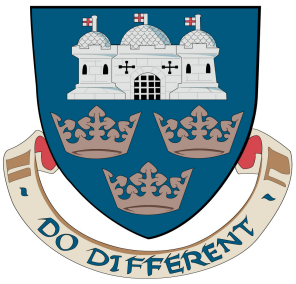
From Data to Knowledge in Secondary Health Care Databases

Joao H. Bettencourt-Silva

A thesis submitted in fulfilment
of the requirements for the degree of

Doctor of Philosophy

School of Computing Sciences
University of East Anglia



July 2014

To

Beatrice

&

Mariana ,

to my family, and *in memoriam* of those we lost.

Abstract

The advent of *big data* in health care is a topic receiving increasing attention worldwide. In the UK, over the last decade, the National Health Service (NHS) programme for Information Technology has boosted *big data* by introducing electronic infrastructures in hospitals and GP practices across the country. This ever growing amount of data promises to expand our understanding of the services, processes and research. Potential benefits include reducing costs, optimisation of services, knowledge discovery, and patient-centred predictive modelling. This thesis will explore the above by studying over ten years worth of electronic data and systems in a hospital treating over 750 thousand patients a year.

The hospital's information systems store routinely collected data, used primarily by health practitioners to support and improve patient care. This raw data is recorded on several different systems but rarely linked or analysed. This thesis explores the secondary uses of such data by undertaking two case studies, one on prostate cancer and another on stroke. The journey from data to knowledge is made in each of the studies by traversing critical steps: data retrieval, linkage, integration, preparation, mining and analysis. Throughout, novel methods and computational techniques are introduced and the value of routinely collected data is assessed. In particular, this thesis discusses in detail the methodological aspects of developing clinical data warehouses from routine heterogeneous data and it introduces methods to model, visualise and analyse the journeys that patients take through care. This work has provided lessons in hospital IT provision, integration, visualisation and analytics of complex electronic patient records and databases and has enabled the use of raw routine data for management decision making and clinical research in both case studies.

Contents

List of Tables

List of Figures

Acknowledgements

I am extremely grateful and fortunate to have been supervised by Professor Vic Rayward-Smith (School of Computing Sciences, UEA). His contributions, motivation, and wisdom have made my experience truly rewarding, both academically and personally. The ideas that shaped this project were born out of a series of meetings with Vic, and it was thanks to his guidance and support that I was able to secure funding from the UEA and the FCT (Foundation for Science and Technology, grant number SFRH/BD/43770) to undertake this research. *Gratias maximas tibi ago.*

I am also most thankful to Dr. Beatriz de la Iglesia (School of Computing Sciences, UEA), for her supervision, support and thorough advice throughout this work. Her motivation, guidance, and willingness to give her time so generously was very much appreciated especially as my PhD drew to an end. *Muchas Gracias.*

I would like to thank Professor Simon Donell (UEA and N&NUH) for his supervision, guidance, and invaluable insights, in particular at the early stages of this project. I am also grateful to have met and worked with Professor Phyo Myint (UEA, N&NUH and University of Aberdeen), whose enthusiasm, wisdom and support enabled me to widen the applicability and reproducibility of this work, and provided opportunities for further research. *Thank you.*

The expertise and guidance of Professor Colin Cooper (Prostate Cancer Genetics, UEA) was pivotal in shaping some of my work and stimulated new ideas; Dr. Jeremy Clark's (Prostate Cancer Genetics, UEA) input and contributions were also very much appreciated. Many thanks also to Mr. Robert Mills (Urology Consultant, N&NUH)

for his guidance and expertise in prostate cancer, and for his help in making sense of complex data at very crucial points in this project; to Dr. Laszlo Igali (Histopathology Consultant, N&NUH) for his advice, insights and motivation; to Dr. Trevor Tickner (Clinical Biochemistry Consultant, N&NUH); and to Dr. Hugo Baillie-Johnson (Oncology Consultant, N&NUH). I am very grateful to all for the time they have spent discussing this project with me. *Thank you all.*

I would also like to honour the memory of Professor Rick Jones (Chemical Pathology and Health Informatics, Leeds University) whom I had the pleasure to meet at conferences and to collaborate with, and who inspired much of the work on clinical data warehousing in this thesis.

The interdisciplinary nature of the research carried out in this thesis led me to meet and work with a vast number of people from the Norfolk & Norwich University Hospital, the Norwich Medical School, the School of Computing Sciences, and the School of Biological Sciences at the University of East Anglia. From hospital consultants to junior doctors, service managers to administrators, information analysts to IT staff, and fellow researchers and scientists, they all have, in many ways, contributed to this work. I am especially thankful to the outstanding staff at the Norfolk & Norwich University Hospital, the Stroke Team, Urology, Pathology, IT and Information Services. The work carried out here would not have been possible without their help and patience. *Thank you all.*

Finally, I must express my gratitude to Beatrice and to my mother Mariana, for their unconditional love, patience and support throughout these reclusive years. I shall try to refrain from mentioning "routinely collected data", "heterogeneous", "multiple sources", and "pathways" for a while, and find suitable replacements.
Merci & Obrigado.

Chapter 1

Introduction

Tempora labuntur, tacitisque senescimus annis.
Tempora mutantur, nos et mutamur in illis.

Time is ticking and we grow older through the silent years.
Times change and we change with them.

– Ovidio, Fasti, 8 AD; William Harrison, Description of England, 1577.

The last decade has seen an exponential growth in the amount of electronic data collected by public and private institutions across the world. Large databases now store records from a wide range of interactions, from everyday purchases to hospital admissions and social interactions. The primary uses of this data are largely intertwined with the core institutional activities (sales, service provision, reporting) while its secondary uses only received interest later. Organisations are now actively seeking ways to harness the statistical power of their data to optimise operations, increase productivity, reduce costs, or understand behaviour. As a result, descriptive and predictive analytical techniques have been applied across a wide range of areas with varying degrees of success. The most prominent

obstacles to the success of such techniques are not purely technological. Instead, the manner in which information is recorded, its meaning, context and format are challenges that often require an interdisciplinary approach. Institutions are now attempting to link and integrate information across distinct service-area databases and in some cases the efforts are cross-institutional. It is expected that this *big data* is able to generate new knowledge that will transform policy making, improve service provision, understand behaviour, increase productivity or further research and innovation.

This is particularly true in health care, where the untapped potential of the vast amounts of health data could unearth new knowledge that could revolutionise the way in which we practise medicine. Access to linked clinical, social and administrative information could reduce costs and optimise the delivery of health care services; Integrated electronic medical records (EMR) and clinical decision support systems can improve and change evidence-based medicine and its timeliness; knowledge discovery and predictive modeling tools and techniques have the potential to identify vulnerable patients or factors leading to poor clinical outcomes before these occur; patients could access their data, engage with the services, and explore their own and others' journeys through care. The potential for revolutionising the delivery of care by searching hidden patterns within large databases are considerable and are increasingly thought to be feasible with the advent of *big data*.

In 2012 an estimated 500 petabytes of digital health care data was spread across the globe and 25,000 petabytes are expected by 2020 [10]. Fueling this steep growth are, among others, an increasing number of patients visiting health services, the increasing burden of disease [11], but also advances in bioinformatics, genomics, computing, and in mobile health. In the early 1990s, the Human Genome Project was one of the first projects to explore the limits of available data processing technology and by 2003 a whole human DNA sequence was stored

electronically. That is 3 billion letters of genetic code. The development of computing techniques in medicine, however, dates back to the early 1960s, where the first database systems appeared. In the 1970s, the first clinical system accepting medical images for angiographies was implemented [12]. Later, clinical decision support and expert systems began to appear (MYCIN [13], Internist-I [14]). In the last two decades, health systems and providers have been making a slow yet steady progress in adopting computational tools and techniques and digitising medical records.

In the UK, over the last decade, the National Health Service (NHS) programme for Information Technology (NPfIT) has further boosted *big data* by introducing electronic infrastructures in hospitals and GP practices across the country. In addition, the growing prevalence of chronic disease is adding pressure to health services and multiplying data [11]. Over the last 5 years, the English NHS alone saw a 13% increase in hospital admissions to a total of 15.1 million admissions in 2012-13 [15]. If each admission accounts for only 1 kilobyte of information that would be 15.1 gigabytes of new admissions data just last year, then multiplied by imaging, referrals and any other procedures and tests. This ever growing amount of data, most of it resulting from routine service activities, promises to introduce novel ways of exploring and understanding the services and their processes, and to boost research and data mining activities. However, underlying technical and social challenges still need to be addressed before biomedical data can have its full influence on health care [16].

Several government projects and initiatives both nationally and internationally have focused on *top-down* approaches to improve and innovate the ways in which health data is recorded, organised and used [17; 18; 19]. Conversely, evidence-based *bottom-up* approaches revealing current practices and data are scarce. In an increasingly digital world, such approaches are now becoming feasible and could provide additional evidence for making decisions about the ways in which

data and systems are implemented and organised.

Despite national and local initiatives to organise health data and the adoption of standards, medical data continues to be infamous for its complexity and heterogeneity [20; 21] and hospitals arguably provide one of the most challenging environments for the development and application of computational techniques. Indeed, state-of-the-art data or process mining algorithms often need to be adapted to cope with the uniqueness of medical data. The research field of health and biomedical informatics has dealt with some of these issues and pioneered solutions. In fact, health informatics research has boosted other research in computing sciences and introduced new tools and techniques with wider applicability (for example, the Entity-Attribute-Value data model [22]). However, persistent issues and challenges still remain [16]. This thesis will address them by exploring the journey from raw data to knowledge in a large English NHS hospital with a catchment area of up to 822,500 people. Further introductory notes as well as the research objectives, contributions and the thesis outline are given in the next sections.

