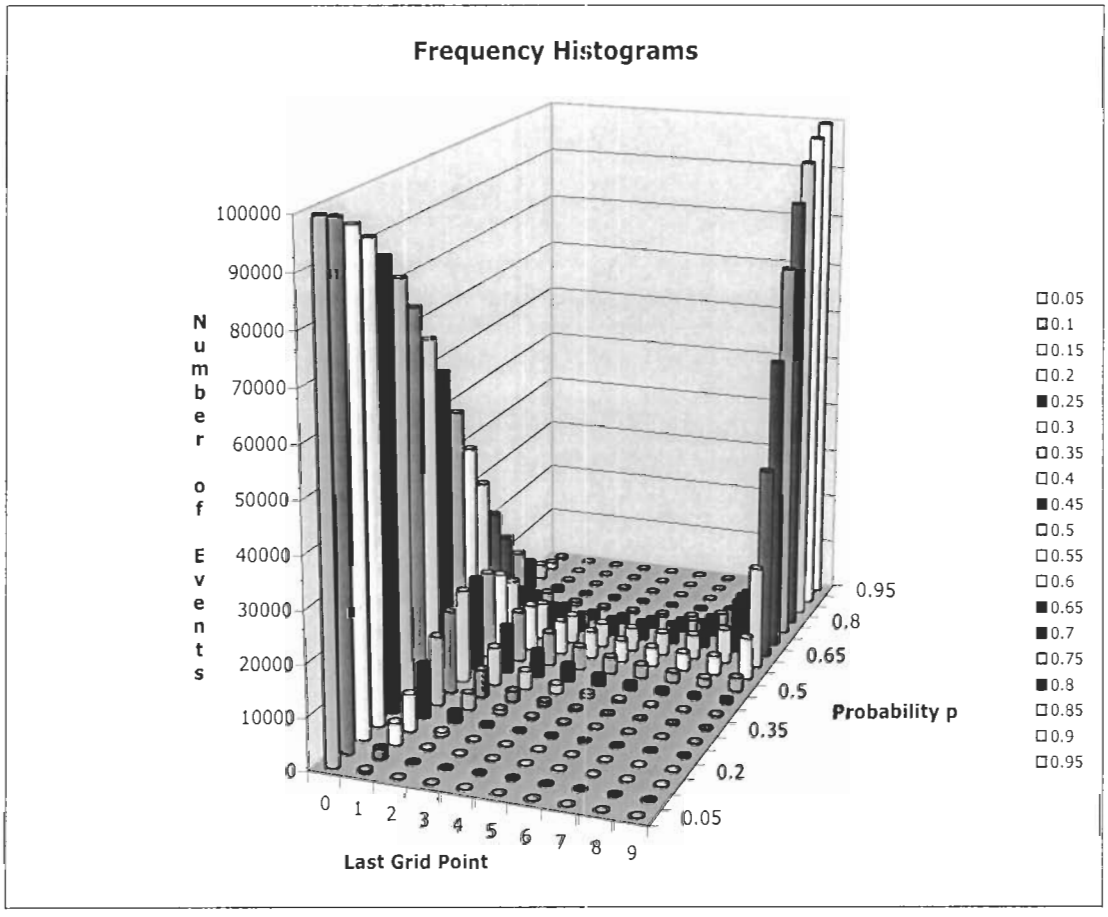
Again using Monte Carlo simulation based on one million trials, for values of p between 0.05 and 0.9, incrementing by 0.05 (Table 3.1) we obtained frequency histograms of the test statistic (last grid point along the cross-section) (Figure 3.4).

Table 3.1: Values obtained from Monte Carlo simulations of pre-invasive bronchial epithelial lesion growth based on a million trials.

p\Grid point	0	1	2	3	4	5	6	7	8	9
0.05	99484	506	10	0	0	0	0	0	0	0
0.1	98040	1887	69	3	1	0	0	0	0	0
0.15	95522	4200	258	20	0	0	0	0	0	0
0.2	91862	7192	816	107	23	0	0	0	0	0
0.25	87117	10394	1991	387	87	21	3	0	0	0
0.3	81696	13658	3293	943	297	80	22	7	4	0
0.35	74905	16306	5388	2077	831	290	119	48	23	13
0.4	67248	18089	7522	3544	1800	842	447	256	124	128
0.45	58975	18254	9308	5223	3141	1893	1185	796	564	661
0.5	49327	17425	10034	6650	4558	3267	2512	1941	1641	2645
0.55	40085	14931	9159	6672	5335	4406	3751	3536	3784	8341
0.6	30933	11103	7174	5471	4675	4498	4386	4812	6746	20202
0.65	22605	7125	4551	3474	3197	3360	3781	5110	8777	38020
0.7	15455	3979	2168	1613	1526	1735	2297	3847	9201	58179
0.75	9836	1904	766	468	449	562	942	2308	7540	75225
0.8	5712	734	197	91	69	107	266	887	5064	86873
0.85	2912	186	25	7	7	12	43	255	2847	93706
0.9	1226	27	2	0	0	0	3	47	1128	97567
0.95	300	3	0	0	0	0	0	3	280	99414

The last grid point along the cross-section was used as test statistic and the probability values of p were between 0.05 and 0.9, incrementing by 0.05. The histogram obtained from these simulations is found in Figure 3.4.

Figure 3.4: Frequency histograms representing last grid point along the cross-section of a million simulations with probability values *p* between 0.05 and 0.9, from Table 3.1.



Source: © Alma Barranco-Mendoza 2005

Using these results, we can estimate the value of *p* used to obtain a given sample cross section.

A natural method to use with this amount of information is the method of maximum likelihood. For example, from the list of frequencies, for the lattice showed on Figure 3.3 we found the maximum likelihood estimate of *p* to be 0.6 (that is, the probability function of the test statistic is maximized at p=0.6).

3.8 Modelling Lesions as Contact Processes

There are obvious limitations with the model proposed above, however, it does give us some insight in the behaviour of the process:

It can be viewed as a discrete time Markov chain. In this section we develop a continuous time model for this process. Markov chains in continuous time are defined by the rates

$$q(j,i) = p(i,j)Q$$

at which jumps occur from state i to state j, where Q is a constant representing the total jump rate and $p(i,j)$ the transition probability at each point of a Poisson process with rate Q.

The finite dimensional distributions of the process at state $s(t)$ at time

t is described by the probabilities $P(s_1(t) = i_1, \ldots s_n(t) = i_n)$, for each choice of a finite number of sites $s_1, \ldots, s_n$ and of possible states $i_1, \ldots, i_n$.

The total configuration at time t is described by giving the state of each site s_i. An initial distribution for the process that does not change in time is called a "stationary distribution" [Liggett. 1985].

If there is a stationary distribution that concentrates on configurations that have infinitely many sites in each possible state then we say that *coexistence* occurs. In most cases in which coexistence occurs there will be a translation invariant stationary distribution where $P(s(t) = i)$ is a constant $u[i] > 0$ that we will call the density of type i [Liggett. 1985].

Clustering occurs if for each *s* and *r* the probability of seeing one type of particle at *s* and a different type of particle at *r* converges to 0 as *t* tends to infinity. [Liggett. 1985]

The *contact process model* was first introduced by Harris in 1974 [Harris. 1974]. In this model, each site in the square lattice is either occupied (in state 1) or vacant (in state 0) and follows the conditions:

i. An occupied site becomes vacant at a rate δ; and

ii. a vacant site becomes occupied at a rate equal to the fraction of the four nearest neighbours that are occupied.

Much research has been done on these types of models ([Liggett. 1985], [Liggett. 1997], [Durrett. 1980], [Durrett and Griffeath. 1982] [Dickman *et al.* 2002], to name a few), but perhaps the most important result on contact process is the *Complete Convergence Theorem*:

When the contact process does not die out then it will converge to the stationary distribution that is the limit starting from all 1's. [Durrett. 1992]

An immediate consequence of this is that the only stationary distributions for the process are [Durrett. 1992]:

i. the limit starting from all 1's,

ii. the trivial stationary distribution, *E*, which assigns probability one to the all 0's configuration, and

iii. $\{(p \times i) + (1-p)\} \times E$

An interesting modification to this model was presented by Durrett and Levin in 1994 when they proposed that the behaviour of stochastic spatial models could be determined from the properties of the mean field ODE [Durrett and Levin. 1994].

Going back to our model, we wish to rephrase it in terms of a modification to the contact process model. Define the lattice to be at most 10 cells in the diagonal since this was assumed to be a characteristic of the lung tissue. Each site in the lattice is either occupied by an abnormal cell (in state 1) or vacant (in state 0) and follows the conditions: [Barranco-Mendoza *et al.* 1999]

i. a vacant site becomes occupied at a rate equal to λ times the fraction of the four nearest neighbours that are occupied, and

ii. an occupied site becomes vacant at a rate equal to δ times the fraction of the four nearest neighbours that are vacant,

where $\lambda, \delta \leq \lambda < 1$, is the rate at which abnormal cells split and δ, $0 \leq \delta < 1$, is the small probability of an abnormal cell being displaced from the site by a healthy cell. These constraints allow us to model the situation in which both normal and abnormal cells coexist (at least in the early stages). It is easy to see that if $\delta \geq 1$ the process would die out, i.e., there would be total recovery; and if $\lambda \geq 1$ then the abnormal cells would take over the entire tissue.

In practice, the vacant sites are not actually vacant but occupied by healthy cells that can be displaced by abnormal ones. However, since we are

currently only concerned about the growth of abnormal cells, considering the healthy sites vacant simplifies the problem considerably.

Condition (ii) is necessary since there is a very small probability that a healthy cell may displace an abnormal cell. When this occurs, since we are not allowing cells to drift in any direction on the plane, the new cell will be pushed out of the 2-D plane. Hence, for this simple 2-D model, this situation is resolved by setting $\delta = 0$, i.e., once a site is occupied by an abnormal cell, it will never become vacant. However, the condition $\delta > 0$ must be considered when introducing drift as well as in the 3-D model.

Let S = {finite subsets of Z^2}. If $A \in S$ is the initial set of abnormal cells, then, let $\xi_A(t)$ be the set of sites occupied by abnormal cells at time *t*.

We can rewrite the above as Markov processes $(\xi_A(t))_{t \geq 0}$ with jump rates given by: [Barranco-Mendoza *et al.* 1999]

Equation 3.4

$$\xi_A(t) \rightarrow \xi_{A \cup \{s\}}(t+1) \text{ (where } s \notin \xi_A(t)) \text{ at rate } \lambda \mid \{r \in \xi_A(t) : \| r - s \| = 1\} \mid,$$

$$\xi_A(t) \rightarrow \xi_{A - \{s\}}(t+1) \text{ (where } s \in \xi_A(t)) \text{ at rate } \delta \mid \{r \notin \xi_A(t) : \| r - s \| = 1\} \mid,$$

where ||s|| is the distance from *s* to 0, i.e., the rate at which a site becomes occupied by an abnormal cell is dependent on the cardinality of the set of sites occupied by abnormal cells adjacent to the current site.

Note that if $\delta = 0$ and $\lambda = 1$, then our model is a finite version of Richardson's growth model presented in [Richardson. 1973]. There Richardson showed that if $B(t)$ is the set of sites occupied at time t, then $B(t)/t$ clusters to a limiting shape, which is roughly but not exactly circular.

Since we assumed that the process started with a single abnormal cell at the origin, then we are only interested in the process $\xi_0(t)$.

3.9 Extensions

In [Williams and Bjerknes. 1972] Williams and Bjerknes presented a model of skin cancer, improved later by Bramson and Griffeath in [Bramson and Griffeath. 1980] and [Bramson and Griffeath. 1981], that follows a similar approach to ours. The main difference is that, by the nature of the problem, skin cancer growth was modelled only as sidewise splitting on the basal layer in such a way that the surface folded onto a torus. On the other hand, recall that our model is just a 2-D simplification of a model of very small pre-cancerous epithelial lesions that have not gone yet into metastasis. It is very constrained since we are modelling the three layers only as cross-sections of the bronchial epithelium, which forces the model to be restricted to a finite height along the diagonal. Based on this approach, we can represent our model using a partial differential equation (PDE) to describe the stochastic process: [Barranco-Mendoza *et al.* 1999]

Equation 3.5:

$$\frac{\partial n}{\partial t} = \sigma \Delta n + \lambda n + \mu_y \frac{\partial n}{\partial y}$$

where n represents the number of abnormal cells in the region of interest.

Summarizing the various parts of the equation, the term:

$$\sigma \Delta n = \frac{\partial^2 n}{\partial y^2}$$

represents the diffusion equation, and models the random movement of the abnormal cells in the tissue over time, where σ is the diffusion constant.

$$\lambda n$$

The birth rate of the abnormal cells, n, is controlled by the parameter λ. Clearly this term allows the number of abnormal cells spawned at any given time to grow linearly with the current number of abnormal cells.

Since there is a natural upwards drift of the cells in the tissue, we use the term:

$$\mu_y \frac{\partial n}{\partial y}$$

to model this phenomena. Notice that this term depends on the distribution of the cells in the vertical direction. Since we can expect more cells to drift upwards if there are more cells clustered near the basal layer of the tissue than elsewhere.

To model other aspects of the biological processes occurring in the diseased tissue, additional terms are required. For example, the term:

$$\mu_x \frac{\partial n}{\partial x}$$

can be added to model the lateral drift of the cells; the rate can be controlled through the parameter μ_x.

Finally, we need to set suitable boundary conditions. Setting the Dirichlet conditions:

$$n(0) = 1, n(10) = 0$$

at the bottom and top, respectively, of the cell layer. This makes sense physically since in this simple model we only allow one abnormal cell at the basal layer, and assume that once cells reach the top layer they die. The boundary conditions to model the sides of the region are more complicated. A possible solution would be to use moving boundary conditions at these edges, so that as the lesion expands the boundaries would also expand.