1.1 A Consilience of Disciplines

Health and biomedical informatics is arguably the most comprehensive term encompassing the core and applied research that deals with the "optimal use of information, often aided by the use of technology to improve individual health, health care, public health, and biomedical research"[4]. However, this discipline can also be defined as biomedical, medical or health informatics and the "adjective problem" [4; 23] is ongoing. Nevertheless, one of the first definitions was coined by Robert Greenes and Edward Shortliffe in 1990: "Medical informatics is the field that concerns itself with the cognitive, information processing, and communication tasks of medical practice, education, and research, including the

information science and the technology to support these tasks"[24]. In 1995, Perry Miller defined medical informatics as a field that focuses on the creative use of computers in support of patient care, medical education and biomedical research [25]. In the early 2000s, Enrico Coiera gave a more encompassing definition of health informatics [26] as the rational study of "the way we think about patients, and the way that treatments are defined, selected and evolved"; "how we organise ourselves to create and run health care organisations"; and "how clinical knowledge is created, shaped, shared and applied".

Emerging as a distinct academic entity, health informatics brings together clinicians, scientists and engineers to find solutions for a wide range of problems pertaining, in most cases, to the optimal use of information and knowledge in biomedicine and health. The field is, therefore, broad and several subcategories have been proposed. William Hersh has recently published a diagram that illustrates the major subcategories of the field [4] (Figure 1.1).

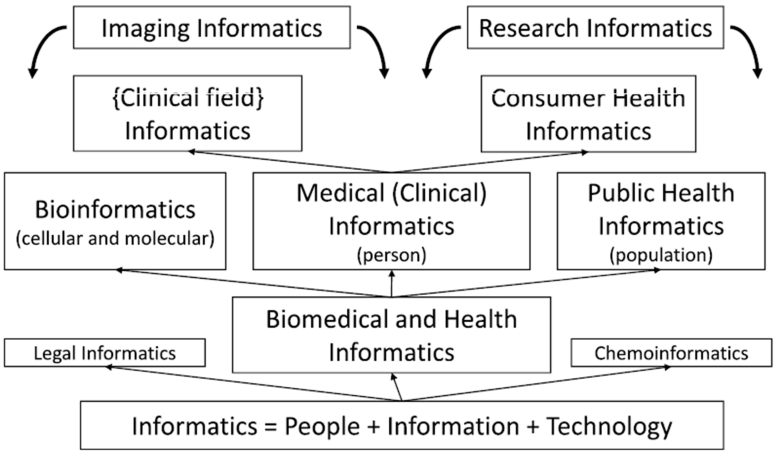


Figure 1.1: Major subcategories of the informatics field, from [4].

The activities across the subcategories may share similar goals with other established sciences such as epidemiology, that aim to study patterns, causes, and effects of health and disease in populations. Indeed, epidemiological methods and tools have been implemented in several health informatics projects such as hospital infection control software [27; 28; 29]. However, despite authors stressing the benefits of interdisciplinary work, there is still scant evidence of a strong collaboration between such sciences in the literature. As E.O. Wilson envisaged in [30], a *consilience* of disciplines providing the synthesis and linkage of knowledge from different fields of human endeavour is key to future advances in research. It is crucial, therefore, that the sciences be able to understand each other and be open to collaborative work in a way that medical informatics, a *consilience* between medicine and computing, has already demonstrated with some success.

One of the early efforts to describe the science in medical informatics was that of Charles Friedman. In 1995, Friedman introduced the *tower of achievement* as one way of representing the creative work in this field [31]. Friedman's tower is widely accepted scientifically and it is comprised of four levels [31]:

- At the bottom of the tower lies the formulation of models for acquisition, representation, processing, display, or transmission of biomedical information or knowledge;
- Then comes the development of innovative computer-based systems, and using these models, that deliver information or knowledge to health care providers;
- Next, the challenge is to install such systems and make them work reliably in functioning health care environments;
- Finally, at the apex of the tower, lies the study of the effects of these systems on the reasoning and behavior of health care providers, as well as on the organization and delivery of health care.

Friedman's tower is exceptionally accurate, in particular when developing clinical decision support or expert systems, and it demonstrates how different actors working in medical informatics need to work together and contribute equally in health care settings. Recent technological advances, however, have resulted in a rising interest in applications and solutions that expand beyond the realm of health care providers. There is now a growing interest in linking information available in health services with other organisations, biological databases, or even individuals through mobile devices and social media. This would lead to an involvement of social scientists so as to help illuminate the complexity of the interactions between users and systems or to play a "crucial role in the discovery of the biases that are intrinsic to digital data" and "in the construction of convincing stories about what those data reveal" [32].

This, however, is only possible if complete and accurate data elements are collected in a consistent and canonical form. Despite years of health informatics research, the unique nature of medical data and systems has, among other less technical factors, continued to hinder substantial progress. This is particularly true in hospitals and it is investigated in this thesis.

1.2 Origins and Nature of Information in Medicine

The organisation of medical knowledge is among the oldest applications of classification, dating to Aristotle's efforts in biology and formal descriptions [33]. The subsequent history shows increasingly detailed classifications of causes of death in the 16th century London Bills of Mortality and recently, the World Health Organisation International Classification of Diseases (ICD), the SNOMED Clinical Terms, and even the Health Level-7 (HL7) for data interchange. The origins of data in medicine rely on classification systems, but it is the collection of data points, the practice of medicine and the continuous communication between prac-

titioners that generate most of the data and allow comprehensive studies to be carried out. The benefits of collecting information on disease have been thoroughly demonstrated in epidemiology. An early example is that of 1854's cholera outbreak in Broad Street, London, where John Snow relied on a collection of geographical and other information on mortality to trace the origin of the outbreak. Since then, epidemiology has continued to play a crucial role in informing medicine and public health.

To a general practitioner or a hospital clinician, collecting information from patient observations is essential to making a diagnosis and deciding on treatment. Such information comprises patient symptoms and complaints, histopathology and biochemistry findings, medical imaging techniques, and even genomics information. Medical *datum*, therefore, derives from multiple sources and comes in a variety of ways: numerical measurements, recorded signals, images, narrative text, genes and discrete data. This data is stored across the services' databases and used primarily to support the delivery of care. In some cases, however, the data is heavily bound to commissioning or billing activities which tend to improve its quality, but neglect other potentially relevant clinical information.

Information is also generated by the need of clinicians to communicate with one another, and some of the properties of the communications media may not account for such interactions [34]. However, it has been noted that, for example, computerising the communication of laboratory requests from, and reports to, clinicians leads to improved efficiency throughout the cycle of reporting and requesting [34], and further work is required to make use of new technology in laboratories. Hospital information systems (HIS) can play an important role in achieving greater efficiency and improved communication.

In 1984, Peter Reichertz gave a lecture on "Hospital Information Systems past, present and future" [35]. At the time, hospital systems had already been in use for over a decade. Reichertz foresaw most of the core developments in HISs (pic-

ture archiving, data processing and transmission, and linkage of expert systems to databases) and indicated what the goals and tasks of information systems should be. The most important goal of a hospital information system is to "map the real environment into a formal representation, through the bottleneck and restriction of data acquisition" [35]. HIS would, therefore, encompass the following functionalities: acquisition of information, processing of information, and presentation of information.

Beyond the necessary system management functions of HISs, Reichertz also foresaw additional aspects. Such as problem solving tasks that would require the development of expert systems using artificial intelligence and statistical methods [35]. Reichertz also anticipated that further architectural changes would be needed as HIS become health information systems (regional or global rather than local). In the UK, the latter has been, to some extent, a reality as most NHS hospitals currently submit data for national audits or for overall commissioning or planning and performance management. However, data that is not required by clinical audits or other programmes is unlikely to be consistently collected or its quality assessed. Indeed, the information available in HIS, its completeness and quality, are not well understood and researchers and clinicians can have unrealistic expectations about the extent in which such data can be re-used. Quality data available in such systems, however, has the potential to be re-used for clinical research although it is often difficult to access.

In hospitals, the autonomy of each department and clinical speciality introduce additional challenges and heterogeneity to the systems and the quality of the data [21]. Having distinct data owners, targets and objectives, hospital departments behave in a similar fashion to different organisations and, albeit often sharing an internal patient identification number, they provide an extremely challenging environment for the integration and analysis of linked health data. The retrieval and uses of routinely collected hospital data is a central topic in this thesis as are

the transformations needed for this voluminous raw data to become knowledge.

Figure 1.2 shows the *knowledge pyramid*, an enlightening illustration of the perspective on the volume and value of data. The top levels of the pyramid see the most valuable knowledge and decisions, and at the lowest level lies complex heterogeneous data [20; 21]. A recent definition of medical knowledge is given by Paul Taylor [18] as "the principles or heuristics that we abstract from experience and use to guide future action". In turn, the process of finding useful information and patterns in data is called *knowledge discovery in databases* and data mining is the use of algorithms to extract such information and patterns from databases [36]. The latter have been characterised as "highly heterogeneous with respect to the data models they employ, the data schemas they specify, the query languages they support, and the terminologies they recognize" [21].

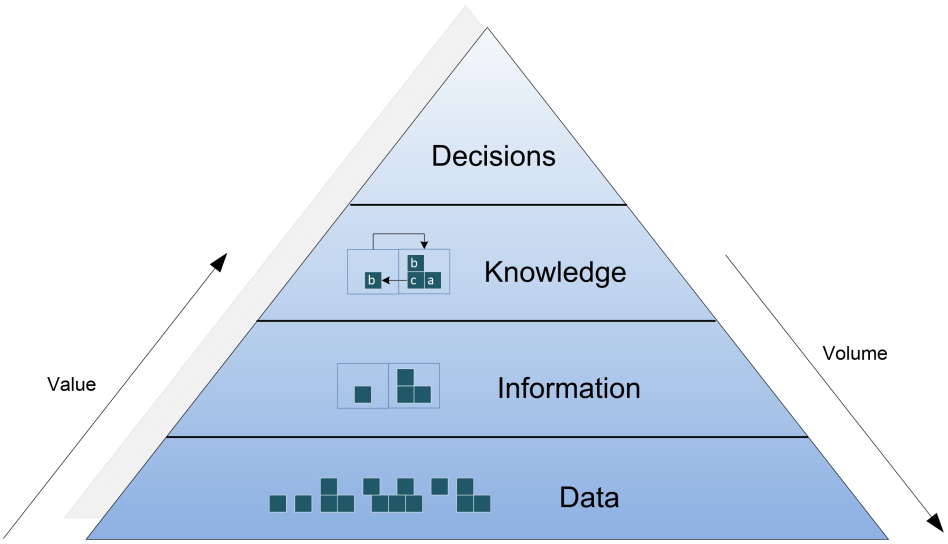


Figure 1.2: The knowledge pyramid.

In addition to the above, there are issues with data quality across five dimensions (completeness, correctness, concordance, plausibility, and currency) and,

currently, there is little consistency or generalisability in the methods used to assess them [37]. Assessments of the quality of the data often expose several types of errors and in some cases these may not just be typographical or missing information. In 1999, *To Err is Human* was one of the first high impact reports that described patient safety and the extent of human errors in medicine [38]. The report revealed that between 2% to 4% of all deaths in the United States were being caused by medical errors. In England and Wales, medical error (or adverse events) accounted for 2.2% of all episodes in 2004 [39]. While progress has been made in this respect [40], errors are part of the nature of medical data, and the role of computerised order entry and decision support systems in preventing them is still precarious [41].

The complexity and heterogeneity of medical data is present throughout its own life cycle: from when it is first created with its initial (or lack of) structure, through all the changes needed for it to be communicated effectively, and up to the point when it is forgotten in hospital databases or libraries. This thesis will explore the journey of retrieving, cleaning, linking and analysing the digital footprints left behind by patients as they visit (and re-visit) hospital.

1.3 Objectives and Problem Domain

There are great expectations about the potentials of *big data* stored routinely in hospitals but the extent to which this data can be used and studied is largely unknown. The primary objective of this thesis is to study the journey from routine data to knowledge in a secondary health care centre, and to investigate the extent of the secondary uses of this data. The journey from data to knowledge is made by traversing critical steps, each addressing specific secondary objectives:

- Investigate and develop methods to extract routinely collected data from multiple heterogeneous data sources and to store this data in integrated repositories;
- Explore and introduce existing and novel techniques to preprocess routinely collected data;
- Study and develop data modeling techniques suitable for complex, heterogeneous and patient-centric data so as to model patients' journeys through care;
- Develop methods, tools and techniques to model, integrate, visualise and analyse such data.

The setting used in this research is the Norfolk & Norwich University Hospital (NNUH) NHS Foundation Trust, a large academic teaching hospital in England with a catchment area of up to 822,500 people. The time period selected in this study is over 10 years, from 2001 to 2011. This time period was selected to capture any changes in data and in systems over time.

Routine clinical data is recorded across several hospital systems but is rarely analysed or linked. Upon linkage and integration, it is expected that important and useful information can be obtained and benefit patients, their doctors and health service planners. Methodologies, data mining and other computational techniques are tested and developed in this thesis, and the importance of clean integrated health records is demonstrated using two clinical case studies: prostate cancer and stroke.

Prostate Cancer

Prostate Cancer is the most common male cancer in the UK [42] and one that varies in behaviour between individuals and over time; it may be an incidental finding which never causes any problems or it may change to a progressive and lethal cancer, which can be treated to some extent with various degrees of success. A major problem is when it escapes from control with hormonal drugs and metastasises to bone and, at present, there is no reliable way of predicting which cancers will behave aggressively and which will not. With regards to mortality, about 90% of men diagnosed with localised prostate cancer (confined within the prostate) live for more than 5 years after diagnosis and between 65-90% live for at least 10 years. This contrasts with locally advanced cancer (spreading outside the prostate capsule) where 70-80% live for at least 5 years after diagnosis, and with advanced cancers (metastatic) where only 30% survive 5 years [42].

In recent years, there has been a generalised increase in reported incidence but, despite this, the mortality rates have been on the decline [42; 43; 44]. Nevertheless, the economic burden of prostate cancer will continue to rise due to increased diagnosis, diagnosis at an earlier stage and prolonged survival [44]. It has been reported that new strategies need to be devised to improve the efficiency of health care provision for this type of cancer in order to tackle the increasing burden [44]. Prostate Specific Antigen (PSA), a biochemical marker used clinically for prostate cancer detection and prognosis, is associated with substantial overdiagnosis and excessive treatment [45], which makes its utility as a screening test controversial, and warrants the need for further studies. Further background information on prostate cancer is provided in Appendix B.

The case study on prostate cancer looks at the utility of routinely collected hospital data in understanding the patients' journeys through care. This case study is of particular relevance given the chronic nature of cancer and the propensity of patients to use multiple hospital resources (radiology, histology, radiotherapy

and theatre, among others) over longer periods of time. This requires data points to be collected over large time windows and across multiple systems, which, in turn, provides a sound environment for observing data quality. In addition, this case study has also carried significant clinical interests and gathered a team of experts including a urology consultant, prostate cancer geneticists, a consultant oncologist, a histopathologist and a chemical pathologist. Even though the initial aims of this study were to explore clinical findings, the bureaucratic, complex and time-consuming nature of the work leading up to the analyses hindered this goal. Nevertheless, work on clinical studies is ongoing and descriptive and summary statistics are included in this thesis.

Stroke

Stroke is the third most common cause of death and the first cause of long-term disability in the UK [46]. Every year in the UK more than 150,000 people suffer a stroke and the costs of treatment and care for the NHS was of 8 billion pounds in 2010 (approximately 5% of the total UK NHS costs [46]). Stroke incidence rises steeply with age and despite better preventative measures, the total number of strokes will continue to rise in the UK [47]. Furthermore, stroke care in the UK is far from ideal. Patients have worst outcomes when compared to other European countries [48] and the situation has become widely known following the recent publication of the Royal College of Physicians' National Stroke Audit [49].

Stroke is often regarded as an acute and episodic disease in secondary care yet it leads to long-term disability and its causes are longstanding. The literature now emphasises the need for prevention and a reorientation of practice so that stroke can be considered a chronic disease with acute events [50; 51]. Indeed, appropriate prevention can lower the risk factors leading to stroke events [51]. From a data perspective in hospitals, stroke contrasts sharply with cancer due to its acute and episodic nature. As a result, it is expected that a greater number of clinical parameters can be found in secondary care databases. This is particularly true

during a patient's stay which lasts, on average, 16 days across all stroke subtypes in England [15].

In addition, the Norfolk & Norwich University Hospital has been collating a in-house stroke register database since 1998. This database, however, is disconnected from hospital systems, its data is input by data entry clerks, and its platform and role have changed over time. Appropriate data models will be explored, and this thesis will also investigate methodologies and techniques to enable this database to be linked to other hospital systems and used for further clinical research. Furthermore, clinical studies using data mining and other statistical and epidemiological methods were undertaken during this research and others are still ongoing, and not included in detail in this thesis. A full list of contributions is given in the next section.

1.4 Summary of Contributions

In this thesis, the journey from data to knowledge in a secondary health care centre is explored in detail and novel computational methods for the retrieval, linkage, integration, modeling visualisation and analysis of routinely collected data are introduced. This has not been previously researched in full, and attention has only been given to segments of the journey. The contributions of this work are summarised below:

- A novel methodology for extracting patient-centric data from multiple heterogeneous data sources was developed and published in a leading health informatics journal [52]. This methodology was developed using the prostate cancer case study, tested with the stroke study and work is ongoing for it to be applied to other hospitals. The methodology also resulted in the development of the largest and most comprehensive repository of prostate

cancers in Norfolk.

- Over 10 years of Hospital Information Systems and their data from a large NHS hospital were studied. This was explored in parallel with the above methodology and provides a comprehensive list of challenges that needed to be addressed for the systems' data to be integrated.

- Tools and techniques for the linkage and preprocessing of routinely collected data were introduced, including an algorithm for extracting relevant text from histopathology reports, and a novel imputation method for continuous values. These techniques were applied extensively to the routinely collected data, contributed to ensuring its quality, and have wider applicability.

- Methodological steps were proposed to develop clinical data warehouses with routine hospital data from HISs. This work led to the development of the new Norfolk & Norwich Research Stroke & TIA Register, a unique integrated repository of strokes and transient ischaemic attacks (TIA) with over 25 thousand patients from 1998 to present. Both the stroke register and the prostate cancer repository were developed to enable high quality cohorts of routinely collected data to be selected for clinical, epidemiological and health services research. A publication in a health informatics journal is expected (Appendix E).

- The concept of a data-driven patient-centric pathway was introduced to describe the journey that patients take through care based on the available electronic data. A pathways data model was proposed and tested, and the feasibility of current data and process mining techniques to analyse the pathways was explored. A research paper has been submitted to a leading health informatics journal (Appendix E).

- A new clinical decision support system was developed based on the prostate cancer pathways. A novel graphical representation was introduced along

with software that aggregates similar sequential pathways and compares the outcomes of given patients' profiles. Both the pathways and the software have a wider applicability in other domains and an application to stroke is planned.

- Clinical studies were undertaken while others are still ongoing. These are not described in detail in this thesis yet they are included in the list of research publications resulting from the work carried out in this thesis (Appendix E).

1.5 Thesis Outline

Chapter 2 focuses on the problem of multi-source data collection from multiple heterogeneous sources and summarises over 10 years of hospital information systems; this is the first step of the journey from data to knowledge, in which routine data is retrieved from multiple hospital information systems. Chapter 3 covers data preprocessing, record linkage, and data warehousing; these are techniques required for quality assessment and data cleansing before any analyses can be undertaken. Chapter 4 deals with the modeling of prostate cancer pathways and the tools and techniques to analyse them; these include algorithms and the development of clinical decision support and other software. Each chapter critically evaluates its own topic and concluding remarks are given in Chapter 5. Not all chapters refer to the stroke and prostate cancer case studies, and this is summarised in Figure 1.3, along with relevant information pertaining to each chapter.