There is a natural extension to using a 3-D lattice for the discrete model. In the proposed partial differential equation model a complete analysis of the system needs to be done to compare the results with those obtained from the presented stochastic model. To analyse the 3-D model, we will require another parameter to handle movement of cells in this third axis. Hence, if μ_x and μ_z are the rates of lateral drift at which cells move along the *x*- and *z*-axis, respectively, then: [Barranco-Mendoza *et al.* 1999]

Equation 3.6:

$$\frac{\partial n}{\partial t} = \sigma \Delta n + \lambda n + \mu_y \frac{\partial n}{\partial y} + \mu_x \frac{\partial n}{\partial x} + \mu_z \frac{\partial n}{\partial z}$$

3.10 Conclusions and Limitations

The behaviour of cancer cells growth and development involves many different factors and modelling it requires a very complex system such as the one developed by Clem *et al.* in [Clem *et al.* 1997a]. However, this model is so complex that it cannot be used to do any mathematical analysis of the growth process nor be used for diagnostic purposes. To gain some insight to effectively understand the process, simplified models of the growth process are very useful tools. In this chapter, two models —one discrete stochastic model and one PDE model— were proposed to solve a 2-D simplification of the original problem posed by Clem and Giraud in [Barranco-Mendoza *et al.* 1997]. Using Monte Carlo simulations, we observed from the experimental results that at $p = .64$ there is a threshold effect. Any smaller values of p result in mainly small configurations, i.e., the abnormal cell division stops or dies out, whereas any larger p results in most configurations having some cells that reach the last level, which can be interpreted that the abnormal cells have a sufficiently aggressive growth that would warrant medical treatment. We also showed that by modelling the cell growth as Markov or contact processes we can expect that if the cell growth B(t) is the set of sites occupied at time t, then B(t)/t clusters to a limiting shape, which is roughly but not exactly circular. This behaviour is observed both in Clem's model —which is considerably more complex than ours and against

which data our results were compared to— and in real-life tumour growth hence confirming that our model exhibits behaviour comparable to that of the real-life system.

These simplified mathematically tractable systems can provide a lot of insight into the fundamental processes involved and can eventually be used as part of the development of a system that could aid in the early detection of lung cancer. However, the main limitation of this model is that it only addresses the growth of one lesion at a time. It is very likely that one patient may have several lesions present; one lesion not developing cancer does not preclude others from developing it. Also, in order to determine the rate of growth, each lesion would have to be observed at regular intervals to quantify its growth, which would require regular bronchoscopies —an intrusive and costly procedure— for the patient. In actual medical diagnosis, whenever a lesion is identified in a bronchoscopy, many times the entire lesion is removed. Thus, this defeats the usefulness of the model, as that particular lesion will not longer continue its growth.

As we explained in the previous chapter, there are many additional factors that impact whether a patient will develop cancer or not. Therefore, a model that only involves the structural analysis of a single lesion is not sufficient as an effective diagnostic tool. In the upcoming chapter we addressed this issue with the development of the PPBM, a multidisciplinary biological knowledge representation model, which establishes a framework to represent multiple types of patient information —such as cancer history, smoking history; imaging, serum,

sputum, and genetic data— and their relationships in terms of constraints. We later present how PPBM can be used in the development of a prototype system for early cancer diagnosis.

CHAPTER 4: PROBABILISTIC PROPERTY-BASED MODEL FOR MULTIDISCIPLINARY BIOLOGICAL KNOWLEDGE REPRESENTATION

4.1 Introduction

As mentioned in chapter 2, a lot of the CAD research is being done on the image analysis field [Gur *et al.* 2004], however, at the early development stages of the lesion, the information that can be obtained from lung imaging analysis (X-ray, CAT scans, MRI, PET scans, etc.), is quite limited, as we explained. Even though some have been able to detect lesions smaller than 1 mm [Nagamoto *et al.* 1993], [Brown *et al.* 2003] the early diagnostic information offered is not determinant as the structural differences between lesions that would become malignant and those that would remain benign or will not develop further, as explained above, are minimal. To completely understand the evolution of normal epithelium into invasive neoplasia would require the understanding of the genetic relationship of the cells in a pre-invasive neoplastic lesion during the development into invasive cancer. In recent years, biological research has been done in the area of cancer genetics that has shown that cancer results from an accumulation of key mutations in expanding clones originating from tissue-specific stem cells [Marx. 2003]. The recent availability of the human genome sequence, and the development of high throughput genomic technologies and methods for isolating selected cell populations have started to give us the

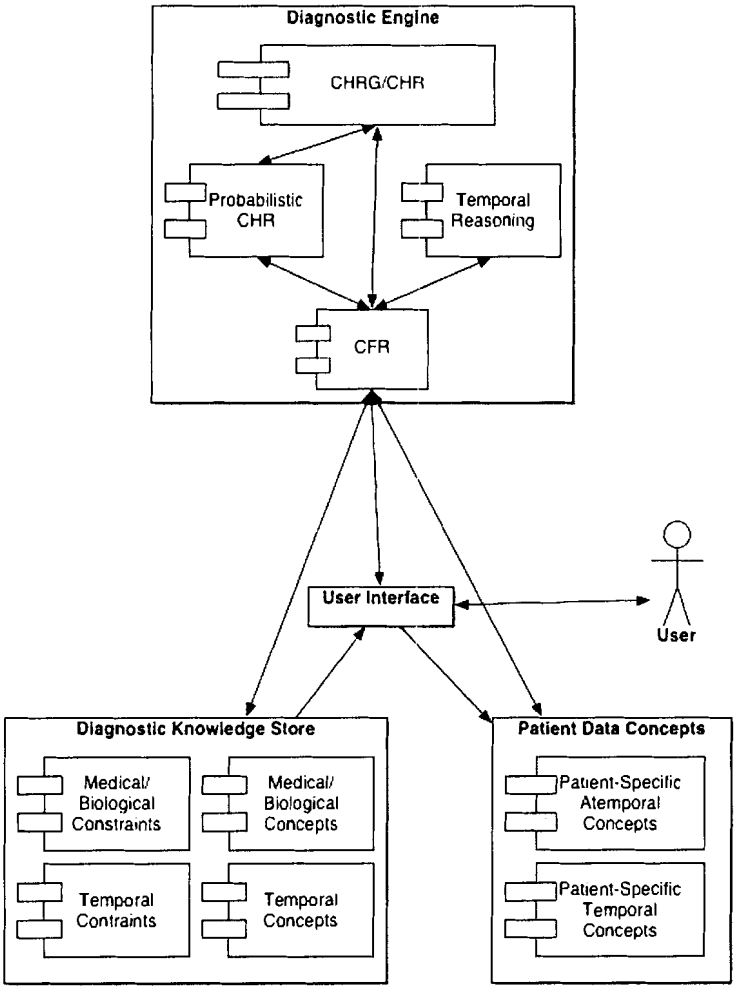
opportunities for understanding how human cancers develop. This information will drastically improve cancer diagnosis and treatment through the discovery of disease-specific molecular targets. As well, recent research has been focusing on the detection of biomarkers obtained from serum and sputum proteomic analysis. [Gealy *et al.* 1999]

4.1.1 Objective

Bringing together all of this multidisciplinary information about the development of the disease at the early stages along with the patient's medical history and behavioural risk analysis, could give the medical professionals a more accurate diagnostic of the probability of the lesions advancing to cancer or not. However, up to now, there has been little research in terms of computational methods to assist in the integration and analysis of all the genetic and molecular information along with the radiological, serum and sputum data. As well, for a system to have a true impact in the medical field it should be able to provide some kind of diagnosis even if given incomplete patient information, as not all tests can or will be done on said patient at a given time. Also, it has to take into account the cost —monetary (e.g. prohibitively expensive), emotional (e.g. too stressful or traumatic), and physical (e.g., too many side effects) — of a particular medical procedure or test, and weight it against the actual benefit it would provide in terms of improving the accuracy of the diagnosis. Our research focused on addressing this with the development of the multidisciplinary property-based model for multidisciplinary biological knowledge representation or PPBM. The high-level architecture of the PPBM (Figure 4.1) consists of 4 main

components: User Interface, Patient Data Concepts, Diagnostic Knowledge Store, and Diagnostic Engine. These components will be described in detail in the following sections.

Figure 4.1: PPBM High-Level Architecture Diagram



4.2 Methodology

The key difficulty in the development of a multidisciplinary system is the very different characteristics that each element of data involved in the analysis presents. Very soon we came to the realization that we needed to address a complex knowledge representation problem before being able to develop a diagnostic system that could integrate all of this data. Given the very different nature of each data element we determined that it was necessary to use a high level of abstraction to be able to capture a generalized way of representing all the elements consistently in order to be able to express both their independent characteristics and the relationships between them. Therefore, the next step we decided to follow was to look into constructivist theories and knowledge representation models based on those, which would allow the flexibility required to address concepts so distinct amongst themselves.

The main idea behind Constructivist Theory, introduced by Bruner in 1973, is that "learning is an active process in which learners construct new ideas or concepts based upon their current/past knowledge" [Kearsley. 2004a]. During this process, "the learner selects and transforms information, constructs hypotheses, and makes decisions, relying on a cognitive structure to do so" [Kearsley. 2004b]. The formation of new concepts from known ones is also central to cognitive logic, and also to logic programming and its recent sub-paradigm, Constraint Handling Rules (CHR). [Fruhwirth. 1998]. More recently, in [Dahl and Voll. 2004], the authors take advantage of these natural connections to develop a cognitive model of knowledge construction, Concept Formation Rules

(CFR), that can be directly executed through a specialized system implemented using CHRs. Based on this model, we can propose a biological knowledge construction model and expand on it by incorporating the use of concept probabilities which would enable us to quantify likelihood impact probability based on individual concepts and the constraints between them.

4.2.1 Concept Formation Rules

CFR is a directly executable new cognitive model of knowledge construction inspired in constructivist theory as well as in recent natural language processing methodologies.

> *In this model, problems, events, feelings etc. that are on focus (that is, in our consciousness at any given time) trigger a (partly or wholly) unconscious search in the knowledge store for those pieces of information that relate to the problem. Once found, these can be put together to draw new knowledge from them. A rule appropriate for modelling the formation of the new knowledge might look roughly as follows:*
>
> *c1, c2, ...ci → newc.*
>
> *where the ci's and newc are concepts expressed as logic atoms. We call these rules* concept formation rules. [Dahl and Voll. 2004]

This system accommodates user definition of properties between concepts as well as user commands to relax their enforcement; accepts concept formation from concepts that violate principles that have been declared as relaxable, and produces a list of satisfied properties and a list of violated properties as a side effect of the normal operation of the rules.

CFR bases its approach on a constraint-based formalism called Property Grammars (PG). [Bes and Blache. 1999], [Bes *et al.* 1999], [Blache. 2000]

PG's main objective is to represent in separate ways syntactic, semantic, pragmatic, and other information of different kinds, while being able to process them simultaneously if needed. That is, PG rules can combine parsing elements of the same level of representation together, or if needed, elements of different levels of representation (e.g., syntactic elements plus pragmatic elements) in the same rule. In [Dahl and Blache. 2004], the authors explain that "[i]n this approach, syntactic structure is not expressed in terms of hierarchy, but only by means of relations between categories. Such relations do not have any topological constraints; they can for example be crossed. Moreover, only relations between objects are used for describing a category. As a consequence, the notion of constituency is no longer relevant for the description process: a category is specified by a set of properties rather than by a set of constituents. In other words, the fact that several categories belong to a network of relations indicates that they characterize an upper-level category. A syntactic category is then described by a set of properties, which represent relations between other categories (lexical or syntactic)." The main goal of PG is to make explicit all the different relations that can exist.

The following types of information obtained by this approach can be summarized by the following properties: [Blache. 2000]

- *Constituency (const): Defines the maximal set of categories constituting a syntactic unit. This property allows the determination of the non lexical categories that will appear in the characterization of a given input.*

- *Obligation (oblig): Specifies the set of compulsory, unique categories. Such categories correspond to the heads.*

- *Unicity (unic): Set of categories which cannot be repeated in a phrase*
- *Requirement (req): sub-categorization, which indicates the co-occurrency relations between categories or sets of categories.*
- *Exclusion (exclude): The impossibility of co-occurrency between categories.*
- *Linearity (prec): Linear precedence constraints.*
- *Dependency (dep): Dependency relations between categories.*

We have identified that these are the type of properties required to express the relations between the biological data involved in the diagnostic process. Here are some biological examples obtained from some lung cancer tumour marker characteristics [Wu and Nakamura. 1997]:

- *Constituency*: Allows the determination of the categories that will appear in the characterization of a patient's personal data:

  `const(Personal_data)={Patient_ID,Age,Race,Gender}`

  The corresponding categories can be dominated in the hierarchy by *Personal_data*.

- Unicity: A patient's personal data cannot have repeated categories:

  `unic(Personal_data)={Patient_ID,Age,Race,Gender}`

- *Linearity*: In our medical domain, linearity is used most frequently to represent temporal precedence. E.g., the change in size and shape of tumour as detected by chest X-ray or CT scan can be represented this way. The tumour must progress to a certain size before detection by the

methods stated above. In other words, the tumour will develop from a minute size to a size large enough to be detected:

$$prec(Tumour\_size\_past, Tumor\_size\_current)$$

which can then be used to determine if there has been tumour growth.

Another example for linearity is the increase in concentration of a particular marker with time:

$$prec(Marker\_concentration\_past, Marker\_concentration\_current)$$

which can be used to determine if there is an increase in concentration values from a session with the physician to the next session are usually indicative of tumour progression.

- *Requirement*: E.g., the presence of the following serum marker (neuron specific enolase, NES) at values greater than 22 ng/mL. Adults have a normal range of 1-22 ng/mL. Children have a normal range value of 1-12 ng/mL. In cases with SCLC, 60-80% of patients have a mean value of 37 ng/mL:

$$req(NES\_normal\_range \leq 22) = \{Age > 12\}$$

$$req(NES\_normal\_range \leq 12) = \{Age \leq 12\}$$

- *Exclusion*: E.g., prostate specific antigen (PSA) is tissue specific i.e., it is the major protein found in seminal plasma. The high concentration of PSA-ACT in serum (4-10 ng/mL) would indicate prostate cancer and not that of lung cancer:

exclude(lung_cancer)={Serum_marker=PSA-ACT,Current_marker_range>4}

- *Obligation*: E.g., the major marker for SCLC is neuron specific enolase and can occur in the blood stream when the SCLC is lysed. Therefore lysis of the cells must occur:

  oblig(lysis) = {Serum_marker=NES}

- *Dependency*: E.g., the detection of NES in the blood stream is usually an indication of malignant disease state of SCLC i.e., metastasis. The prognosis of SCLC is enhanced with detection of thymidine kinase:

  dep(Test_thymidine_kinase,Serum_marker=NES)

Therefore, we used CFR as the basis to implement our model for cancer diagnosis. [Barranco-Mendoza *et al.* 2004]

4.2.1.1 Advantages of CFR

Given the multidisciplinary nature of our data, CFR offers a sufficiently high level of abstraction to represent this biological knowledge in terms of concepts. Each data element, or concept, will be conformed of its own unique properties and the properties between it and other concepts. As rules become unified, new concepts are added to the knowledge store. As most of the rules in the knowledge store will not be used at a given point in the analysis, the rule selection process is important. As explained in [Dahl and Blache. 2004], CFR "solve[s] this problem by using a system that constantly looks for information that matches any new information arriving (i.e., an initial problem's concepts, or the

concepts formed while trying to solve it), and triggers those rules whose left hand side matches at least some of the new information."

Another advantage to CFR is the flexibility it gives to relax constraints, allowing the completion of the analysis even with incomplete information. This addresses the issue raised in 4.1.1 of proving diagnosis even with incomplete patient information. The list of violated properties serves interesting purposes, such as providing a list of suggested follow-up tests to improve the accuracy in the diagnosis. The way this is accomplished is explained in [Dahl and Blache. 2004]:

> *[R]ather than inflexibly allowing for a concept to be formed if a test succeeds and disallowing its formation if that test fails, we single out those tests for which we want to allow flexibility as properties. Properties are like any other test, except that their failure does not result in the rule itself necessarily failing: the concept will still be formed, and two lists will be associated with it: a list of the properties that the concept satisfies (S) and a list of those which it violates (V).*
>
> *This allows us to construct possibly incorrect or incomplete concepts, plus the information regarding in which way they are not totally warranted. The user then has all the information pertaining to the construction of a particular concept and can interpret these results in a much more informed, holistic way than if the degree of randomness or vagueness had been blindly computed from those assigned a priori to each individual piece of a reasoning puzzle.*

CFR transforms its knowledge store as a side effect of applying a concept formation rule. CFR can address the three types of knowledge store transformation: (a) new concept coexisting with the concepts that participated in its formation; (b) removal of redundant concepts; and (c) new concept replacing old concepts that led to it.

Another interesting feature is the possibility of making hypotheses in the form of *assumptions*. Assumptions are atoms syntactically marked as such that are available for consumption during the computation. These assumptions give the opportunity of developing "what if" scenarios, a useful feature when dealing with possible diagnostic scenarios, while still having certain unknowns. These assumptions are automatically backtracked upon if they lead to impasses or inconsistencies.

4.2.1.2 Limitations of CFR

CFR, even though it can be used to determine constraints between concepts, it does not deal with probabilistic analysis. Being able to provide the likelihood that the patient will or will not develop cancer as part of our diagnosis is the main objective, therefore CFR will be expanded to handle this probabilistic analysis.

Another issue to address is that a medical condition such as cancer does not remain static throughout time. In fact, diagnostic and treatment are highly dependent on the changes of the patient's condition throughout time. Hence, a module to handle temporal reasoning must be included.

4.2.2 Temporal Reasoning in Diagnostic Systems

Most clinical tasks require measurement and capture of numerous patient data. Physicians who have to make diagnostic or therapeutic decisions based on these data may be overwhelmed by the number of data if the physicians' ability

to analyse and reason about the data does not scale up to the data-storage capabilities. Most stored data include a time stamp in which the particular datum was valid; an emerging pattern over a stretch of time has much more significance than an isolated finding or even a set of findings. Experienced physicians are able to combine several significant contemporaneous findings, to abstract such findings into clinically meaningful higher-level concepts in a context-sensitive manner, and to detect significant trends in both low-level data and abstract concepts. Thus, it is desirable to provide short, informative, context-sensitive summaries of time-oriented clinical data stored, and to be able to answer queries about abstract concepts that summarize the data. Providing these abilities would benefit both a human physician and an automated decision-support tool that recommends diagnostic measures based on the patient's clinical history up to the present.

Such concise, meaningful summaries, apart from their immediate value to a physician, could, in the future, support the automated system's further recommendations for diagnostic or therapeutic interventions, provide a justification for the system's or for the human user's actions, and monitor plans suggested by the physician or by the decision-support system. To achieve this, a meaningful summary cannot use only time points, such as dates when data were collected; it must be able to characterize significant features over periods of time, such as "2 months of increasing lesion size."

In [Combi and Shahar. 1997], the authors explain that:

> *Temporal reasoning has been used in medical domains as part of a wide variety of generic tasks [...], such as diagnosis (or, in general, abstraction and interpretation), monitoring, projection, forecasting, and planning. These tasks are often interdependent.* ***Projection*** *is the task of computing the likely consequences of a set of conditions or actions, usually given as a set of cause–effect relations. Projection is particularly relevant to the* planning task *[...].* ***Forecasting*** *involves predicting particular future values for various parameters given a vector of time-stamped past and present measured values, [...] given the values up to and including the present.* ***Planning*** *consists of producing a sequence of actions for a care provider, given an initial state of the patient and a goal state, or set of states, such that that sequence achieves one of the goal patient states. Possible actions are usually operators with predefined certain or probabilistic effects on the environment. The actions might require a set of enabling* preconditions *to be possible or effective. Achieving the goal state, as well as achieving some of the preconditions, might depend on correct* projection *of the actions up to a point, to determine whether preconditions hold when required.* ***Interpretation*** *involves abstraction of a set of time-oriented patient data, either to an intermediate level of meaningful temporal patterns, as is common in the* ***temporal-abstraction*** *task or in the* ***monitoring*** *task, or to the level of a definite diagnosis or set of diagnoses that explain a set of findings and symptoms, as is common in the* ***diagnosis*** *task. Interpretation, unlike forecasting and projection, involves reasoning about only past and present data and not about the future.*

When modelling time whether for reasoning about or management of time-oriented clinical data there are several decisions that need to be taken, based on the needs of the domain of our model. These decisions involve determining how time will be modelled to represent time points (instants, e.g. x-ray on May 12, 2003) and time intervals (e.g. length of chemotherapy treatment); relative time (makes reference to its context, e.g. experiences coughing after smoking) and absolute time (an absolute position such as a calendaric time, e.g. biopsy on March 13, 2005); linear time (set of time points is completely ordered; e.g., most clinically-oriented databases); branching time (can happen in diagnosis,

projection or forecasting) and circular time (when recurrent events happen, e.g. chemotherapy treatment every week.) [Cukierman and Delgrande. 2000],

Allen's interval algebra [Allen. 1984], [Allen. 1983] has been widely used in medical informatics to model temporal relationships. [Combi *et al.* 1995] proposes some extensions to Allen's basic thirteen interval relationships [Allen. 1983] (direct relations and their inverse: before, after; meets, met by; starts, started by; during, contains; overlaps, overlapped by; finishes, finished by; equals).

There are two main types of temporal relationships: qualitative (e.g. bleeding after coughing) and quantitative (e.g. bleeding 5 minutes after coughing). [Combi *et al.* 1995]

There are two main approaches to modelling temporal entities in medical informatics: (a) adding a temporal dimension to an existing object, and (b) creating task-specific, time-oriented entities.

Adding a temporal dimension to existing an existing object is an approach that originated from database research. The main idea is to add a temporal dimension at the tuple or object level or at the attribute or method level. In [Das and Musen. 1994] the main focus is to model time-oriented clinical data to permit the DBMS to store and manage this type of data. In [Combi *et al.* 1995] the concept of "temporal assertion" was introduced. The main idea is to model both instant- and interval-based information in a homogeneous way.

The approach of creating task-specific, time-oriented entities originated mainly from artificial intelligence in medicine. Modelling different temporal features of complex, task-specific entities is the main idea of this approach. The needs of the relevant temporal-abstraction and, in general, temporal-reasoning tasks define the temporal entities.

In [Kahn *et al.* 1991a] the concept of a Temporal Network (TNET) was formally introduced and [Kahn *et al.* 1991b] later extended it by the Extended TNET, or ETNET model. Here, a T-node (or an ET-node) models task-specific temporal data at different levels of abstraction. I[Keravnou and Washbrook. 1990] introduces findings, features, and events to distinguish various types of instantaneous and interval-based information, which is patient-specific or general.

4.2.2.1 PPBM Temporal Requirements

Our specific model will require addressing temporal issues mostly dealing with change of symptoms/conditions from session to session for a specific patient. At the moment, the input does not involve analysis of medical records in natural language hence we only have to deal with absolute and not relative time. Our data will be represented in terms of CFR concepts, which can be viewed as database tuples. Therefore, we will follow the approach to modelling temporal entities by adding a temporal dimension to our existing data. To do this, we will use time stamps at the tuple level. We will also generate constraints that would incorporate Allen's interval algebra [Allen. 1984], [Allen. 1983] in our analysis. Our model's timeline, as it stands at the moment, is mostly linear and it is as

such that we make that assumption in the initial formalization of PPBM. [Barranco-Mendoza *et al.* 2005] Yet, we will be taking into consideration the possibility of branching and circular time in future research.

As we expressed above, there are a vast number of temporal issues that should be addressed but were beyond the scope of the initial formalization of the PPBM, here presented. However, these issues will be addressed by the author and incorporated into the model in future research.

4.2.3 Probabilistic Analysis of Medical Data

When analysing medical data for purposes of diagnosis we require a way to quantify the impact the different symptoms and patient characteristics will have on the diagnostic of a particular disease. Considering each symptom to have the same weight in the occurrence of a disease is unrealistic as, for example, age may not be as determinant in a patient developing lung cancer as having a history of heavy smoking is.

Typically, when building a CAD system, knowledge engineers rely on a variety of sources that include expert knowledge (interviews with pathologists, etc.), literature, available statistics, and databases of relevant cases. Very often, the structure of the model is elicited from experts and the numerical probabilities are learned either from available databases or literature results. A common approach to quantifying the data is done by using score systems due to their simple applicability [Ohmann *et al.* 1995],[Franke *et al.* 1999]. Unfortunately, they are not very effective when we need to take into account interdependencies

among the values, which are input to them when trying to decide an actual application case; a drawback that is overcome by the more powerful probabilistic systems. In the following sections we will explain how score systems work and discuss how a probabilistic approach to them can be followed. Then we will present how we will implement this approach in PPBM using Probabilistic CHR [Frühwirth *et al.* 2001]. We will also discuss some important aspects to consider when engineering a knowledge base composed by data from multiple sources.

4.2.3.1 Basic Considerations of Diagnostic Systems

In [Schramm and Fronhofer. 2003], the authors identified some general elements that must be considered in the modelling of a diagnostic system. First of the relationships between diagnoses (diseases) and the attributes (symptoms) must be determined and represented explicitly.

> *This means that we have a finite set of* ***variables*** *—symptom variables and, in addition, diagnosis variables— with each having a finite set of values. The* ***symptom variables*** *describe properties/symptoms/attributes relevant for the diagnosis task. [...] The values of the* ***diagnostic variable*** *define a classification of the possible diagnostic results [...] based on the values of the symptom variables. [...] This relation can be specified by a* ***method of judgement****. Depending on this method we may either get a 'classical' relation or a somehow fuzzy one, i.e., we may have yes-no judgements on tuples [...] or a more fine-grained judgement, e.g. scores or probability measures. Ibid.*

We can denote the symptom variables —or symptom concepts in the context of PPBM— by $S_i\ (1 \le i \le m)$ each of them associated with the set of their values $S_i = \{s_{ij} \mid 1 \le j \le k_i\}$.

In a similar way, we denote the diagnosis variable —or diagnosis concept— as $D = \{d_1,...,d_{kD}\}$.

Hence, a diagnostic system should model a relation on the tuple space $\Omega = S_1 \times \cdots \times S_m \times D$, based on the above considerations. *Ibid.*

For the purpose of our system, we would require a fine-grained method of judgment on the tuples $\langle s_{1j_1}, s_{2j_2},\ldots, s_{mj_m}, d_h \rangle$ based on probability measures.

However, given the nature of our domain, there may be times that a certain symptom may either not occur or may not be possible to observed. Therefore, our method of judgement should not be restricted to judging only individual tuples but we are interested in judgements on arbitrary subsets of Ω.

Using probability theory terminology, we define Ω to be an **event space** with its power set as **set of events** or **event algebra**.

So, if one considers the subset $\breve{S}_i \subset S_i$ (symptom concept) or the subset $\breve{D} \subset D$ (disease concept) one can denote $\langle \breve{S}_i \rangle := S_1 \times \cdots \times S_{i-1} \times \breve{S}_i \times S_{i+1} \times \cdots \times S_m \times D$ and $\langle \breve{D} \rangle = S_1 \times \cdots \times S_m \times \breve{D}$. So, $\langle \breve{S}_{i_1},\ldots,\breve{S}_{i_n} \rangle$ with $\breve{S}_{i_j} \in S_{i_j}$, which is the intersection of the sets $\langle \breve{S}_{i_j} \rangle$.

Therefore, the expression $\langle s_1,\ldots,s_m,d_h \rangle$, where $s_i \in S_i$ and $d_h \in D$ and which can also be written as $\langle \vec{s}, d_h \rangle$, is called an **elementary symptom event** in Ω, (we call $\langle s_i \rangle$ a **simple event**) and all the general events are sets of elementary

events. [Schramm and Fronhofer. 2003] Following Schramm's notation, we will represent a **conditional event** E'|E as E⇨E'.

To determine the conditional judgement for our CAD system, we can assume that $\vec{s}$ is an event that represent the set of symptoms' values of a particular patient and we want to determine if this patient suffers from disease $d_h \in D$ (d_h = lung or oral cancer, in our case). Therefore, we want to determine the judgement of the event $\langle \vec{s}, d_h \rangle$ compared to the event $\langle \vec{s}, d_g \rangle$, which is other diseases (or, in our case $d_g = \overline{d}_h$, i.e., absence of lung or oral cancer) in view of the same symptoms' values $\vec{s}$. Hence, we are interested in the judgement of the conditional event $\vec{s} \Rightarrow \langle \vec{s}, d_h \rangle$ (or $\vec{s} \Rightarrow d_h$ for simplicity) in comparison to the judgement $\vec{s} \Rightarrow \langle \vec{s}, \overline{d}_h \rangle$ (or $\vec{s} \Rightarrow \overline{d}_h$). In medical terms this could be read as: "if I know that a patient is showing the symptom values $\vec{s}$, what can I say —in view of this knowledge— about his risk of having the illness d_h?" [Schramm and Fronhofer. 2003]

4.2.3.2 Score Systems

Score Systems are based on a set of attributes, which each one of them has a value or set of values associated. When a score system is applied to a concrete case, the scores that correspond to the attribute values in case are added up. Once this is done, if the sum obtained falls in a particular *score interval*, then, the diagnostic decision associated with that particular interval is proposed.