	Chapter 2	Chapter 3	Chapter 4
	Multi-Source Data Collection	**Preprocessing, Linkage and Data Warehousing**	**Pathways Modelling, Mining and Visualisation**
Case Study	Prostate Cancer Stroke	Prostate Cancer Stroke	Prostate Cancer
Study Focus	Hospital Information Systems and their data; Data Retrieval and development of study specific data bases	Tools and techniques for Data linkage, preparation and Quality improvement; Clinical Data Warehousing	Data Modelling, Data-driven pathways, Analysis and Mining Techniques
Tools & Techniques	Novel Methodology for Multi-Source Data collection	Data Quality and Preprocessing Techniques; Data Warehousing Methodology	Novel Data Model for Pathways; Novel Framework for Integration and Visualisation; Assessment of Mining Techniques; Data Quality; Decision Support
Infrastructure	Prostate Cancer Data Repository; Norfolk & Norwich Stroke & TIA Register		Prostate Cancer Visualisation and Integration Software; Decision Support tools
Additional Material	Appendix A Information Systems and Data Collection Details	Appendix B Prostate Cancer Background	Appendix C Additional Data on Pathways and Mining; Appendix D Summary of Prostate Cancer Trends

Figure 1.3: Thesis Layout.

Chapter 2

Multi-Source Data Collection

This chapter explores in detail the cumbersome process of accessing and retrieving patient-centric data from multiple hospital information systems (HIS) for the creation of a data repository for future data analyses.

Section 2.1 introduces the problem of multi-source data collection from secondary care centres; section 2.2 provides the background and preliminary work carried out for the development of a methodological process for data collection.

The process is explained in section 2.3 whilst section 2.4 describes the development of a semi integrated data repository from the collected data. The process roadmap was developed using the case study on prostate cancer, and validated by a second study, on stroke (section 2.5).

Discussion and evaluation of the data collection approach are given in section 2.6.

2.1 Introduction

The use of electronic medical records (EMR) in clinical research has already been envisaged by the medical informatics community [53; 54; 55]. Indeed, different types of EMR systems can be used to select and, in some cases, retrieve data for epidemiological studies, clinical trials or other clinical study data management systems [56]. In longitudinal or other observational studies, however, legacy patient information may be collected from regional or national data sources such as registries. The latter is often federated with data from primary, secondary and tertiary care centres which are, in turn, required to submit the datasets and respective mandatory fields. However, such regional or national approaches generally target a rather specific dataset with limited parameters and uses. This reductionist approach to data collection may not fulfil the needs of other, more complex, retrospective studies requiring more detailed clinical parameters.

In the National Health Service (NHS) in the UK, primary care specialists work in the community and refer patients to secondary or tertiary care centres (hospitals and specialised treatment centres respectively). Perhaps because of their organisational structure, secondary and tertiary care centres generally implement Hospital Information Systems (HIS). The latter were initially designed to support and monitor patient care, specific medical tasks and hospital management [35]. Only recently has attention been given to clinical research [57] and other secondary uses of such patient data [19].

Historically, HIS developed from central to modular and distributed systems [58]. Central and modular systems perform better in homogeneous environments, but this poorly reflects the reality of medical information in hospitals, which is heterogeneous and complex [20; 58]. Similarly, the distribution of information processing and uniqueness of medical data [20] pose obstacles to data integration [58]. Commercially available HIS often focus on administrative tasks and lack

knowledge-based functionality [59]. According to Sujansky, hospitals may opt for implementing several commercial departmental systems, creating islands of information across various departments [21], which further hinders the retrieval of patient-centric data. This problem is augmented by a lack of semantic interoperability (i.e. clinical coding), particularly in outpatient events in the NHS [60]. Nevertheless, in-house projects such as the Oxford Clinical Intranet [61] have been able to link and centralise patient information from multiple HIS and other sources. Other in-house projects have also achieved integrated centralised systems [62] but they rely on strong organizational support, funding, resources, and a software design strategy focused on the caregivers information needs [62]. Some of these systems may be used to extract cohorts of patient-centric data, however, in most cases, such functionalities are not supported, particularly in commercially available systems. Apart from cases where there is strong organisational support and a forward-thinking strategy, most hospitals will have environments that pose similar challenges to those identified by Sujansky [21]. Furthermore, it is expected that future projects carrying out inter-organisational data linkage exercises will face similar challenges, and hospitals with multiple heterogeneous systems provide a sound environment to study how patient or person-centered data can be collected and linked.

Collecting patient-centric data and EMRs from one or multiple HIS in a secondary care centre is often reported to be an *ad hoc* process, and it is poorly described in the literature. Indeed the description of methods for data collection has been mentioned as a key point for the improvement of the quality of the literature [63].

A methodology for data collection was constructed based upon expert consultation and driven by a case study on prostate cancer. Previous work on knowledge discovery in databases (KDD) methodologies [64; 65], KDD methodologies tailored to medical domains [5], and work on clinical data integration [6], were important in shaping the methodology which is described in detail in section 2.3.

2.2 Background and Preliminary Work

The methodology for collecting patient-centric data from multiple HIS was drafted by initial consultation with domain experts as well as input from relevant segments of other KDD methodologies. Because collecting data from multiple sources ultimately results in data integration, input from previous work on this topic [6] was also crucial. The forthcoming sections will address such topics in detail.

2.2.1 Knowledge Discovery Methodologies

In medical data mining, large data-sets of prepared data allow the discovery of patterns in sets of records. For these patterns and rules to be successfully derived and contextualized, a knowledge discovery methodology (or process model) is essential. KDD methodologies have been developed to provide guidance to data mining and other data analysis and knowledge discovery projects using large databases. According to Cios [66], the goal of designing a DMKD methodology is to come up with a set of processing steps to be followed by practitioners when executing their DMKD projects [66]. The benefits of such process can be measured in terms of development time, reliability, efficiency and overall cost [5].

Methodologies reviewed include Fayyads 9-step model [67], Debuse et al. KDD Roadmap [64], Cabena's 5-step model [68], the Cross Industry Standard Process for Data Mining (CRISP-DM [65]), and Cios 6-step data mining and knowledge discovery [5]. The latter has been applied to medical data mining exercises and is the most appropriate to be explored in detail in this section. Nevertheless a comparison between the abovementioned methodologies is given later in this section.

2.2.1.1 The 6-Step DMKD

The 6-step data mining and knowledge discovery (6-step DMKD) is a modified version of CRISP-DM (used primarily in business contexts) and has been successfully applied to medical problem domains [69; 70; 71]. The methodology, shown in figure 2.1, comprises the following steps:

Understanding the medical problem domain
This is an essential step where a list of well defined outcomes is put together. Indeed a key goal of this step is to translate medical goals into data mining goals. However, the latter can only be achieved after a previous understanding of the problem domain and hence a background survey of the domain is often necessary. This step should also identify problem requirements, restrictions and determination of success [5] and involvement from the domain experts is expected.

The following tasks are defined within this step:

- Define the problem
- Determine medical objectives
- Identify key people
- Learn about current solutions to the problem

Understanding the data
Data should be retrieved and a preliminary analysis is performed in this step. It may be necessary to obtain ethics and governance credentials from the institutional review board and/or local, regional or national ethics committees to grant access to the data and/or data sources. The data may also need to be anonymized, de-indentified or encrypted [20] and summarisation (described below) may be of help in achieving this.

An understanding of the data, its usefulness, along with their semantics and syntax, is expected in this step. This leads to analyses of completeness, relevance, missing values, attribute values, extremes, among others.

Data preparation

The 6-step DMKD states that preparation of the data is a key step of the entire DMKD project [5]. Indeed this step has been previously acknowledged as the hardest step in a KDD process [72]. In this step, decisions must be made regarding which data to input in the data mining algorithms and significance and correlation tests will be performed together with sampling of the database [5]. Furthermore, this step will also account for the cleansing of the data, which includes, for example, handling of missing values or noise.

Once clean, the data can be further processed by feature selection and extraction algorithms (to reduce dimensionality), by derivation of new attributes (such as by discretisation), and by summarisation of data (granularisation) [66].

Overall, it is important for the information to be accessible from a computational point of view not only because medical data is heterogeneous in nature [20], but also because many data mining methods cannot cope with voluminous or malformed data.

The results of such changes may be new attributes, values or records that should meet the specific input requirements for the data mining algorithms [66]. Training and test sets can be used to make sure that data mining techniques work and present interesting results. Indeed training sets are often used to develop the specific parameters required by the techniques [36].

Data mining

In this step, data mining techniques are applied to the data. The techniques may have been previously selected in the preceding steps, however, other algo-

rithms can be applied if necessary. There is a wide range of techniques including: Bayesian methods, neural networks, clustering, machine learning, decision trees, among others. The scalability of DM techniques is important because some may not work with large volumes of data.

Evaluation of the discovered knowledge

The results obtained from previous step are interpreted here. This step is often performed along with a domain expert that should confirm the importance and utility of the results. It may be necessary to go back to step one (Understanding the problem domain) or to step four (Data mining) of the DMKD process. This recursive feature of the 6-step DMKD process allows the identification of alternative actions that could improve the results.

Using the discovered knowledge

The involvement of domain experts is also essential in this step. In fact, this is the step where database owners or problem domain experts decide how to use the discovered knowledge. The discovered knowledge, when interesting, may result in research publications of novel clinical findings [72] or changes and optimisation of the involved services or systems, among other possible outcomes. A plan to monitor the implementation of the discovered knowledge is also created here, and the entire project documented [66].

Figure 2.1 shows the 6-Step DMKD methodology and the relations between steps. Regarding time spent on each step, data preparation is the most time/effort consuming step (45%), followed by data mining (15%), understanding the data (15%), and understanding the problem domain (15%). Data mining may, however, may require more effort than the other two steps that equally stand at 15%.

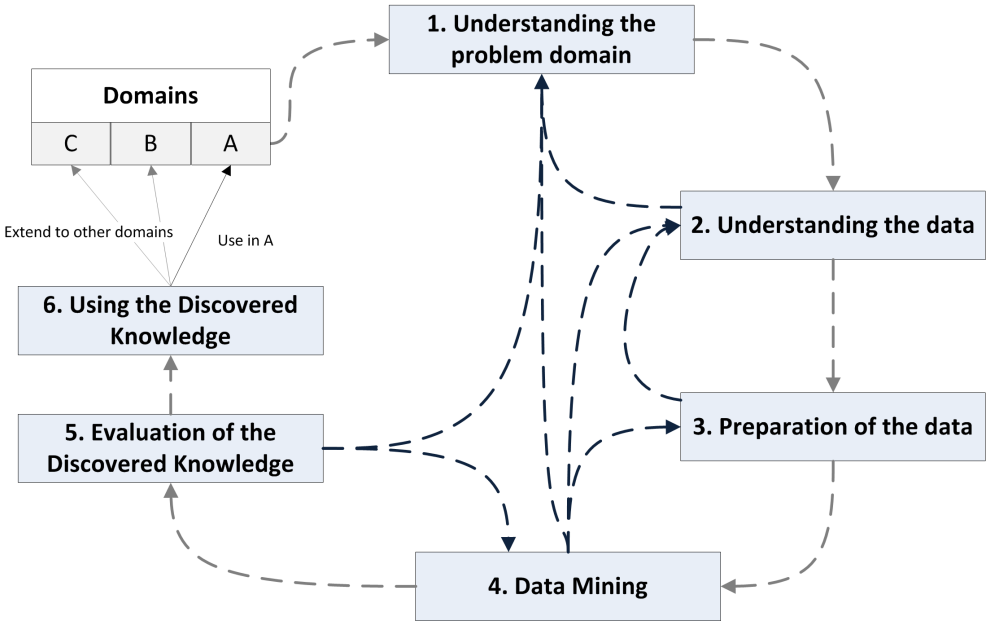


Figure 2.1: Cios 6-Step Data Mining and Knowledge Discovery Methodology adapted from [5].

2.2.1.2 Comparing Methodologies

Five DMKD methodologies were reviewed and the 6-step was selected as an example because it is tailored to medical domains [69; 70; 71]. A summary of the steps in each methodology reviewed is given in Table 2.1. Transversely, the overarching common steps across methodologies are:

- **Domain understanding**

  The data understanding step is present in almost all methodologies reviewed and is part of domain understanding or data preparation. It is an essential step as semantics and syntactical differences are often reported, particularly in medical domains [73]. Some methodologies [5; 65], identify this step on its own, whilst the Cabena's 5-step methodology [68] misses it entirely as it assumes this as previous knowledge.

- **Data preparation**

  Data preparation, also be referred to as data cleansing or preprocessing, is present in all methodologies and it can be included as a step of its own or broken down into several smaller steps [67].

- **Data mining**

  Data mining, also referred to as modelling or data analysis, is defined as a single step across all methodologies.

- **Evaluation of the discovered knowledge**

  The process of evaluating the discovered knowledge is, in all methodologies, broken down into an analysis, evaluation or interpretation phase and deployment or application of the knowledge.

Cabenas 5-step methodology is too broad and it may not be ideal for medical domains unless it is used by physicians within a health care organisation, in a situation where the data and the problem are already well understood. In fact, the incompleteness of Cabenas process model has already been reported by Hirji [74]. Conversely, Fayyads 9-step model [67] is complete and thorough but it does provide less freedom to the data miners. For instance, data miners may want to reduce dimensionality and project the data before cleaning it, and not the opposite.

It would also be expected that methodologies would account for the process of seeking credentials with the health care organisations or the retrieval and potential linkage of the study data, which is not the case on any of the reviewed methodologies.

Overall, the most detailed methodologies may be more useful and efficient to less experienced data miners, or in environments where little is known about the problem domain and its data. Simpler methodologies such as Cabena's 5-step

may be more efficient to those who are experienced and where knowledge of the domain is well understood.

A well balanced methodology is the 6-step DMKD which provides enough freedom to the data miners, is based on the widely accepted and used, CRISP-DM [65], and has been successfully applied to several medical domains. Another important feature of this methodology is the additional recursion between steps (illustrated in figure 2.1).

With regards to the application of the above methodologies, and using the Elsevier citation index, Fayyad's 9-step is by far the most widely cited in academic papers [75]. Other methodologies have been substantially less cited (approximately 40 times less than Fayyad's) in equal measure. A reason for this difference lies in the fact that Fayyad's paper describing the methodology also introduced the most commonly used definitions of DM and KDD [75]. This reveals that a search of citation indexes, even if thoroughly carried out, may provide insufficient accuracy in determining the number of applications of the methodologies.

However, two polls carried out by KDnuggets (`http://www.kdnuggets.com`) in 2002 and 2004 revealed that the majority of respondents use the CRISP-DM methodology or their own methodology. Applications in medical domains have been reported in all methodologies, however, the 6-step has been used almost exclusively in this area [75].

The overall common steps in all methodologies are illustrated in Table 2.1 and provide the framework to embrace the data collection exercise presented in detail this chapter (section 2.3).

The analysis of the main methodologies guiding knowledge discovery or data analysis projects depicts that data collection (retrieval or extraction) has never been mentioned as a step in its own right. Indeed it is often assumed the data is

Step	Cios 6-Step [5]	Fayyad 9-Step [67]	CRISP-DM [65]	Cabena 5-Step [68]	KDD Roadmap [64]
Domain Understanding	Understanding the problem domain	Understanding domain and identifying DMKD goals	Business understanding	Business objective determination	Problem specification
	Understanding the data	Creating target data set	Data understanding	Data preparation	Resourcing
Data Preparation	Preparation of the data	Data cleansing and preprocessing	Data preparation		Data cleansing
		Data reduction and projection			preprocessing
		Matching goal to a particular data mining method			
		Exploratory analysis and model hypothesis selection			
Data Mining or Analyses	Data mining	Data mining	Modelling	Data mining	Data mining
Knowledge Evaluation & Application	Evaluation of the discovered knowledge	Interpreting mined patterns	Evaluation	Analysis of results	Evaluation
	Using the discovered knowledge	Consolidating discovered knowledge	Deployment	Knowledge assimilation	Interpretation
					Exploitation

Table 2.1: Comparing steps across DMKD methodologies.

readily available, or made available by the domain experts. Researchers carrying out projects in secondary care centres using multiple HIS are faced with the additional workload of retrieving and collating data, which are non-trivial tasks.

The data collection methodology presented in section 2.3 can be seen as an expansion of the data preparation step, or ultimately a new step to come between domain understanding and data preparation. Although some emphasis is given to the neighbouring steps, this work focuses primarily on the method for data collection since this has received the least attention in the literature.

2.2.2 Data Integration

Data collection and preparation in KDD methodologies in medical environments may be achieved by following an information pipeline architecture described in [6]. The pipeline, as illustrated in figure 2.2, channels data from their sources to an operational data store (ODS) where the information is stored before it is further validated, cleaned and merged during the Extract-Transform-Load (ETL) process [6] to create a core database. The functionality of an ETL process can be summarized in the following tasks:

- **Extraction**: the identification of relevant information at the source side and the extraction of this information;
- **Transformation**: the customisation and integration of the information from multiple sources into a common format and the cleansing of the resulting data set;
- **Loading**: the propagation of the data to the data warehouse and/or data marts, or in this case the propagation to the operational data store.

The ETL process is frequently used in building data warehouses and will be discussed again later in chapter 3. The core database implements a data model and is used to draw more specific data marts for reporting, visualisation, analysis and data mining [6].

The method for data collection presented focuses firstly on the journey of extracting data from a single source, and secondly, on how the retrieved data is used to build an ODS. The process can be repeated to include multiple information sources. Some of the issues of working with multiple sources have been addressed in [6; 21; 76; 77]. Furthermore, global query systems have been developed [76; 77]

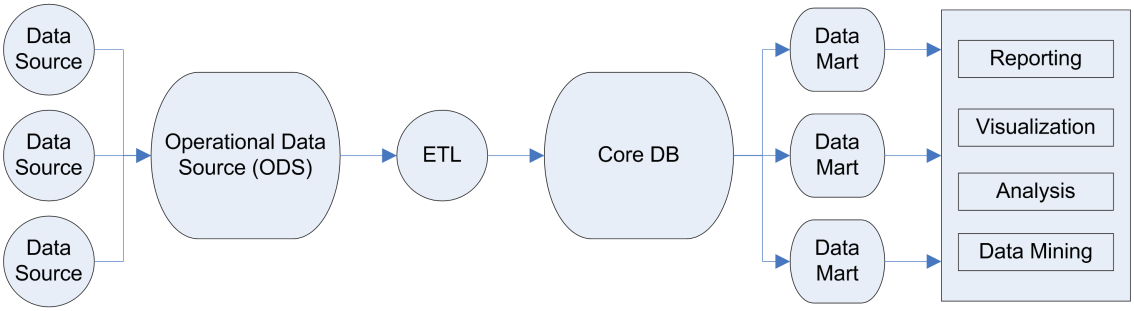


Figure 2.2: Information pipeline architecture, adapted from [6].

to facilitate the retrieval of information from multiple databases in distributed enterprises. However, such issues are often approached from a database power-user point-of-view, where a privileged access to the underlying database systems exists together with a high level of technical proficiency. Ethical issues, data-ownership, credentials and hospitals' organisational structures may pose obstacles to gaining access to systems and patient-centric data.

Nevertheless the use of automatic and intelligent tools and techniques to link and extract data in the next chapter.

Knowledge of the different types of heterogeneities in distributed database systems [77] is key for setting important goals when working in such environments. These types include:

- Technical differences in database systems
  - Structure, where different data models provide different structural primitives;
  - Constraints, where different data models may implement different referential integrity rules;
  - Query languages, where retrieval and manipulation of query results may differ.