For example in [Ohmann *et al.* 1995] the authors explain their score system for diagnosing appendicitis:

> *For instance, in the medical domain, there may be the symptom/attribute 'body temperature' with (discrete) values 'low', 'normal', 'high' and 'very high'. To each attribute value a numerical value —its **weight** or **score**— is assigned (see Table [4.1])*
>
> *Thus, for instance, a proposal for a medical treatment is established on the basis of symptoms found with a patient and which are represented by a list of attribute values.*

Table 4.1: "Ohmann Score" for the diagnosis of appendicitis

In case of negative answers the scores are zero. Patients are diagnosed as having appendicitis if score sum ≥ 12, they are interned in case of 6 – 12, and are sent home in case of ≤ 6. (RLQ: right lower quadrant of abdomen (as seen from the patient).)

Symptom/attribute	Score, if yes
Tenderness in RLQ	4.5
Rebound tenderness	2.5
No micturition	2.0
Continuous type of pain	2.0
Number of leucocytes ≥ 10000	1.5
Age > 50 years	1.5
Relocation of pain to RLQ	1.0
Rigidity	1.0

Source: based on Table 1 in [Ohmann *et al.* 1995]

Score systems can formally be defined as follows:

> *It consists of a set of **variables** (**attributes**)* $S_i (i = 1, \ldots, m)$. *Each* S_i *can be identified with its set of variable values* $\{s_{i1}, \ldots, s_{iki}\}$ $(k_i > 1)$. *We denote by* $\vec{s}$ *a tuple of values* $\{s_1, \ldots, s_m\}$ *with* $s_i \in S_i$.
>
> *Moreover, for each variable* S_i *exists a set* $W_i = \{w_{i1}, \ldots, w_{iki}\}$ *of nonnegative **weights** or **scores** and a bijective **score function*** w_i: $S_i \rightarrow W_i$. *We also have a **(global) score function** we defined as* $w(\vec{s}) = \sum_{i=1}^{n} w_i(s_i)$.

Finally, there are ***score intervals*** *given by a set of* ***border values*** $b_1 < ... < b_{kT}$, *a decision variable T with values* $\{t_1,...,t_{kT}\}$ *and a decision function t which maps a sum of scores* $w(\vec{s})$ *to* t_i *iff* $b_{i-1} < w(\vec{s}) \leq b_i$ *(with* $b_0 := -\infty$*).* [Schramm and Fronhofer. 2003]

In our case, a simple score system is not the best option as our data requires establishing constraints that involve more than one concept. However, in the section below we present a probabilistic approach to score systems that can meet our constraint requirements.

4.2.3.3 Probabilistic Approach

As mentioned before, Score Systems have the limitation that they are not very effective when we need to take into account interdependencies among the different attributes and their values. Probabilistic Diagnostic Systems (PDS) [Devore. 1991] help us overcome that limitation.

PDS use a function, *P*, —which follows the laws of probability to assign probabilities to the events or symptoms— as the method of judgement.

Let us define our event algebra by mapping every elementary event —a finite set in our case— to a nonnegative real number such that the sum of functions values of *P* over all elementary events is equal to 1. This is called a *P-measure.* [Devore. 1991]

Let us define, for every event *E*, where *E* is the unique union of elementary events $e_1,...,e_n$:

$$P(E) = \sum_{i=1}^{n} P(e_i)$$

A set *E* together with a P-measure *P* is called a *P-space*. [Devore. 1991]

In [Schramm and Fronhofer. 2003] the authors showed how to transform score systems into slightly larger, yet simple, probabilistic systems derived from them.

The main idea behind their approach is to derive a set of constraints to be satisfied by a respective P-measure.

They accomplished this by:

i. Extending the symptom space $\Sigma = S_1 \times \cdots \times S_m$ to $\Omega = S_1 \times \cdots \times S_m \times D$, where $D = \{d, \bar{d}\}$. [1]

ii. Defining a set of P-measure on Ω based on the symptom values $\vec{s}$ as a judgement of $\vec{s} \Rightarrow d$, which leads to the constraint

$$^{c}P(\vec{s} \Rightarrow d) = \frac{w(\vec{s})}{\hat{w}_{\max}} \quad [2]$$

for all $\vec{s} \in \Sigma$, where $\hat{w}_{\max} = \max(w(\vec{s}) \mid \vec{s} \in \Sigma\}$ and

$$^{c}P(\vec{s}) > 0, \forall \vec{s} \in \Sigma. \quad [3]$$

iii. For arbitrary events $\vec{s}, \vec{s}' \in \Sigma$, $w(\vec{s}) > w(\vec{s}') \Leftrightarrow {}^{c}P(\vec{s} \Rightarrow d) > {}^{c}P(\vec{s}' \Rightarrow d)$. [4]

In their paper, Schramm and Fronhöfer showed that this P-measure is consistent. It also showed can handle incomplete symptom sets proving the theorem below in [Schramm and Fronhofer. 2003]:

[A]ssume for each $S_i = \{s_{i1}, \ldots, s_{ik_i}\}$ *a positive normalized weight function* y_i, *i.e.,* $y_i(s_{ij}) > 0$ *and* $\sum_{j=1}^{k_i} y_i(s_{ij}) = 1$. *Next [...] extend* w_i *to a*

function on the power set S_i *by defining for all subsets* $\breve{S}_i = \{s_{ij_1},\ldots,s_{ij_n}\} \subset S_i (n \leq k_i)$

$$w_i(\breve{S}_i) = \left(\sum_{p=1}^{n} y_i(s_{ij_p}) \bullet w_i(s_{ij_p})\right) / \sum_{p=1}^{n} y_i(s_{ij_p})$$

[D]efine for a subset $I = \{i_1,\ldots,i_n\} \subset \{1,\ldots,m\}$ *and for its complement* $\bar{I} = \{1,\ldots,m\} \setminus I$ *the* ***partial symptom event*** $\vec{p}_I = \langle \breve{S}_{i_1},\ldots,\breve{S}_{i_n} \rangle$ *(with* $\breve{S}_{i_j} \subset S_{i_j}$*).*

[E]xtend [the] global score function to $w(\vec{p}_i) = \sum_{i \in I} w_i(\breve{S}_i) + \sum_{i \in \bar{I}} w_i(S_i)$ *and extend [the] translation by the following additional constraint*

$$^{c}P(\vec{p}_i \Rightarrow d) = \frac{\sum_{j \in I} w_j(\breve{S}_j) + \sum_{j \in \bar{I}} w_j(S_j)}{\hat{w}_{\max}} . \qquad [5]$$

For a $s_{ij} \in S_i$*, [...] denote by* $\bar{s}_{ij}$ *the set of all values of* S_i *besides* s_{ij}*, i.e.,* $S_i \setminus \{s_{ij}\}$*. [D]efine* $a_{ij} = w_i(s_{ij}) - w_i(S_i)$ *and* $b_{ij} = w_i(\bar{s}_{ij}) - w_i(S_i)$*.*

Theorem: *Given a P-measure P which satisfies the constraints [2],[3], and [5]. With the weight functions* y_i *taken as the marginal distributions of the* S_i *derived from P, then holds that the* S_i *are marginal independent if all* $a_{ij} > 0$*.*

In the medical domain, the above theorem corresponds to the scenario where the value of a particular symptom is neither known nor completely unknown but it could be confined to a subset of the particular symptom.

4.2.3.4 Representing Probabilities in PPBM

As mentioned above, PPBM is based on CFR, which is implemented using CHRs. Probabilistic CHRs (PCHR) [Frühwirth *et al.* 2002a], [Frühwirth *et al.* 2001] is an extension of CHRs. The main objective of PCHRs is to allow the probabilistic weighting of rules, that is, specifying the probability that a particular rule is applied. As defined by the authors, "PCHR is characterised by a

probabilistic rule choice: Among the rules that are applicable, the committed choice of the rule is performed randomly by taking into account the relative probability associated with each rule." [Frühwirth *et al.* 2002a]

PCHR bases its implementation on source-to-source transformation (STS) [Frühwirth *et al.* 2002b] in which users write STS programs and, during compilation, these STS programs manipulate other programs. CHR rules get translated into relational normal form. This is done by introducing special CHR constraints for the components of a rule. These components are head, guard, body and compiler pragmas. These pragmas are the components that contain the probability or weight of each particular rule.

PCHR's syntax and operational semantics are described as follows: [Frühwirth *et al.* 2002a]

> *Syntactically, PCHR rules are the same as CHR rules but for the addition of a weighting representing the relative probability of each rule:*
> ***Definition [4.1]***
> *A probabilistic simplification CHR is of the form* $H \Leftrightarrow_p G \mid B$ *and a probabilistic propagation CHR is of the form* $H \Rightarrow_p G \mid B$ *where* p *is a nonnegative number.*
> *[...]*

Definition [4.2]
The transition relation $\mapsto_{\tilde{p}}$ *for PCHR is indexed by the normalised probability* $\tilde{p}$ *and is defined as follows:*
Simplify
$H' \wedge D \mapsto_{\tilde{p}_i} (H = H') \wedge G \wedge B \wedge D$
if $(H \Leftrightarrow_{p_i} G \mid B)$ *in* P *and* $CT \models \forall(D \rightarrow \exists\bar{x}(H = H' \wedge G))$.

Propagate
$H' \wedge D \mapsto_{\tilde{p}_i} (H = H') \wedge G \wedge B \wedge H' \wedge D$
if $(H \Rightarrow_{p_i} G \mid B)$ *in* P *and* $CT \models \forall(D \rightarrow \exists\bar{x}(H = H' \wedge G))$
where

$$\tilde{p}_i = \begin{cases} \dfrac{p_i}{\sum_{r_j} p_j} & \text{if } \sum_{r_j} p_j > 0 \\ \dfrac{1}{n} & \text{otherwise} \end{cases}$$

where the sum $\sum_{r_j} p_j$ *is over the probabilities of all rules* r_j *which are applicable to the current constraint in the current state and the number of applicable rules is* n.

4.2.3.5 Knowledge Engineering Using Data from Different Sources

It is critical to take close consideration of the process of the development of the knowledge base when discussing CAD systems, as it is the data and its representation what drives the entire diagnostic process.

In particular, when building probabilistic models, the most intimidating task in the knowledge engineering process is obtaining the numerical parameters. Many authors combine various sources of information such as textbooks, databases, statistical reports, expert advice, etc. to accomplish this task. In [Druzdzel and Diez. 2003], Druzdzel and Díez show how "the criteria 'do not combine knowledge from different sources' or 'use only data from the setting in which the model will be used' are neither necessary nor sufficient to guarantee

the correctness of the model." *Ibid.* In their paper, they offered a method for determining when knowledge form different sources can be combined safely into the general population causal model, as well as explaining how to use available subpopulation data to "build a model specific to a certain subpopulation characterized by a known variable X, assuming the selection value $X = x_s$." *ibid.*

The initial step in their approach is to design a causal graph based on the literature and the knowledge elicited from the expert, with nodes representing the disease and its symptoms as well as any other variable that impacts the model, such as population biases, e.g., data obtained only from a hospital setting; and the directed vertices representing the causal links. This graph will be used as a guide for determining how to combine data from different sources.

> ***Theorem [4.1]*** *Given a selection variable X_s in a Bayesian network and a node X_i (other than X_s), such that X_i is not and ancestor of X_s, the conditional probability distribution of X_i given $pa(X_s)$ is the same in the general population and in the subpopulation induced by value x_s, i.e., $Pr(x_i|pa(x_i), x_s) = Pr(x_i|pa(x_i))$.* [Druzdzel and Diez. 2003]

This theorem, proved in the paper, allows us to identify, from the conditional probabilities in a certain subpopulation, which have remained unaltered by the selection process and, hence, are unbiased. This allows us to introduce such parameters into the general-population model.

To build a subpopulation model, the corresponding causal graph must meet the following criteria:

> ***Definition [4.3]*** *A graph is linearly ordered for X_s iff*
> $\forall X_i, X_i \in \{X_s\} \cup anc(X_s), \exists X_j, X_j \in pa(X_i), \exists X_k, X_k \in pa(X_i)$
> $\Rightarrow (X_j = X_k) \vee (X_j \in pa(X_k)) \vee (X_k \in pa(X_j))$

This property can be phrased as follows: if X_s or and ancestor of X_s (say X_i) has two parents (X_j and X_k), then one of the two must be a parent of the other. Obviously,if each ancestor of X_s has only one parent, then the graph is linearly ordered for X_s.

Definition [4.4] *A causal Bayesian network is linearly ordered for X_s if its graph is linearly ordered for X_s.*

Theorem [4.2] *Given a Bayesian network is linearly ordered for X_s, for each configuration x_R of the variables $X_R = X\backslash\{X_s\}$, it holds that* $\Pr(x_R \mid x_s) = \prod_{i \neq s} \Pr(x_i \mid pa(x_i), x_s)$. [Druzdzel and Diez. 2003]

This method can only be applied to linearly ordered graphs. However, it is always possible to make a graph linearly ordered for X_s by following the algorithm proposed in [Druzdzel and Diez. 2003]:

1. *make an ordered list of **X** such that* $\forall i,\ pa(X_i) \subseteq \{X_1, \ldots, X_i\}$
   [which can be read as parents of X_i must be numbered before X_i.]
2. $A \leftarrow X_s$
3. *while A has parents*
   a. *B ← last node in pa(A), according to the list created in step 1,*
   b. *For each node C in pa(A)\{B},*
      If link C → A is not in the graph, add it,
   c. *A ← B*
   end while.

Therefore, any implementation of the knowledge base for PPBM, must take into consideration the above requirements to prevent the introduction of source biases into the parameters of the data.

4.3 Results

As proof of concept, we implemented a prototype of the PPBM. As mentioned before, the high-level architecture of the PPBM (Figure 4.1) consists of 4 main components:

- User Interface
- Patient Data Concepts
- Diagnostic Knowledge Store
- Diagnostic Engine

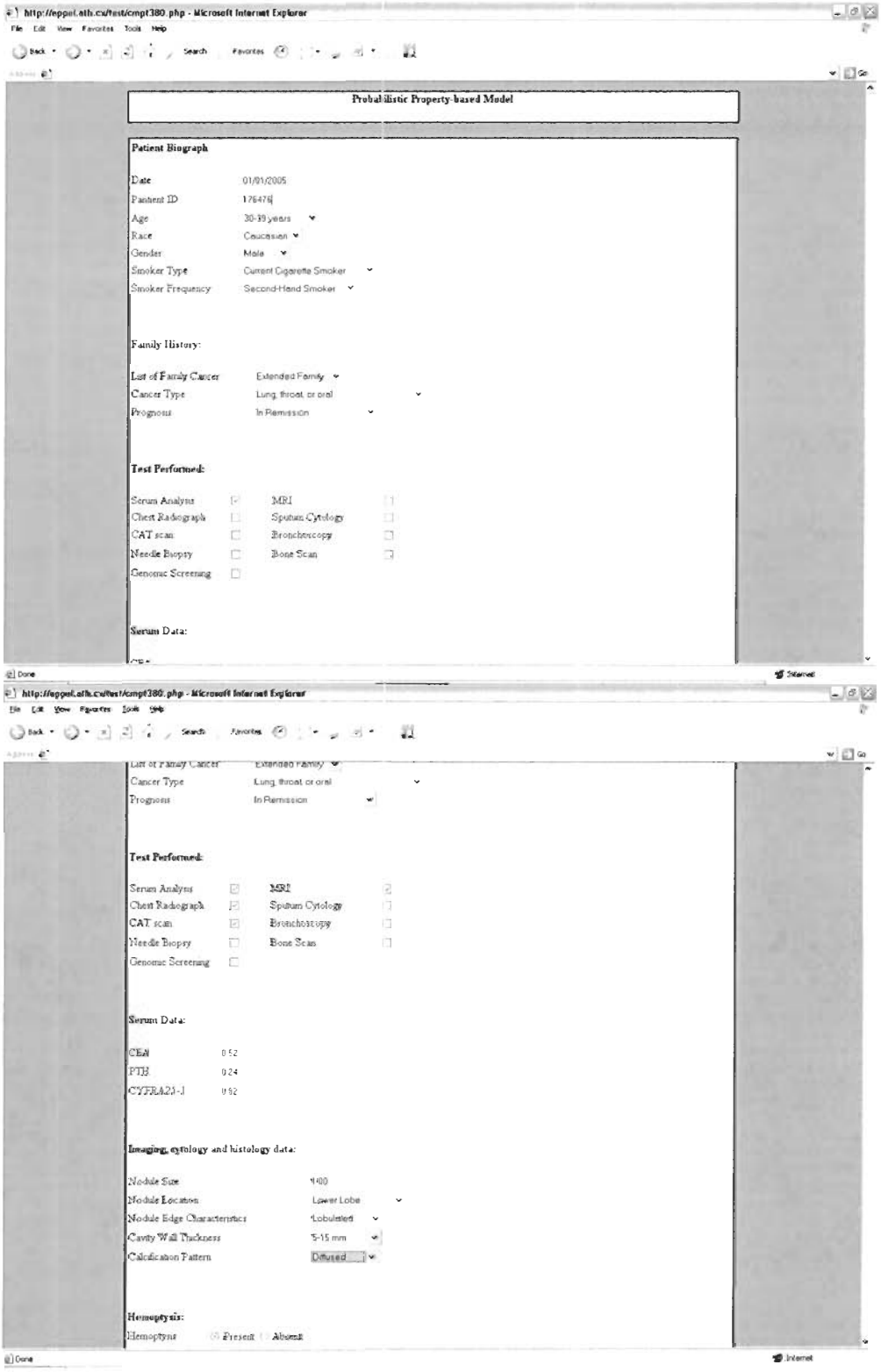
4.3.1 User Interface

We considered the patient's age, smoking history, malignancy history, radiological, serum and sputum data as part of the input concepts to our model (see Fig. 1). This component enables the user to enter the patient's available information in a simple way and generates a date-stamped session knowledge base (KB), which will be added to the Patient Data Concepts.

The user interface was developed using mySQL/PHP. The reason we decided on this implementation was that it creates a more dynamic program. All required concepts and possible values are retrieved from the Knowledge Store and are used to populate the interface controls (e.g. Dropdown Lists, Radio Button Lists). This generates the HTML page dynamically, being driven by the database. Therefore, the user can create a concept group (e.g., Family History) and give it different web controls (e.g., a dropdown list of relationships, a dropdown list of types of cancer). Then a corresponding entry is entered into a

table that keeps track of the values of each type of web control, where the user can specify the different types of relationships possible. (Figure 4.2)

Figure 4.2: PPBM User Interface

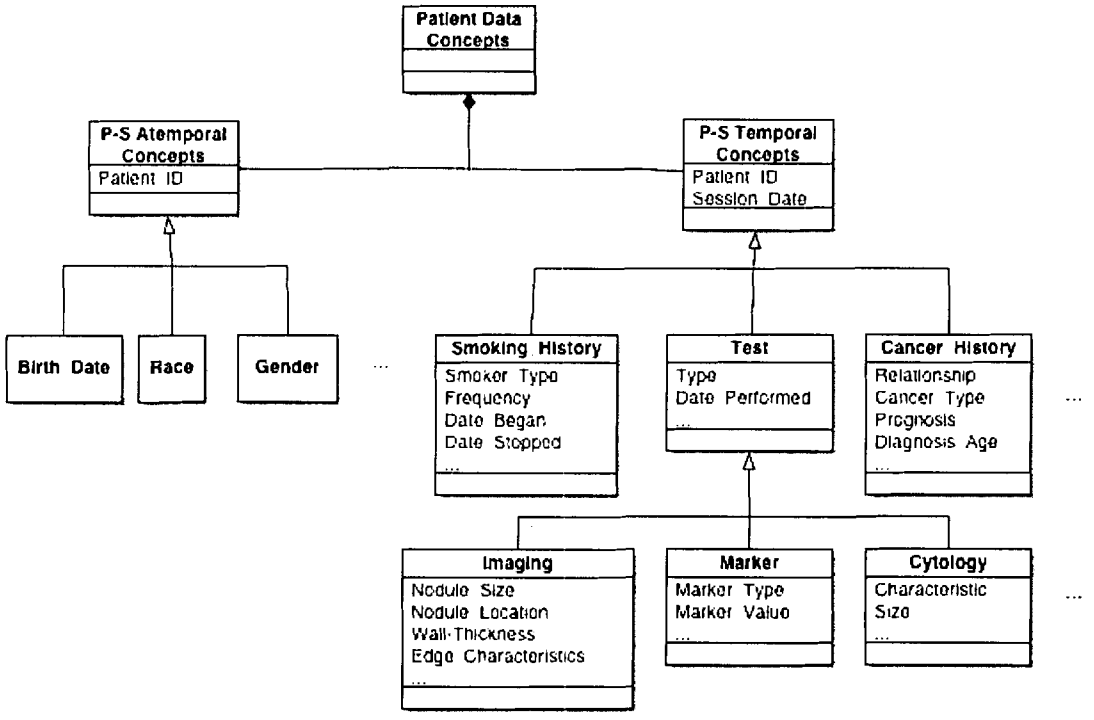


4.3.2 Patient Data Concepts

The Patient Data Concepts (PDC) is where the collection of data for a particular patient is stored. There are two main classes of concepts stored in the PDC: Patient-Specific Temporal Concepts (PSTC) and Patient-Specific Atemporal Concepts (PSAC) (Figure 4.3). In each of them, the patient's condition state is registered based on the available information at each particular session. The concepts associated to PSAC are those that do not change from session to session such as race, gender, birth date, etc. On the other hand, PSTCs may experience change from session to session, e.g., some laboratory tests may not be performed or the values may change as time passes or treatment occurs.

Figure 4.3: Patient Data Concepts Hierarchy Diagram

The diagram shows the hierarchical relationships of some sample concepts that may be present in the PDC at a given time.



To represent this knowledge, we have chosen to follow the CFR approach by defining each piece of diagnostic information as a *concept*. A *concept*, from an object-oriented approach, could be roughly considered as a class. All these PDCs follow a very similar general predicate syntax format:

$$\texttt{concept\_name(Patient\_ID, attribute}_0\texttt{, attribute}_1\texttt{, ..., attribute}_n\texttt{)} \quad (1)$$

Patient_ID is the key attribute, which is unique to each patient. This addresses some critical personal privacy issues to which medical data is subject to and which would be violated if other private identifiable information such as patient's name were stored. For PSTCs, $attribute_0$ = *Session_Date* is the date of the present session, which is stored in the format YYMMDD (two last digits for

year, two digits for month and two digits for days). The rest of $attribute_j$, $1 \leq j \leq n$, are the characteristics unique to a particular concept. For example, the concept

```
serum_data(1232, 050112, CYFRA21-1, 0.5, true)          (2)
```

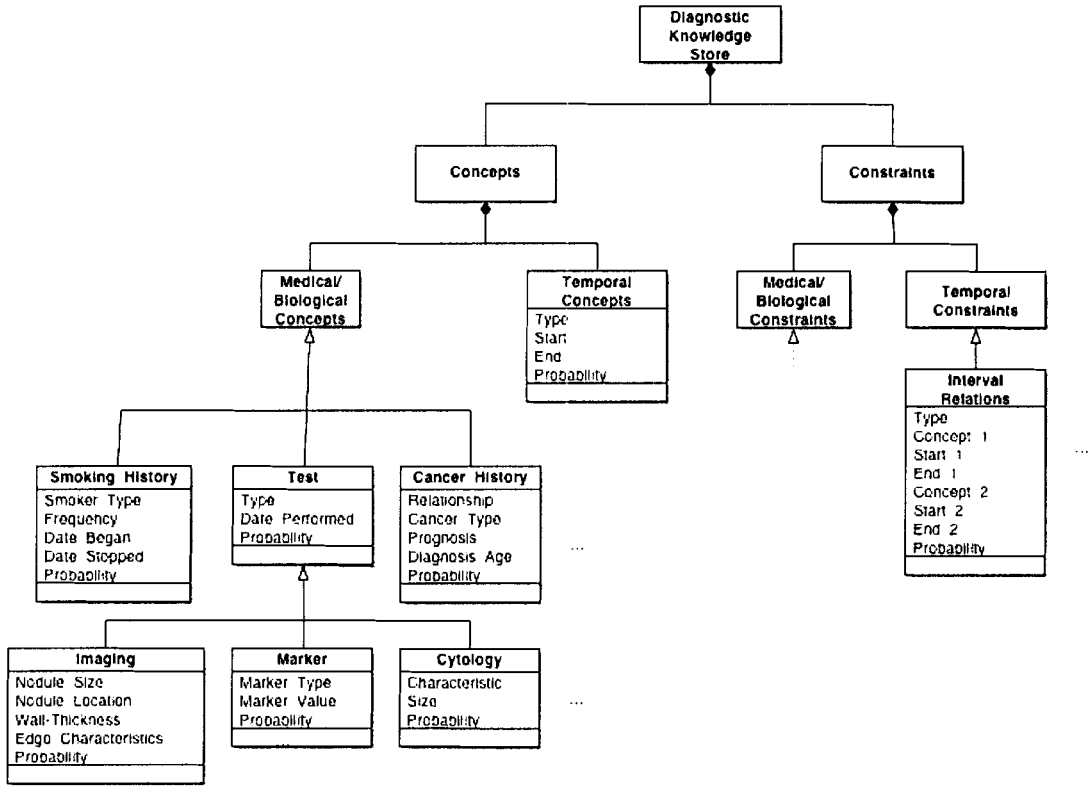
would represent, based on (1) the concept *serum_data*, where $attribute_1$ = *Serum_marker_type*, $attribute_2$ = *Value*, and $attribute_3$ = *In_range.* Therefore, (2) can be interpreted as the lab results obtained from the blood sample of the patient with *Patient_ID* = 1232 on the *Session_Date* = January 12, 2005. The results show that *Serum_marker_type* = CYFRA21-1 had a level *Value* = 0.5, which makes *In_range* = true as it is within the abnormal marker range.

4.3.3 Diagnostic Knowledge Store

The Diagnostic Knowledge Store (DKS) is the part of the KB that includes the properties that should be evaluated for each input data element as well as the relations amongst them. The DKS includes four types of concept rules: Medical/Biological Concepts, Temporal Concepts, Medical/Biological Constraints and Temporal Constraints. (Figure 4.4)

Figure 4.4: Diagnostic Knowledge Store Hierarchy Diagram

The diagram shows the hierarchical relationships of some sample concepts that may be present in the DKS at a given time.



The diagnostic analysis —performed in the Diagnostic Engine— is given as a likelihood probability of cancer that is calculated, as explained above, as a function of the concepts and constraints used in the analysis. As well, the diagnosis will list those diagnostic properties that were satisfied and those that were not using the *relax* rules. For example:

```
const(Prob), age(Patient_ID, Age, Prob), smoking_history(Patient_ID, _,
Smoker_type, Frequency, Prob), serum_data(Patient_ID,
Serum_marker_type, Value, In_range, Prob) <=> marker(Patient_ID,
Marker_type, In_range, P, B), acceptable(marker(Patient_ID,
Marker_type, In_range, P), B), Probability(Patient_ID, P, Prob, B),
acceptable(probability(Patient_ID, P, Prob), B)|
possible_lung_cancer(Patient_ID, true, Prob).

relax(marker(Patient_ID,Marker_type,In_range,P,B)).                    (3)
```

This rule evaluates for a patient with ID *Patient_ID* if a specific biomarker, *Marker_type*, found in serum data is within a certain value range for a patient with an age of *Age* who is a *Smoker_type* smoker (depending on the number of cigarettes or cigars smoked daily). If true, then the diagnosis of possible lung cancer is going to be true with a probability increase of ***P*** (where ***P*** is a function of the patient's age, health history, and this particular biomarker presence). But if we relax the requirement of the presence of the biomarker, then the system can evaluate patient records that do not have this particular information and report in the diagnosis listing that this information was not included in the record, which could be valuable information as recommended follow-up tests for that particular patient.