- Semantic Heterogeneity

  This type of heterogeneity pertains to the meaning, interpretation, or intended use of the same related data across systems [77].

- Autonomy

  Different organisational entities and their independent control over different systems are the cause of this type of heterogeneity. In particular, autonomy involves the design, communication and execution of database systems components.

Indeed to address some of the above issues the concept of metadatabases evolved from the traditional data dictionaries [76; 78]. In traditional data dictionaries, information about the data (i.e. metadata) is gathered to facilitate software development and the control of data sharing among different programs and systems as well as supporting the lifecycle of a database system [78]. Metadatabases were developed to tackle higher levels of management and control of information models (metadata), and to include information on process models and business rules [78]. This is particularly important since medical information is essentially bound to the context of its production [55]. Indeed, one of the outputs of the data collection process described here is a metadatabase comprising of metadata files for each data source.

One of the benefits of using a metadatabase in the context of the work carried out is to aid the critical review and evaluation of not only the data and their sources but also the future quality of any studies that can be derived from the collected data. It has been suggested [56] that this evaluation should include the protocols, record layout and codes, data entry instructions, published material, analyses, technical reports, and the carrying out of appropriate completeness and validity studies, all with respect to the specific context of the study [56].

The two sections below define the problem specification (section 2.2.3) and explain the initial process of selecting and identifying the data sources (section 2.2.4. These two sections may be regarded generically as domain understanding. It is, however, in section 2.3, that the approach for data collection is thoroughly described. Data cleansing and integration are discussed in section 2.4.

2.2.3 Problem Specification and Ethical Approval

The criterion used includes patients diagnosed with prostate cancer (ICD C61.X). The aim of this chapter is to use the inclusion criteria and present a methodological approach to data collection from multiple sources in a hospital. A research protocol comprising this information was agreed with the research team and credentials were sought and approved by the local research ethics committee and research governance committee.

This time consuming process involves ethical and research governance approvals of a research protocol. Patient sensitive data such as names and addresses are excluded, only key patient identifiers are used for linkage. The research protocol included an extensive literature survey on the problem domain, and the application for credentials also required a list of information systems (data sources) to be used in the study; however, this was not fully known *a priori* and subsequent amendments to the original documents were necessary.

2.2.4 Identifying Data Sources

In order to provide an accurate and holistic view of each patient, the aim was to collect data from as many authoritative sources as possible, where information on a patient with prostate cancer exists. When so much data is collected and analysed, less biased results are expected as a better understanding of comor-

bidity factors and the pathway of a patient is possible. Nevertheless it has been argued that requiring data with no practical importance creates a source of misleading information [6]. Efforts need to be put towards the definition of practical importance or relevance of data required for the study.

The study on prostate cancer involves a wide range of data from several departments and in order to accurately link this data across systems and departments, a comprehensive EMR of a patient is needed. Data elements for the sole purpose of validating or linking data need to be included as well as other fields or datasets that may have an impact on the understanding of the problem domain. At a first stage, the primary concern was to gather a list of relevant and authoritative data sources containing relevant data to the study. Data elements for each data source will be identified later.

A conceptual data model (CDM) may also be of use in the identification of the variables for the study and those that are outside of it [6] but this requires previous knowledge of the sources and data elements. The approach presented here assumes little or no previous knowledge of the systems, database schemata, or any other metadata. The cohort definition and an authoritative list of data sources define the boundaries of the study data. The process of identifying the sources relied on the cohort definition and the work carried out in section 2 as well as input from domain experts and hospital staff.

Ranking relevant, authoritative databases is a research topic in modern information retrieval and is dependent on a given query input. Methods have been developed to work with text databases, which involve building inverted indices of terms for each database and using similarity measures between the databases and a user query. This would be an interesting approach involving the formulation of a query that captures the study population. However, it does rely on preprocessing that cannot be achieved without full access to the underlying database systems.

The aim is to work with local hospital data and, for this reason, national or regional data sources are disregarded for the most part within the proposed methodological approach. Nevertheless, this is later discussed as a way of validating the study data.

Data Source (Information System)	Abbreviation	Department
Picture Archiving Communication System	PACS	Radiology
Patient Administration System	PAS	Administration
Oncology System	ONC	Oncology
Biochemistry & Histopathology	LAB	Histopathology
Operating Theatre	OPT	Operating Theatres
Radiotherapy	RAD	Radiotherapy Physics
In-hospital Cancer Registry System	CRE	Information Services
Local Cancer Registry (external)	CR	East Anglia Cancer Registry
Orthopaedics System	ORT	Orthopaedics

Table 2.2: Data Sources (Information Systems) identified and the respective department.

At an initial stage, gathering information on clinical coding was important so that cohorts of target patients can be retrieved first. This is important because not all sources provide accurate clinical coding or another way of identifying the target population. As described in table 2.2, out of the data sources used in this project, only one contains homogeneous clinical coding for diagnoses and was used as the primary source for data collection on prostate cancers. Alternatively the local cancer register (CR) could have been used to select patient data, however, at this stage we are interested in the sources within the hospital. The use of the local cancer register is later discussed in this thesis.

2.2.4.1 Selected Data Sources

The Picture Archiving Communication System (PACS) is a system dedicated to radiological imaging. In this case study, only textual information (medical reports) present in this system needed be retrieved. The hospital Patient Administration System (PAS) contains important appointment information for all patients as well as clinical coding (ICD for diagnoses and OPCS for procedures) for inpatients. The Oncology department system (ONC) is an in-house developed system which relies primarily on textual reports. Diagnoses, treatments and history are available in this system. The biochemistry & histopathology system (LAB) includes important biochemical information such as the prostatic specific antigen (PSA) and histopathology reports with respective tumour markers.

The operating theatre (OPT) system will contain information on procedures and the radiotherapy system (RAD) system will include radiotherapy data such as the number of fractions (sessions) and the body site being treated. The orthopaedics system (ORT) contains information on any pathological fractures from advanced metastatic prostate cancer.

The in-hospital cancer registry system (CRE) was introduced in 2007 to aid the management of cancer pathways for which targets have been set by the government. This data was previously uploaded to the National Cancer Waiting Times Database (NCWTDB) by a lengthy process (comparable to an ETL) and the purpose of the new system introduced in 2007 was to facilitate this as well as produce further information for planning and performance as well as clinicians. This system was not used in this project due to the fact that it would not have good quality data until 2010. However, the use of the data on NCWTDB is explored later.

A detailed description of the hospital information systems used is available in Appendix A.

In the UK, another resource that may help the initial selection of some data elements necessary for the study, without previous access to the systems, is the NHS data model and dictionary. The data model and dictionary were developed as an attempt to create a reference point for assured information standards throughout NHS services [79]. Although this service focuses primarily on defining minimum data sets for standardising health management data such as commissioning, audit and performance measuring, other attributes such as PSA levels are now available. This data model was solely used for understanding data elements from particular national datasets such as the NCWTDB.

2.3 Data Collection

Each of these steps will be described in the following sections and, when applicable, a summary table will present the relevant information gathered from the case study at each step. The overall process is repeated for each data source identified and new data sources may be identified or excluded from the initial list. A travel log of the journeys should be kept. The order in which the identified sources are put through the process is important as not all systems will necessarily provide a way of identifying the target patients (as arose in the case study).

The process in Figure 2.3 accepts a data source (HIS, database system or equivalent) as input and returns metadata and a study dataset as outputs. In order to reach the outputs, four major stages are followed: system understanding (described in detail in Section 2.3.1), data understanding (Section 2.3.2), extraction preparation (Section 2.3.3) and extraction & evaluation (2.3.4).

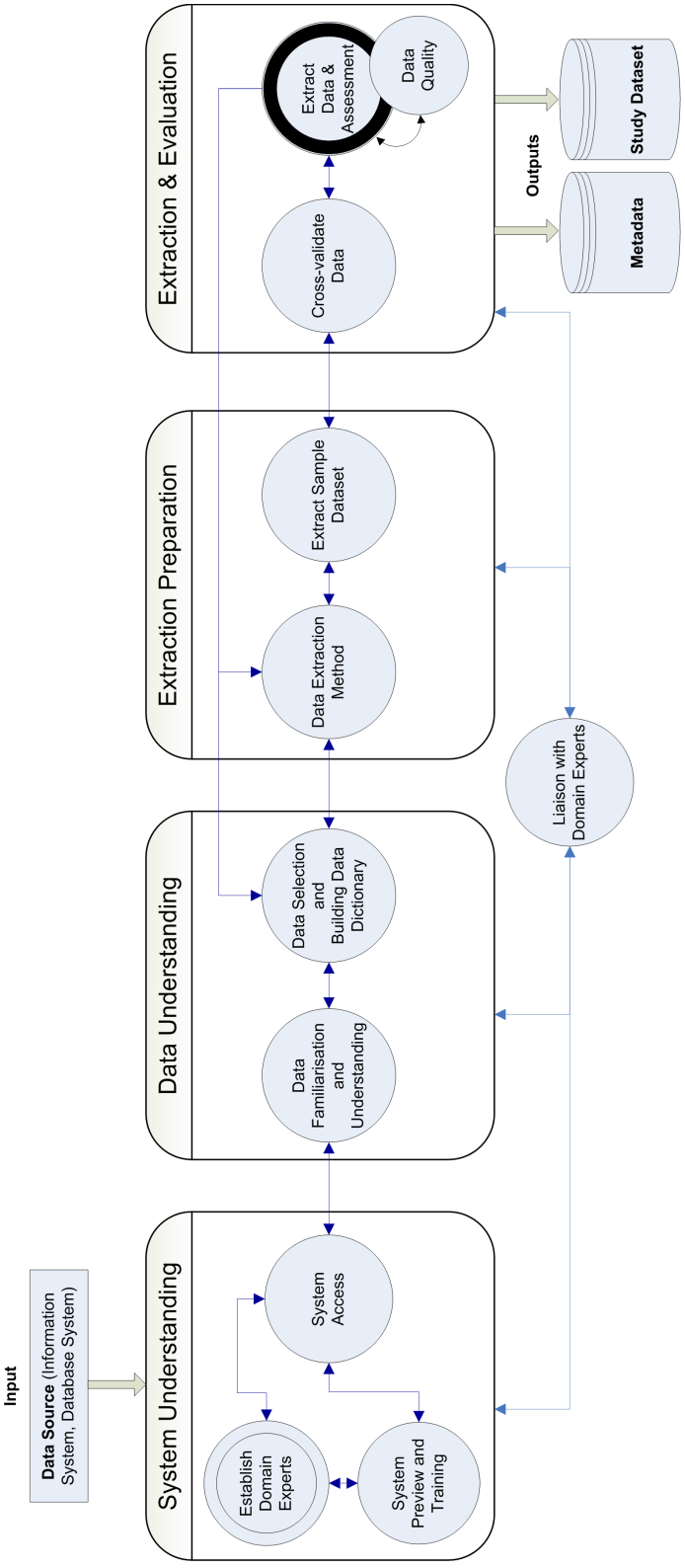


Figure 2.3: Process to extract data from a single source.

2.3.1 System Understanding

2.3.1.1 Establish Domain Experts

The process described above begins with liaising with the domain experts (system managers or experienced system users such as data analysts). Ultimately this allows further assessment as to whether the system does indeed contain relevant data for the study. System experts may overlook the data needs of the research study. Further investigation can be sought by previewing the system. Indeed, when available, system training and preview can speed up the process of understanding the system. When a system is found unsuitable or redundant, the list of identified sources in section 2.2.4 is revisited and the process in figure 2.3 ended. Domain experts may also be able to point out suitable data extraction methods, details on how to obtain access to the system, and information on when the system was implemented.

In this step, important information to be obtained from the system expert is a data flow chart detailing user actions and system interactions. A context level diagram is often sufficient to understanding how information flows to and from a system. Such information is later important when analysing data, particularly to understand patients pathways. The investigation of whether a system feeds data to others is an important part of this task as it can potentially make one of the systems redundant. During this study, it was found that most data fed across systems are demographics coming from the PAS but this was not implemented on all systems. When this was implemented, however, this minimised the risk of double entry as well as improving future record linkage. The frequency of data updates to systems is also important, as well as relevant data changes that may occur on a patient's record after its retrieval.

2.3.1.2 System Preview and Training

Liaising with experts may not be sufficient to determine whether a system is appropriate and previewing the system should help to highlight any relevant data. In some cases, a system preview, often seen as a walkthrough using the graphical user interface (GUI), can be part of a hospitals system training. Indeed, when available, system training is an advantage that should be exploited. Depending on the quality and depth of the training, the information gathered can facilitate understanding of the system, building a data flow diagram, understanding the data (step 2.3.2.1 and 2.3.2.2) and obtaining system access (step 2.3.1.3).

With respect to the case study, formal hospital training was undertaken on PAS, PACS and LAB. The first was a day of foundation training and exercises and a day of training on a data extraction tool (an On-line Analytical Processing (OLAP) system), the second a one-to-one session with the domain expert and the third was an online interactive course. At the end of each training session, system access credentials were provided. Training is not always available or appropriate for researchers. For example, there are PAS training modules tailored to particular administration tasks, some irrelevant to the scope of this research. There may, however, be training in expert data extraction tools that are normally provided to IT staff and information analysts. As described above, this was the case for PAS and, although limited to extracting and linking administration data, it allowed an initial understanding of the data extraction tool and the data elements. There was no formal training for the ONC system but the system was previewed during the first meeting with the domain experts.

2.3.1.3 System Access

This is a crucial step as without system access data collection can not be achieved. Obtaining credentials to access the system (read-only accounts) may require previous system training (covered in step 2.3.1.2 above). In the study, this was the case for three systems: PAS, PACS and LAB. By reaching this step, sufficient information about the system will have been gathered and a summary of the metadata pertaining to the three initial steps of the process can now be produced. Table 2.3 describes the summary metadata for all four information systems in the case study. The metadata produced during the above steps pertains to the systems specifications. At this stage the metadata is derived from expert consultation only. Further, more technical, details are gathered throughout the remaining steps of the process.

Metadata - Understanding the System	Examples from the case study
Appropriateness and authoritativeness of system	All identified systems are authoritative and contain crucial information on prostate cancers.
Data flow chart	Overall understanding of hospital information systems data flow.
Data extraction method (if known)	PACS - No particular method identified, users suggest manual extraction to spreadsheets; PAS - Business Intelligence software; ONC - Database back-end, liaise with manager; LAB - Business Intelligence software.
Preview and system training	Previewed all systems, training on PAS, PACS,LAB.
System credentials	Read-only access to the systems was issued.
Key data elements present (if known)	Hospital number, NHS number, Date of Birth present on all systems.
System Updates (impact)	All systems were live and updates occurred daily, but no impact on study data.
System Limitations	The LAB system was introduced in 2003 and there is no electronically available data prior to this date. The PACS system went live in October 2001 and backdated reports were uploaded to the system in a slightly different format (visible on the GUI).
System Live Date	Date when system went live.
Last System Update Date	Date when system was last updated.

Table 2.3: Summary of items gathered when understanding the system.

2.3.2 Data Understanding

2.3.2.1 Data Familiarisation and Understanding

The identification and ranking of the relevant data fields from the source is essential [66] and can first be achieved by familiarisation using the systems GUI together with any data extraction tools when available. A set of examples to achieve database familiarisation has been described in [64] and, from those, the following were considered important:

- Database field type determination (categorical or numerical)
  Numerical types include discrete and continuous. Categorical types include ordinal (with an implied order) and nominal (no implied order). As pointed out in [64], in some cases, categorical fields may include numerical values that should be treated as such, and this is important metadata to add to a data dictionary. As data types may differ in software packages, it is also important to deal with this accordingly, making sure that the original data is not altered in any way, which can affect correlations. Perhaps a common mistake is to assume different software packages, database systems, or other information retrieval systems handle decimal places in the same way.

- Determination of database field semantics
  Two or more fields, although different, may be based on the same or similar measurements [64]. An understanding of the data fields is therefore important and detailed explanations should be included in the data dictionary. It is also important to expand any abbreviated field names.

- Determination of field value semantics and their plausibility [5; 66]
  Knowledge about the meaning of the field value is essential to understanding the data; it can be used to spot outliers and erroneous values, and in some cases, to handle missing values. This is important as it should be considered

whether missing information means that exposure or outcome has not taken place or whether it is indeed a missing value [56] and, if so, whether it is possible for it to be recovered or guessed.

- Reliability
  Field reliability impacts data quality. Although at this stage there may be no collected data to accurately assess field quality, it is still important to manually check for data completeness and determine, at an initial stage, whether the field is reliable or not based on an inspection of the field throughout different patient records.

Because full data integration can only be achieved by overlapping focal data elements [6], these need to be sought for each data source. This may be difficult to achieve at the very first iteration of the process and relies on identifying the fields that allow linkage. The most common are hospital number, and other ubiquitous patient details throughout systems such as date of birth and contact details. Input from domain experts greatly facilitates this. When the data collection process runs at least two complete iterations, hence evaluating more than one system, it is possible to compare the focal data elements across sources metadata.

As previously mentioned, if there is no preceding list of target patients (as with the case study of patients suffering from prostate cancers) it is important that these can be identified in the system. This is achieved by identifying clinical coding fields or textual fields where it is possible to retrieve clinical diagnoses. When clinical coding is stored in a consistent yet textual data field (such as a clinical report), text mining and natural language processing techniques may be used to retrieve the correct information. In this case study, the ONC source contains full text reports, but a single one line text field indicates the primary site (primary diagnosis).Therefore, a way to retrieve prostate cancers in this oncology system was to run a database query for where the word prostate appears at least once

in the primary site field. This method was accurate, mainly because the data source has a restricted domain, but also because of the various ways in which the diagnoses are written do not abbreviate the word prostate. Occasionally, tumour grading or PSA markers are also included in this data field, but are inconsistent throughout records and further text mining on adjacent text fields would have to be carried out to retrieve such information consistently. The PAS system, however, uses ICD-10 and procedure (OPCS) codes but only for inpatients. The two other systems (LAB, PACS) do not provide consistent diagnosis information throughout patient records.

Ranking is a process that relies on selecting data elements which have a predictive value in assigning a correct identity to an individual. This is essential to ensure linkage is possible at a later stage, once a data repository comprising all datasets has been compiled. The most common data elements are the basic demographic details (patient number, age or date of birth), event dates, event types and coding.

Rank	PAS	PACS	ONC	LAB
1	Hospital and NHS Number	Hospital and NHS Number	Hospital and NHS Number	Hospital and NHS Number
2	Date of Birth	Date of Birth	Date of Birth	Date of Birth
3	Episode Start Date	Study Date	Date Registered	Date of Entry
4	Diagnosis Code 1	Procedure Code	Primary Site	Test Code
5	Diagnoses Codes	Modality	Diagnosis	Test Data

Table 2.4: Data fields ranking from the different data sources in the case study.

An example from the case study is given in Table 2.4 in which the most authoritative data elements from each data source have been identified and ranked. The uniqueness of the value corresponds to high authoritativeness and thus a high rank. Data elements such as date of birth or age at episode are not considered to be unique but ensure a more accurate linkage when used together with a patient identifier. During the case study, and due to patient confidentiality restrictions, only patient identification numbers and date of birth were used as the most authoritative data elements. The record date (episode, diagnosis, procedure or other) was also important as it allowed selecting the date range of the cohort

and linkage validation of certain events. Deterministic record linkage can be used at a later stage to link particular datasets, after the latter have been compiled into the operational data store (ODS).