Several of the probabilities used in the prototype were based on the research findings in [Cummings *et al.* 1986] and [Gurney. 1993], which determined likelihood ratios for the incidence of various clinical and radiographic features of a lung nodule, where

Likelihood Ratio = Probability in patients with disease / Prob. in subjects without disease
= Sensitivity/(1 – specificity) (4)

Then using Bayes' theorem, where

Current Odds = Prior Odds x Likelihood Ratios; (5)

determined the probability of malignancy in solitary pulmonary nodules.

4.3.4 Diagnostic Engine

As mentioned earlier, Property Grammars rely exclusively on constraints. Therefore, for the implementation of the Diagnostic Engine (DE), we use a

specific constraint programming language called Constraint Handling Rule Grammars (CHRG) described in [Christiansen. 2001] on top of CHR. Using CHRGs also gives the possibility of dealing with assumptions [Christiansen and Dahl. 2004], [Dahl and Tarau. 2004], —hence incorporating some hypothetical reasoning— since CHRGs include assumptions whereas CHR do not. CHRGs are implemented using SICStus Prolog.

The basic mechanism in constraint satisfaction problems is to find, for a given set of variables, an assignment that satisfies the constraint system. In the problem addressed here, the variables are taken from the set of categories. An assignment is given from an input (i.e. the PDCs to be parsed). Starting from the set of categories corresponding to the information available for a patient's particular session, all possible assignments (i.e. subsets of categories) are evaluated. When a DKS category is characterized, it is added to the set of categories to be evaluated. This approach is basically incremental in the sense that any subset of categories can be evaluated following a bottom-up approach. This means that adding other categories can complete an assignment ***A***. When DKS categories are inferred after the first step of the process, it is then possible to complete the first assignments (made with PDC categories) with new DKS ones.

The role of selection constraints is central to this approach. As shown before, they allow for the selection of the characterized category. This is due to the fact that such constraints are local to this category. Moreover, in some cases they have a global scope over the category: their satisfiability value (i.e. satisfied

or violated) cannot change for a given category whatever the subset of constituents. As soon as the constraint can be evaluated, this value is permanent. For example, when a linearity or a dependency constraint is satisfied, adding new constituents to the category cannot change this fact. Other kinds of constraints have to be re-evaluated at each stage. For instance, when adding a new category, we need to verify that unicity and exclusion are still satisfied. These are also called filtering constraints. As opposed to selection constraints, the realization of a DKS category cannot be inferred from their evaluation. Their filtering role is in the sense that they rule out some construction.

The principle consists in completing original assignments with new categories when they are inferred. As long as the evaluation of selection constraints (as soon as this evaluation can be performed) is valid through a complete assignment, whatever its constituents, it is not necessary to re-calculate it. In other words, when an assignment *A* is made by completing another assignment *B*, *A* inherits the set of selection constraints of *B*.

As mentioned above, with CHRG as the core component of our DE, we can perform the following operations inherent to the CHR engine:

- Information selection: a side effect of the engine's search over applicable rules.
- Transformation of information: when a rule triggers, it augments the knowledge store with the concept newly formed.
- Hypotheses: made through assumptions.

CFR, another core component of the DE, provides a flexible cognitive structure through relaxable, directly executable properties between concepts called *concept formation rules*, whose guard may include any number of *property calls* for our defined properties. These properties must follow these defined characteristics: [Dahl and Blache. 2004]

> a) A property must be named and defined through the binary predicate *prop*, whose first argument is the property's name and whose second argument is the list of arguments involved in checking, and in signaling the results of checking, the property. [...]
>
> b) Acceptability of a property that has thus been defined must be checked in the concerned rule through the binary system predicate *acceptable*, whose first argument is the *prop* atom with all its arguments and whose second argument will evaluate to either true, false, or a degree of acceptability, according to whether (or how much of) the property is satisfied. [...]
>
> c) In order to relax a property named N (i.e. to allow the derivation of concepts that require it but for which it is not satisfied), we simply write the following:
>
> ```
> relax(N).
> ```
>
> Degrees of acceptability can be defined through a binary version of the relaxing primitive, where L is the *prop* atom with all its arguments and D is a measure of acceptability:
>
> ```
> relax(L,D).
> ```
>
> A list of satisfied and violated properties, together with the degree of violation if appropriate, will be output for each property defined in a given CF program.

The Temporal Reasoning component is an extension of CFR that establishes temporal conditions based on Allen's basic interval relations, defined as CFR properties, as follows (inverse not shown):

```
prop(before,[End_Time_1,Start_Time_2]):- End_Time_1<Start_Time_2.

prop(meets,[End_Time_1,Start_Time_2]):- End_Time_1=Start_Time_2.

prop(starts,[Start_Time_1,End_Time_1,Start_Time_2,End_Time_2]):-
Start_Time_1=Start_Time_2, End_Time_1<End_Time_2.

prop(during,[Start_Time_1,End_Time_1,Start_Time_2,End_Time_2]):-
Start_Time_1>Start_Time_2, End_Time_1<End_Time_2.

prop(overlaps,[Start_Time_1,End_Time_1,Start_Time_2,End_Time_2]):-
Start_Time_1<Start_Time_2, End_Time_1<End_Time_2,
Start_Time_2<End_Time_1.

prop(finishes,[Start_Time_1,End_Time_1,Start_Time_2,End_Time_2]):-
Start_Time_1>Start_Time_2, End_Time_1=End_Time_2.

prop(equals,[Start_Time_1,End_Time_1,Start_Time_2,End_Time_2]):-
Start_Time_1=Start_Time_2, End_Time_1=End_Time_2.
```

CHAPTER 5: CONCLUSION

In this document a model was presented to demonstrate how the growth of pre-invasive neoplastic bronchial epithelial lesions might proceed. The development of this model was based on cytological data obtained from [Clem *et al.* 1992]. By applying a stochastic and a PDE approach it was possible to represent the early development of pre-invasive neoplastic bronchial epithelial lesions as contact processes. Furthermore, it was shown that:

i. Normal and abnormal cells could be represented as two different competing populations (*see Thesis 1* in Chapter 1, shown in Section 3.4) and, hence,

ii. We were able to develop a particle system that modeled the structural behaviour of pre-invasive bronchial epithelial lesions (*see Thesis 2*, shown in Section 3.5).

iii. This tractable system enabled us to determine the likelihood probability of growth from a 2-D section of the lesion (*see Thesis 3*, shown in Section 3.7).

The results of said model showed that the sole structural analysis of independent pre-invasive bronchial epithelial lesions (even though it gives some insight on the lesion growth process, and, in fact, could be used to determine how likely that particular lesion is to develop abnormal growth upon regular

observation) does not provide sufficient evidence to make an accurate diagnosis of the likelihood of a patient developing lung cancer. Only a small fraction of these lesions actually progress to a malignant tumour and this analysis focuses only on a single lesion analysis. While one lesion may not develop cancer, others could. For use in diagnosis, this approach would require regular bronchoscopies performed on the patients (to derive a time course analysis to obtain actual data as the legion progresses in size), which would be costly and highly stressful for them. Based on this insight, a new model had to be developed that would use the results of the first model plus results from other medical tests. As a result, the Probabilistic Property-based Model or PPBM, is proposed for representation and analysis of multi-disciplinary biological data, which could include not only cytological data, but imaging, serum, sputum, and genetic data, as well as it considers personal and lifestyle factors such as race, age, gender, family cancer history, and smoking history. In the PPBM approach, it was shown that:

i. Medical and biological knowledge could be represented in terms of concepts and constraints, regardless of the diversity of their sources (*see Theses 4 and 5*, shown in Sections 4.2.1 and 4.2.3).

ii. Relationships and interactions between multidisciplinary biological data have an impact on the likelihood probability of development of a disease (*see Thesis 7*, shown in Section 4.2.3.5), and

iii. These interactions and relationships can be represented in a relatively simple way in the PPBM in terms of constraint systems (*see Thesis 6*, shown in Section 4.3).

iv. As well, some basic temporal analysis was incorporated, as it is a critical component of the diagnostic process (shown in Section 4.2.2).

A specific proof of concept prototype for assistance in the early diagnosis of cancer was presented.

Except for the work by Clem et al. 1992, 1995, 1997a, and Barranco et al. 1997, 1998, 1999 there are no other reported studies on the development of computational models for the growth of pre-invasive neoplastic bronchial epithelial lesions in the scientific literature. This thesis provides new insight on mathematical modelling of these critical lesions and is a first in the development of a system that would provide a means of unifying the many different medical and biological data sets that are characteristics of bronchial epithelial lesions. It is the intention that these models would aid in the development of a software application that could aid doctors in the early diagnosis and management of lung and other types of cancer.

5.1 Future Research Directions

5.1.1 System for Early Diagnosis of Oral Cancer

The next obvious step in this research is to develop the PPBM prototype into a system to test with real-life data. Dr. Miriam Rosin, primary investigator on an extensive Oral Health Study [BC Cancer Agency. 2004], taking place in British Columbia by the Cancer Control Research program of the BCCRC, has expressed interest in the incorporation of the PPBM to assist with the analysis of the data obtained in this study. The objective of this study is for dentists to screen

patients who present oral lesions that are believed to be early precursors of oral cancer, similar to bronchial epithelial lesions are to lung cancer. Yet, they also need to take into account other medical and lifestyle factors —such as cancer history in the patient's family and smoking history— to make a more accurate diagnosis and not overburden the health system by referring to the oncologist patients that may not really require it. This requirements fit well with the design characteristics of the PPBM. In addition, for a CAD system to support this study, it must be able to allow diagnosis with an incomplete set of tests and refine the diagnosis throughout time as new tests are being performed (another characteristic of PPBM). More over, the system has to take into consideration the cost of tests and weight it against the possible gains in terms of diagnostic accuracy to determine if they should be recommended or not (a factor also taken into consideration in our model).

The biological characteristics of oral cancer in many aspects are quite similar to those of lung cancer and many of the diagnostic tests and environmental/lifestyle factors are as well. Although not as common as lung cancer, the prognosis of oral cancer is also very bad as it is very hard to diagnose at the early stages as well. The design of the PPBM is generic enough to support different types of cancer and, in fact, different disease domains. The close similarities between oral and lung cancer will permit the reuse of some of the knowledge engineering already done in the prototype for lung cancer diagnosis.

For the implementation of this system a more specialized constraints engine may be required. We based our prototype implementation in CHR due to its simplicity, ease and speed of implementation yet, in a production setting, it might not be able to meet the performance requirements of a real-life scenario.

5.1.2 Temporal Reasoning

As explained in section 4.2.2.1, the way the PPBM currently deals with temporal reasoning is solely based on a simple time interval algebra. Several temporal aspects critical to the medical domain such as recurrent or cyclic temporal concepts are not yet being modelled in PPBM. This is another natural area of expansion for this model.

5.1.3 Other Medical Domains

The implementation of the knowledge base for another disease other than lung and oral cancer will validate the PPBM paradigm as an effective and efficient model for multi-disciplinary biological data representation in different medical domains. We have begun the analysis for the implementation of the knowledge base for Type 2 Diabetes in PPBM.

5.1.4 Data Mining/Machine Learning Component

Research in the field of Medical Bioinformatics in the last decade has been characterized by the efforts to bridge the gap between the large amounts of uninterpreted data and the understanding of such data. Thus, the research emphasis is now on data analysis. Data mining, knowledge discovery in databases, and intelligent data analysis, along with machine learning techniques,

are the latest focus areas of medical computing research [Barranco-Mendoza. 2004].

The need for intelligent data analysis in medicine is evident in the following: (i) for instance, to support the analysis of individual patients' raw data of specific knowledge-based problem solving activities such diagnosis, prognosis, monitoring, treatment planning, etc. and (ii) the use of data mining in the discovery of new medical knowledge that can be extracted from collections of example cases. [Lavrac *et al.* 2000]

Hence the design and development of a data mining/machine learning component will automatically analyse the existing patient data and infer new rules or behaviour patterns that would help elucidate determinant factors that may not have been previously identified, or to determine that other factors previously identified as determinant as not as important in the development of lung or oral cancer as previously thought. An approach to do this would be to design a component to do regression and time series analysis from the data represented using PPBM. These two techniques have show to be the most effective analysing data which include probabilistic and time dependent elements. [Elmasri and Navathe. 2004]

5.1.5 Natural Language Interface

As explained above, there are large amounts of uninterpreted medical data, in particular in the form of patient records. Developing a natural language interface/parser to enter knowledge in the PPBM would allow the incorporation of

data not initially predifined. This will prove particularly useful to infer new knowledge through a datamining/machine learning component. To do this, one must consider the use of available standard ontologies such as UMLS. Some of the component elements of PPBM, namely Property Grammars and CHR, have been used in the development of natural language parsers [Blanche. 2000], [Dahl and Blanche. 2004], which would ease the integration with our model.

APPENDIX A: GLOSSARY OF BIOLOGICAL TERMS

Anaplastic Cells: Cells that have been reversed to a more primitive or undifferentiated form.

Biopsy: The removal and examination of tissue, cells, or fluids from the living body.

Cytomorphometric Analysis: Analysis of measurement of external form of cells.

Differentiation: How developed the cancer cells are in a tumour. Well-differentiated tumour cells resemble normal cells and tend to grow and spread at a slower rate than undifferentiated or poorly differentiated tumour cells, which lack the structure and function of normal cells and grow uncontrollably.

Imaging Data: Data obtain from radiographs (X-rays), CT scans, MRI, etc.

Mass Spectrometry: An instrumental method for identifying the chemical constitution of a substance by means of the separation of gaseous ions according to their differing mass and charge.

Metastasis: a: transfer of a malignant tumour from the site of disease to another part of the body b: a secondary metastatic growth of a malignant tumour.

Mitotic Cells: Cells in the process of cell division.

Neoplasm: See Tumour

Nodule: a small abnormal knobby bodily protuberance (as a tumorous growth).

Pleomorphic Tumour Cells: tumour cells that proliferate quickly and are quite different from benign tumour cells.

Pleura: The delicate serous membrane that lines each half of the thorax of mammals and is folded back over the surface of the lung of the same side.

Protoplasm: The organized colloidal complex of organic and inorganic substances, as proteins and water, that constitutes the living nucleus, cytoplasm, plastids, and mitochondria of the cell.

Protoplasmic Framework: See Protoplasm

Radon: A heavy radioactive gaseous element formed by the decay of radium.

Serum Data: Data obtained from analysis of blood samples.

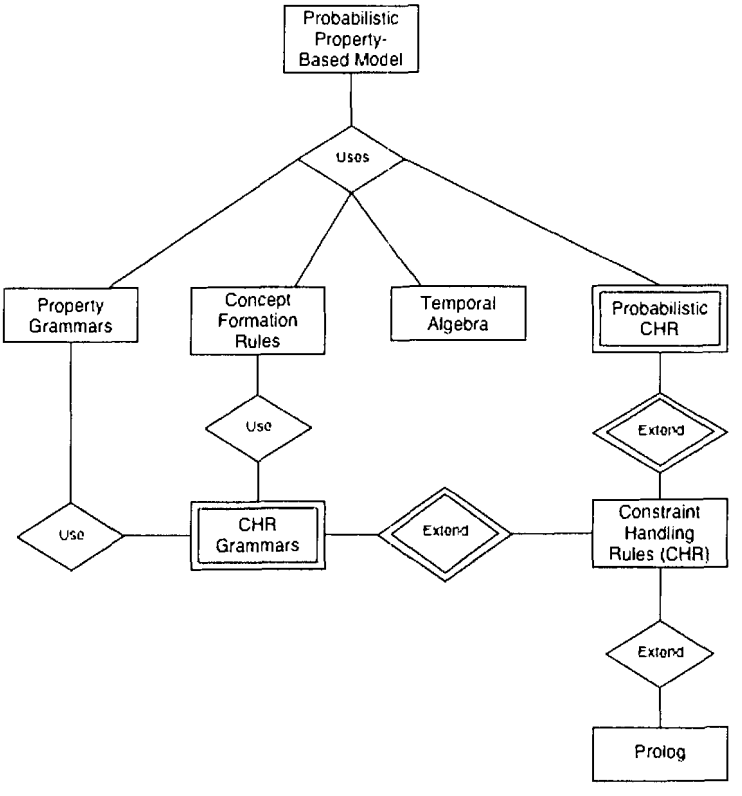
Stroma: the spongy protoplasmic framework of some cells.

Tumour: Abnormal cell growth that has no useful function to the host.

APPENDIX B: RELATIONSHIPS BETWEEN THE FORMALISMS USED IN PPBM

Figure B.1 shows the relationships between the different formalisms used within the PPBM, explained in Chapter 4. The "Use" relationship indicates that a formalism has utilized elements of another. The "Extend" relationship indicates that a formalism includes another one in its totality and have added extensions to it. Relationships are to be read top to bottom and left to right.

Figure B.1: Relationship Diagram of the Formalisms used within the PPBM



APPENDIX C: EXAMPLES OF CONCEPTS AND CONSTRAINTS FROM THE PPBM'S DIAGNOSTIC KNOWLEDGE STORE

Examples of Medical/Biological Concepts

These concepts calculate the likelihood probability score depending on the age of the patient. The concept `birth_date` is from the Patient-Specific Atemporal Concepts and The concept `current_session` is from the Patient-Specific Temporal Concepts in the Patient Data Concepts knowledge base. If the patient is older than 50 years old then there will be an increase in likelihood:

```
birth_date(Patient_ID,_, _, Birth_year),
current_session(Patient_ID,_, _, Current_year) <=> Age =
Current_year – Birth_year: Age =< 50 | age(Patient_ID, Age,
0).
birth_date(Patient_ID,_, _, Birth_year),
current_session(Patient_ID,_, _, Current_year) <=> Age =
Current_year – Birth_year: Age >= 50 | age(Patient_ID, Age,
0.1).
```

This concept determines the likelihood probability score for the concept `personal_data`, which is the collection of Patient-Specific Atemporal Concepts for a particular patient. This is determined by calculating the average probability scores of `age`, `race`, and `gender`.

```
age(Patient_ID, Age, Prob1), race(Patient_ID, Race, Prob2),
gender(Patient_ID, Gender, Prob3) <=> Prob =
(Prob1+Prob2+Prob3)/3 | personal_data(Patient_ID, Age,
Race, Gender, Prob)
```

Example of Medical/Biological Constraints

This constraint calculates the decrease in likelihood probability score based on the gender of the patient if she is female. This constraint will only be triggered if there was already at least one other constraint that set `possible_lung_cancer` to true (which is a greater than zero probability that the patient may have lung cancer, even if the value is very small) and there are no more constraints to be triggered (last constraint so the likelihood probability score only gets decreased once). The concept `gender` is from the Patient-Specific Atemporal Concepts:

```
possible_lung_cancer(Patient_ID, true, Prob),
gender(Patient_ID, Gender, Prob1), no_more_constraints <=>
Gender = 'F' | possible_lung_cancer(Patient_ID, true,
Prob*0.9).
```

This constraint evaluates if a specific biomarker, `Marker_type`, found in serum data is within a certain value range for a patient with an age of `Age`. If true, then the diagnosis of possible lung cancer is going to be true with a probability increase of `P`. This is calculated within the `marker` constraint.

```
age(Patient_ID, Age, Prob1), serum_data(Patient_ID, Serum_marker_type,
Value, In_range, Prob2) <=> marker(Patient_ID, Marker_type, In_range,
P, B), acceptable(marker(Patient_ID, Marker_type, In_range, P), B)|
possible_lung_cancer(Patient_ID, true, Prob+P).
relax(marker(Patient_ID,Marker_type,In_range,P,B)).
```

Notice that if we relax the requirement of the presence of the biomarker, then the system can evaluate patient records that do not have this particular information and report in the diagnosis listing that this information was not included in the record, which could be valuable information as recommended follow-up tests.

REFERENCE LIST

Allen, J. (1984) Towards a General Theory of Action and Time, *Artificial Intelligence*, **23**, 123-154.

Allen, J. (1983) Maintaining Knowledge about Temporal Intervals, *ACM Comm.*, **26** (11), 832-843.

Altman, R.B. (1999) AI in Medicine, *AI Magazine* (Fall).

Baker, S.G., Kramer, B.S., and Srivastava, S. (2002) Markers for early detection of cancer: Statistical guidelines for nested case-control studies , *BMC Medical Research Methodology*, **2** (4), 1471-2288.

Barnett, G.O., Cimino, J.J., Hupp, J.A., and Hoffer, E.P. (1987) DXPLAIN. An Evolving Diagnostic Decision-Support System , *Journal of the American Medical Association*, **258** (1), 67-74.

Barranco-Mendoza, A., Persaud, D.R., Eppel, G., Farrant, B., and Dahl, V. (2005) A property-based model for multi-disciplinary biological knowledge representation and early cancer diagnosis, (submitted)

Barranco-Mendoza, A. (2004) Medical Bioinformatics: A Look at Computer-Aided Diagnosis of Lung Cancer, Ph.D. Depth Research Paper , Simon Fraser University, Burnaby, BC, Canada.

Barranco-Mendoza, A., Persaud, D.R., and Dahl, V. (2004) A property-based model for lung cancer diagnosis, (ed. Gramada, A. and Bourne, P.E.), *Currents in Computational Molecular Biology*, 558-559.

Barranco-Mendoza, A., Clem, C., Gupta, A., Fizzano, P., and Guillaud, M. (1999) Predicting the development of pre-invasive lesions from biopsies, special issue on Control and Estimation in Biological and Medicine Sciences, *Archives of Control Sciences*, **9**, 25-40.

Barranco-Mendoza, A. and Gupta, A. (1998) Modelo Computacional del Desarrollo de Lesiones Neoplásticas Preinvasivas del Epitelio Bronquial, Proceedings of General Congress Computo98 (Medical Informatics track), Mexico City, Mexico.

Barranco-Mendoza, A., Clem, C., and Gupta, A., et al (1997) Modelling Pre-invasive Bronchial Epithelial Lesions, Proceedings 1st Pacific Institute for the

Mathematical Sciences Industrial Problem Solving Workshop, Vancouver, Canada.

BC Cancer Agency (2005) Lung Cancer Diagnosis (accessed March 1, 2005), http://www.bccancer.bc.ca/PPI/TypesofCancer/Lung/Diagnosis.htm.

BC Cancer Agency (2004) Oral Health Study (last updated May 20 2004), http://www.bccrc.ca/ccr/mrosin_oralHealth.html (accessed March 20, 2005).

Bes, G. and Blache, P. (1999) Proprietes et analyse d'un langage, Proceedings of TALN'99

Bes, G., Blache, P., and Hagege, C. (1999) The 5P Paradigm, GRIL/LPL.

Blache, P. (2000) Constraints, Linguistic Theories, and Natural Language Processing, Natural Language Processing - NLP 2000: Second International Conference, Patras, Greece, June 2000. Proceedings, (ed. Christodoulakis, D.N.), *Lecture Notes in Computer Science*, **1835**, 221-233, June 2003, Springer-Verlag GmbH.

Bouhaddou, O., Lambert, J.G., and Morgan, G.E. (1995) Iliad and the Medical House Call: Evaluating the Impact of Common Sense Knowledge on the Diagnostic Accuracy of a Medical Expert System, Nineteenth Annual Symposium on Computer Applications in Medical Care, 742-746, Salt Lake City, Utah, USA, Applied Medical Informatics.

Bramson, M. and Griffeath, D. (1981) On the Williams-Bjerknes tumor growth model, I , *Ann. Prob.*, **9**, 173-185.

Bramson, M. and Griffeath, D. (1980) On the Williams-Bjerknes tumor growth model, II , *Proc. Camb. Phil Soc.*, **88**, 339-357.

Bratko, I. and Kononenko, I. (1987) Learning diagnostic rules from incomplete and noisy data, AI Methods in Statistics, (ed. Phelps, B.), Gower Technical Press, London.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984) Classification and Regression Trees, Wadsworth, Belmont.

Brown, M.S., Goldin, J.G., Suh, R.D., McNitt-Gray, M.F., Sayre, J.W., and Aberle, D.R. (2003) Lung micronodules: automated method for detection at thin-section CT--initial experience, *Radiology*, **226** (1), 256-262, United States.

Buchanan, B. and Shortliffe, E. (1984) Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristics Programming Project, (ed. Anonymous), Addison-Wesley, Menlo Park, Calif., USA.

Canadian Cancer Society (2005) Lung cancer stats (last updated February 3 2005), http://www.bc.cancer.ca/ccs/internet/standard/0,3182,3278_14459_371459_langId-en,00.html.

Cestnik, B., Kononenko, I., and Bratko, I. (1987) ASSISTANT 86: A knowledge elicitation tool for sophisticated users, Progress in Machine learning, (ed. Bratko, I. and Lavrac, N.), Sigma Press, Wilmslow.

Chellappa, R., Chatterjee, S., and Bagdazian, R. (1985) Texture Synthesis and Compression using Gaussian-Markov Random Fields Models , *IEEE Transactions on Systems, Man and Cybernetics*, **15**, 298-303.

Chen, Y., Nixon, M., and Thomas, D. (1995) Statistical Geometrical Features for Texture Classification , *Pattern Recognition*, **4**, 537-552.

Chiou, Y.S.P., Lure, Y.M.F., and Ligomenides, P.A. (1993) Neural network image analysis and classification in hybrid lung nodule detection (HLND) system, Proceedings of the IEEE-SP Workshop on Neural Networks for Signal Processing , 517-526.

Christiansen, H. and Dahl, V. (2004) Assumptions and Abduction in Prolog, Proc. 3rd International Workshop on Multiparadigm Constraint Programming Languages , Saint-Malo, France.

Christiansen, H. (2001) CHR as grammar formalism, a first report, Sixth Annual Workshop of the ERCIM Working Group on Constraints

Clark, P. and Boswell, R. (1991) Rule induction with CN2: Some recent improvements, Proc. Fifth European Working Session on Learning, 151–163, Springer.

Clark, P. and Niblett, T. (1989) The CN2 induction algorithm, *Machine Learning*, **3** (4), 261–283.

Clausi, D.A. and Zhao, Y. (2002) Rapid extraction of image texture by co-occurrence using a hybrid data structure, *Comput. Geosci.*, **28** (6), 763-774, Pergamon Press, Inc.

Clem, C.J., Boysen, M., and Rigaut, J.P. (1992) Towards 3-D modelling of epithelia by computer simulation, *Analytical Cellular Pathology*, **4**, 287-302.

Clem, C.J., Guillaud, M., and MacAulay, C. (1997a) Computer simulation of the 3-D tissue organizational changes associated with development of pre-invasive neoplastic bronchial epithelial lesions, 11th International Congress on Diagnostic Quantitative Pathology, Siena, Italy.

Clem, C.J., König, D., and Rigaut, J.P. (1997b) A three-dimensional dynamic simulation model of epithelium tissue renewal, *Analytical and Quantitative Cytology and Histology*, **19**, 174-184.

Clem, C.J. and Rigaut, J.P. (1995) Computer simulation modelling and Visualization of 3-D architecture of biological tissues, *Acta Biotheoretica*, **43**, 425-442.

Combi, C., Pinciroli, F., and Pozzi, G. (1995) Managing different time granularities of clinical information by an interval-based temporal data model, *Methods of Information in Medicine*, **34** (5), 458-474.

Combi, C. and Shahar, Y. (1997) Temporal reasoning and temporal data maintenance in medicine: issues and challenges, *Comput. Biol. Med.*, **27** (5), 353-368, UNITED STATES.

Cross, G. and Jain, A. (1983) Markov Random Field Texture Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **5**, 25-39.

Cukierman, D. and Delgrande, J. (2000) A Formalization of Structured Temporal Objects and Repetition, TIME '00: Proceedings of the Seventh International Workshop on Temporal Representation and Reasoning (TIME'00), 13, IEEE Computer Society.

Cummings, S.R., Lillington, G.A., and Richard, R.J. (1986) Estimating the Probability of Malignancy in Solitary Pulmonary Nodules, *Am. Rev. Respir. Dis.*, **134**, 449-452.

Dahl, V. and Blache, P. (2004) Directly Executable Constraint Based Grammars, Proc. Journees Francophones de Programmation en Logique avec Contraintes , Angers, France.

Dahl, V. and Tarau, P. (2004) Assumptive Logic Programming, Proceedings of ASAI'04 , Cordoba, Spain.

Dahl, V. and Voll, K. (2004) Concept Formation Rules: An Executable Cognitive Model of Knowledge Construction, Proceedings of the Natural Language Understanding and Cognitive Science Workshop, *ICEIS04*, 28-36, Porto, Portugal.

Das, A.K. and Musen, M.A. (1994) A Temporal Query System for Protocol-Directed Decision Support, *Methods of Information in Medicine*, **33**, 358-370.

Devore, J.L. (1991) Probability and Statistics for Engineering and the Sciences, 709, Third Edition, Duxbury Press, Belmont, California.

Dickman, R., Rabelo, W.R.M., and Odor, G. (2002) Pair contact process with a particle source , *Phys. Rev. E*, **65** (1), 016118, APS.

Druzdzel, M.J. and Diez, F.J. (2003) Combining knowledge from different sources in probabilistic models, *Journal of Machine Learning Research*, **4** (July), 295-316.

Durrett, R. and Levin, S. (1994) The importance of being discrete (and spatial), *Theoret. Pop. Biol.*, **46**, 363-394.

Durrett, R. (1992) The contact process: 1974-1989, Mathematics of Random Media, (ed. Kohler, W.E. and White, B.S.), 1-18, American Math. Society.

Durrett, R. and Griffeath, D. (1982) Contact processes in several dimensions, *Z. fur Wahr*, **59**, 535-552.

Durrett, R. (1980) On the growth of one dimensional contact processes, *Ann. Prob.*, **8**, 890-907.

Early Detection Research Network (2002) The Early Detection Network - Translational Research to Identify Early Cancer and Cancer Risk, Second Report http://www3.cancer.gov/prevention/cbrg/edrn/edrn_report2002.pdf

Elmasri, R. and Navathe, S.B. (2004) Fundamentals of database systems (4th ed.) , Benjamin-Cummings Publishing Co., Inc.

Faugeras, O. and Pratt, W. (1980) Decorrelation Methods of Texture Feature Extraction, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**, 323-332.

Fausett, L.V. (1994) Fundamentals of neural networks: Architectures, algorithms and applications , Prentice Hall, Upper Saddle River, NJ.

Franke, C., Böhner, H., Yang, Q., Ohmann, C., and Röher, H. (1999) Ultrasonography for Diagnosis of Acute Appendicitis: Results of a Prospective Multicenter Trial , *World J. Surg.*, **23** (2), 141-146.

Frawley, W., Piatetsky-Shapiro, G., and Matheus, C. (1991) Knowledge discovery in databases: An overview, Knowledge discovery in databases, (ed. Piatetsky-Shapiro, G. and Frawley, W.), The AAAI Press, Menlo Park, CA.

Fretz, P.C. and Peterson, M.W. (1996) Malignancy in a Solitary Pulmonary Nodule

Fruhwirth, T. (1998) Theory and Practice of Constraint Handling Rules, *Journal of Logic Programming*, **37**, 1-3.

Frühwirth, T., Di Pierro, A., and Wiklicky, H. (2002a) Probabilistic Constraint Handling Rules , *Electronic Notes in Theoretical Computer Science*, **76**, 16, http://www.elsevier.nl/locate/entcs/volume76.html.

Frühwirth, T., Di Pierro, A., and Wiklicky, H. (2002b) An Implementation of Probabilistic Constraint Handling Rules, 11th International Workshop on Functional and (Constraint) Logic Programming (WFLP 2002)

Frühwirth, T., Di Pierro, A., and Wiklicky, H. (2001) Towards Probabilistic Constraint Handling Rules, Third Workshop on Rule-Based Constraint Reasoning and Programming (RCoRP'01) at CP'01 and ICLP'01, Paphos, Cyprus.

Gealy, R., Zhang, L., Siegfried, J.M., Lutetich, J.D., and Keohavong, P. (1999) Comparison of mutations in the p53 and K-ras genes in lung carcinomas from smoking and nonsmoking women, *Cancer Epidemiology, Biomarkers, & Prevention*, **8**, 297-302.

Gotlieb, C. and Kreyszig, H. (1990) Texture Descriptors based on Co-occurrence Matrices, *Computer Vision, Graphics and Image Processing*, **51**, 70-86.

Gur, D., Zheng, B., Fuhrman, C.R., and Hardesty, L. (2004) On the testing and reporting of computer-aided detection results for lung cancer detection, *Radiology*, **232** (1), 5-6, United States.

Gurney, J.W. (1993) Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis, Part I. Theory, *Radiology*, **186**, 405-413.

Hao Chen and Chen, C.H. (2002) Hyperspectral image data unsupervised classification using Gauss-Markov random fields and PCA principle, Geoscience and Remote Sensing Symposium, 2002. IGARSS '02, **3**, 1431-1433, IEEE International.

Haralick, R. (1979) Statistical and Structural Approaches to Texture , *Proceedings of the IEEE*, **67**, 786-804.

Haralick, R., Shanmugam, K., and Dinstein, I. (1973) Textural Features for Image Classification , *IEEE Transactions on Systems, Man and Cybernetics*, **3** (6), 610-621.

Harris, T.E. (1974) Contact interactions on a lattice, *Ann. Prob.*, **2**, 969-988.

Hayashibe, R., Asano, N., Hirohata, H., and et al. (1996) An automatic lung cancer detection from X-ray images obtained through yearly serial mass survey, Proceedings of the International Conference on Image Processing, **1**, 343-346.

Hucka, M., Finney, A., Sauro, H.M., and et al. (2003) The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models, *Bioinformatics*, **19** (4), 524-531.

Hung, J., Lam, S., LeRiche, J.C., and Palcic, B. (1991) Autofluorescence of normal and malignant bronchial tissue, *Lasers Surg. Med.*, **11**, 99-105.

Kahn, M., Fagan, L., and Tu, S. (1991a) TQuery: A Context-Sensitive Temporal Query Language, *Computers and Biomedical Research*, **24**, 401-419.

Kahn, M., Fagan, L., and Tu, S. (1991b) Extensions to the Time-Oriented Database Model to Support Temporal Reasoning in Medical Expert Systems. , *Methods of Information in Medicine*, **30**, 4-14.

Kanazawa, K., Kubo, M., and Niki, N. (1996) Computer aided diagnosis system for lung cancer based on helical CT images , **3**, 381-385.

Kearsley,G. (2004a) Constructivist Theory (J. Bruner) 2004a), http://tip.psychology.org/bruner.html (accessed March 1, 2005).

Kearsley,G. (2004b) Explorations in Learning and Instruction: The Theory into Practice Database 2004b), http://tip.psychology.org (accessed March 1, 2005).

Keravnou, E.T. and Washbrook, J. (1990) A Temporal Reasoning Framework Used in the Diagnosis of Skeletal Dysplasias, *Artificial Intelligence in Medicine*, **2**, 239-265.

Kulikowski, C. and Weiss, S. (1982) Representation of Expert Knowledge for Consultation: The CASNET and EXPERT Projects, (ed. Szolovits, P.), *Artificial Intelligence in Medicine*, Westview, Boulder, Colorado, USA.

Lam, S., MacAulay, C., leRiche, J.C., and Palcic, B. (2000) Detection and localization of early lung cancer by fluorescence bronchoscopy, *Cancer*, **89** (11 Suppl), 2468-2473.

Lam, S., MacAulay, C., Hung, J., and Palcic, B. (1993) Detection of dysplasia and carcinoma in situ with a lung imaging fluorescence endoscope device , *J. Thorac. Cardiovasc. Surg.*, **105**, 1035-1040.

Lavrac, N., Keravnou, E., and Zupan, B. (2000) Intelligent Data Analysis in Medicine, Encyclopedia of Computer Science and Technology, (ed. Kent et al.), **42**, 113—157, Dekker, New York.

Ledley, R.S. and Lusted, L.B. (1959) Reasoning Foundation of Medical Diagnosis: Symbolic Logic, Probability, and Value Theory Aid Our Understanding of How Physicians Reason, *Science*, **130** (9), 9-12.

Liggett, T.M. (1985) Interacting Particle Systems, Springer, New York, USA.

Liggett, T.M. (1997) Stochastic models of interacting systems, *Ann. Probab.*, **25** (1), 1-29.

Lin, J.S., Lo, S.C.B., Hasegawa, A., and et al. (1995) Reduction of false positives in lung nodule detection using a two-level neural classification, *IEEE Trans. Medical Imaging*, **15** (2), 206-217.

Manjunath, B., Simchony, T., and Chellappa, R. (1990) Stochastic and Deterministic Networks for Texture Segmentation, *IEEE Transactions on Acoustics Speech and Signal Processing*, **38** (6), 1039-1049.

Markey, M.K., Tourassi, G.D., and Floyd Jr. C.E. (2003) Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer, *Proteomics*, **3**, 1678–1679.

Marx, J. (2003) Mutant Stem Cells May Seed Cancer, *Science*, **301**, 1308-1310, News Focus.

Michalski, R., Mozeti¡c, I., Hong, J., and Lavra¡c, N. (1986) The multi-purpose incremental learning system AQ15 and its testing application on three medical domains, Proc. Fifth National Conference on Artificial Intelligence , 1041–1045, Morgan Kaufmann, San Francisco, CA.

Michalski, R.S. (1983) A theory and methodology of inductive learning, Machine Learning: An Artificial Intelligence Approach, **1**, 83–134, Tioga Publishing Company, Palo Alto, CA.

Miller, R.A., Masarie, F., and Myers, J.D. (1986) Quick Medical Reference (QMR) for Diagnostic Assistance, *MD Computing*, **3** (5), 34-48.

Miller, R.A., Pople, H.E., and Myers, J.D. (1982) INTERNIST-1: An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine, *New England Journal of Medicine*, **307** (8), 468-476.

Nagamoto, N., Saito, Y., and Sato, M. (1993) Lesions preceding squamous cell carcinoma of the bronchus and multicentricity of cancerationserial slicing of minute lung cancers smaller than1 mm, *Tohoku J. Exp. Med.*, **170**, 11-23.

National Library of Medicine (2004) Fact Sheets, Unified Medical Language System® (UMLS®) 2004).

National Women's Health Resource Center Inc. (2004a) Lung Cancer Overview (last updated October 4 2004a), http://healthywomen.org/content.cfm?L1=3&L2=49&L3=0&SS=0.

National Women's Health Resource Center Inc. (2004b) Lung Cancer Diagnosis (last updated September 14 2004b), http://healthywomen.org/content.cfm?L1=3&L2=49&L3=1.0&SS=0.

NCERx Inc. (2004) Lung Cancer Types (last updated 14 December 2004 2004), http://www.lung-cancer-types.com/.

Ohmann, C., Franke, C., and Yang, Q., et al (1995) Diagnostic score for acute appendicitis , *Chirurg*, **66** (2), 135-141, GERMANY.

Penedo, M.G., Carreira, M.J., Mosquera, A., and Cabello, D. (1998) Computer-aided diagnosis: a neural-network-based approach to lung nodule detection, *IEEE Trans. Medical Imaging*, **17** (6), 872-880.

Phoenix5 (2002a) Definition of differentiation (last updated June 2002 2002a), http://www.phoenix5.org/glossary/differentiation.html.

Phoenix5 (2002b) Definition of anaplasia (last updated June 2002), http://www.phoenix5.org/glossary/anaplasia.html.

Quinlan, J.R. (1993) C4.5: Programs for Machine Learning , Morgan Kaufmann, San Mateo, CA.

Quinlan, J.R. (1986) Induction of decision trees , *Machine Learning*, **1** (1), 81-106.

Quinlan, J.R. (1983) Learning efficient classification procedures and their application to chess end-games, Machine Learning: An artificial intelligence approach, (ed. Michalski, R.S., Carbonell, J.G., and Mitchell, T.M.), Tioga Publishing Company, Paolo Alto.

Richardson, D. (1973) Random growth in a tesselation, *Proc. Camb. Phil Soc.*, **74**, 515-528.

Rumelhart, D.E. and McClelland, J.L. (1986) Parallel Distributed Processing, Vol. 1: Foundations , Rumelhart, D.E; McClelland, J.L., MITPress, Cambridge, MA.

Schramm, M. and Fronhofer, B. (2003) Probabilistic aspects of score systems, *Int. J. Uncertain. Fuzziness Knowl. -Based Syst.*, **11** (Supplement), 51-73, World Scientific Publishing Co., Inc.

Shortliffe, E.H. (1993) The adolescence of AI in medicine: Will the field come to age in the '90s? , *Artificial Intelligence in Medicine*, **5** (2), 93-106.

Smith, G. (1998) Image Texture Analysis Using Zero Crossings Information, *Thesis*, Ph.D., University of Queensland, Queensland, Australia.

Szolovits, P. and Pauker, S.G. (1976) Research on a Medical Consultation System for Taking the Present Illness, Third Illinois Conference on Medical Information Systems , 299-320, Chicago: University of Illinois at Chicago.

Tufo, H.M., Bouchard, R.E., and Rubin, A.S. (1977) Problem-Oriented Approach to Practice, *I. Economic Impact. Journal of the American Medical Association*, **238** (5), 414-417.

Voorhees, H. and Poggio, T. (1988) Computing Texture Boundaries from Images, *Letters to Nature*, **333**, 364-367.

Williams, T. and Bjerknes, R. (1972) Stochastic model for abnormal clone spread through epithelial basal layer, *Nature*, **236**, 19-21.

Willis, R.A. (1967) Pathology of tumors, 4th ed, Appleton-Century-Crofts, NY.

Woolner, L.B., Fontana, R.S., and Cortese, D.A. (1984) Roentgenographically occult lung cancer: Pathologic findings and frequency of multicentricity during a ten year period, *Mayo Clin. Proc.*, **59**, 453-466.

Wu, C.M., Chen, Y.C., and Hsieh, K.S. (1992) Texture Features for Classification of Ultrasonic Liver Images, *IEEE Transactions on Medical Imaging*, **11** (2), 141-152.

Wu, J.T. and Nakamura, R.M. (1997) Human Circulating Tumour Markers: Current Concepts and Clinical Applications, American Society of Clinical Pathologists, Chicago, IL, USA.

Yoshida, H., Keserci, B., and Doi, K. (1997) Computer-Aided Diagnosis of Pulmonary Nodules in Chest Radiographs: Distinction of Nodules from False Positives based on Wavelet Snake and Artificial Neural Network, Proceedings 1st Int. Workshop on Computer-Aided Diagnosis

Zhou, Z.-H., Jiang, Y., Yang, Y.-B., and Chen, S.-F. (2002) Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles, *Artificial Intelligence in Medicine*, **24** (1), 25-36.