2.3.2.2 Data Selection and Building a Data Dictionary

This is a key step where metadata is compiled into a table, and, once all metadata from all data sources is collected, a final metadatabase is created (Section 2.4). This is an iterative step, revisited when the finalised dataset is extracted from a data source. At a first instance, items to include in the dictionary include the focal data elements identified in the previous step. As the process continues, further data items may be added, removed or updated from the dictionary. The typical metadata generated for each field of the data sources in the case study is illustrated in Table 2.5.

Metadata - Data Understanding	**Description**
Original Field Name	Name of the data field (original name from source)
Field ID	Primary key for identifying a particular data field in a metadatabase
Field Description	A detailed description of the field
Data Type (source)	Data type of the field as it appears in the data source
Data Type (converted)	Data type to use in final dataset (may be the same as source data type)
Field Size (source)	Size of the field in characters from the source (when applicable)
Field Size (converted)	Size of the field to include in final dataset (may be the same as source field size)
Expected Outliers (if any)	Any outliers or expected erroneous field values should to be detailed
Field Rank (Importance)	The rank number of this field in relation to its data source
Data Linkage	Often useful to use as a Boolean indicating whether data linkage is possible with this field

Table 2.5: Data fields ranking from the different data sources in the case study.

2.3.3 Extraction Preparation

Preparing a strategy for extraction relies on liaising with experts at the hospital (system administrators, analysts, clinicians) but also, when possible, from the software vendors. There are a number of ways in which information can be extracted from a single source. Perhaps the most common way to query a database for information is by writing and running Structured Query Language (SQL) queries. This may be considered a back-end approach where significant credentials are needed, and, in some systems, a query builder GUI may be available and embedded in the system itself. It may also be worth investigating whether OLAP tools exist for the particular data source as this can act as a more complex query builder, perhaps allowing data from more than just one source to be retrieved. Ultimately, the approaches depend on the access level to a system. In some cases, and indeed in the case study presented in this paper, software programs needed to be developed [73] to extract information from the PACS (further details in Table 2.6). Any software developed to work at a hospital site may need approval by the information technology department, and the process can be time consuming. Table 2.6 illustrates the extraction methods chosen for each data source in the case study: an OLAP tool for PAS and LAB, a back-end SQL query run by the domain expert for the ONC system, and the software developed during the case study to extract data from PACS. As part of the extraction preparation step, sample datasets were also retrieved in order to validate the extraction methods. This was particularly useful to validate the extraction method with the OLAP tool in the case study.

Data Source	Extraction method used in the case study
PAS	OLAP software tool available was used to retrieve the data elements in this source. Training and liaison with experts was needed in order to build queries.
PACS	Because of access restrictions, a software program was developed to copy textual imaging reports from a PACS GUI to a canonical, spreadsheet format (special attention is required to ensure data formats are not lost in this process). The program needs to be assisted by a user, and hence, it is a time consuming method of retrieving data. Later in the project the Radiology System (RIS) was able to pull partial results from PACS and hence this was the preferred method.
ONC	The domain expert agreed to run a back-end SQL query to search for the cohort.
LAB	The same OLAP software as above was used but using a different schema to access a different data source and this required different access credentials.

Table 2.6: Extraction methods for each data source identified in the case study.

2.3.4 Extraction and Evaluation

2.3.4.1 Cross-Validation

Using the method selected in step 2.3.3, a sample dataset is extracted and cross-validated this means that the sample data (or a random part of it) is manually compared to the system data using its interface. This clerical review process ensures that the data coming from the system is indeed as expected. It may also be possible to select a smaller cohort and compare it with patient notes or with independent reference sources (e.g. registries) in which the whole or part of the target population is registered [56]. This can be investigated further in the data quality step but, in the current step, the major concern is to ensure the data retrieval method is extracting the desired data. Query optimisation methods in relational databases [80] may also be useful to increase the performance of certain lengthy operations, especially when working with OLAP systems, but they are hard to implement and may require further expert consultation.

From the case study, it was evident that the retrieval methods often produced an output in the form of comma or tab separated value, and that this may be a source of error. When working with textual data fields, the comma separated value creates a source of erroneous vertical field separation because several text reports include commas as part of their natural language. In some cases, tab separation was appropriate, but in others, rare symbols were used to effectively separate the data fields (e.g. the dollar symbol was found to be a rare symbol in text reports in the case study).

2.3.4.2 Extract Finalised Dataset and Data Quality

Upon a satisfactory cross-validation, the final dataset is extracted using the method selected in 2.3.3. However, the dataset should still be carefully examined for missing or erroneous values, and simple statistical measures produced to provide a basic understanding of the nature of the fields and the data [64]. Indeed, such statistics may suggest that data cleansing is necessary [64] and this is often covered in the next steps of DMKD methodologies (data cleansing and preprocessing). Nevertheless, it seemed important to include a brief step for quality assurance checking at the end of the methodology presented in this paper, also because this may contribute to the completeness of a metadatabase. It is important that the domain experts provide help in this process.

One of the concerns of researchers and epidemiologists working with secondary data is its completeness. Indeed, inaccurate or missing data tend to bias associations towards the null hypothesis [56]. The most relevant data cleansing operations in the course of the case study presented in this chapter were:

- Outlier handling

  Many outliers may be classified as either errors or groups of interest [64] and, at this stage, only erroneous outliers are dealt with as the first should

be addressed by the research study analyses. Expert knowledge may be necessary to determine outliers and the corrective action to apply to each form of outlier [64]. Outliers in numerical data fields may be typically found by plotting the distribution of the field values, or by sorting the dataset.

- Missing data handling
  The missing values may be removed, estimated, or simply marked as missing. Estimation or imputation, however, may need to be performed in a separate (training) dataset so that the impact can be assessed before the data is changed. A simple approach [64] is to replace missing numerical values in a field with the mean over all known examples. It may also be possible to apply rough set theory to reduce the amounts of missing data [81], and more complex, and potentially dangerous approaches, may involve training a neural network to predict missing values using the remaining fields [80]. The next chapter will further discuss the use of data editing and imputation techniques.

When possible, a quantification of the amount of missing and erroneous data in a dataset is important metadata and should be added to the metadatabase. Further methods to evaluate other data quality dimensions is given later in this thesis.

The case study on prostate cancer showed the collected data is of acceptable quality when it comes to missing values (focal data elements had less than 5% missing data), especially on the administrative systems and other systems that are often inspected by the quality team, and from which their data is used for general planning and performance, or part of national data requirements. The case study showed that, for example, the amount of missing data and outliers from two highly ranked fields from the LAB source, hospital number and test data, are 3.22% and 1.66%, respectively, on a total number of records exceeding

320 thousand. It is common, however, to find greater amounts of erroneous and missing data in legacy systems collected by clinicians [5]. It is also common to find non-focal data elements missing, which will not necessarily impact linkage, but may influence the study results.

Further statistics and detail on the quality of the collected data are given in the next chapter and throughout this thesis.

2.4 From Data Pool to Integrated Repository

One of the two outputs of the process described in Figure 2.3 is a metadata table, for each data source. A metadatabase, based on the information collected, is a useful resource to help data integration from multiple heterogeneous sources. This is particularly helpful in building a database schema for the repository based on the collected metadata. An example of such metadatabase from the case study is depicted in Figure 2.4, below. In order to build a data repository or research study database, each dataset should be imported into a password-protected operational data store (ODS). The ODS schema for the case study was built to resemble the organisation of the hospital information systems, where most tables represent a source (HIS) and were built using the collected metadata. Exceptions to this were very large datasets representing sets of biochemistry tests. In these cases a table per test was the preferred choice to maximise database query performance. Database normalisation rules were helpful in the design of the ODS, which is a database that employs the relational model. Despite re-formatting data elements where the extraction method corrupted the data type, no other data transformations were carried out to build the ODS. This was to facilitate further data to be added to the ODS in a longitudinal fashion as well as mapping the host environment. The ODS created for the purpose of the case study will therefore contain missing values, outliers or any other erroneous data as they

appear in their original data sources. The main purpose of an ODS is to act as a data pool from which researchers can query for their study needs. Indeed when following a knowledge discovery process such as the 6-step DMKD, an ODS can be used as the main study database or one from which to derive cohorts from. It is also expected that most of data understanding already occurred as part of the process presented in this chapter. Nevertheless when a particular study dataset is retrieved from the ODS, it still needs to go through a data preparation step as described by knowledge discovery or data analysis methodologies [5; 6; 64]. One of the challenges in retrieving datasets from the ODS is the matching or linkage of records. It is often assumed that deterministic matching (i.e. overlapping of focal data elements) is the most effective technique used on datasets from within an organisation such as a hospital, mainly due to the use of a patient hospital number throughout sources. In contrast, probabilistic matching, which weighs a number of identifiers to ensure pairs of records are from the same individual, is often used to compare records between organisations, where no common identifier necessarily exists between them. Examples from the case study in prostate cancer reveal that there may be erroneous matches due to double entry of a patient or due to other, clerical, inaccuracies, when only simple deterministic methods are used. Record linkage was therefore performed using rules based deterministic linkage of patient hospital number together with date of birth and episode date (these are the top three ranked data fields for linkage). A metadatabase allows a careful and informed choice for linkage, and should also include matching weights, when computed. It may also be required, at this stage, or earlier, for some of the data to be deidentified or encrypted. During the case study patient sensitive information such as names and addresses was never collected. When exporting datasets from the ODS, the database internal identifier was used to replace patient number. In order to carry out appropriate clerical review and other spot checking of preliminarily linked data, a data visualisation tool for the ODS was developed. The latter is helpful to provide a patient-centric view of the records

collected; evidence on how well the records were being merged; an assessment of the correctness of the queries developed; and for communication purposes with the domain experts.

There are also automatic approaches to database schema matching that can be useful for heterogeneous data integration [82]. Schema matching is possible using the metadata collected for each system available in the metadatabase but was not used during the case study.

The integrated repository can be seen as a data warehouse and indeed some of the work carried out to develop the repository is similar to that of data warehousing techniques (such as the ETL process). Nevertheless we will explore data warehousing as well as data linkage techniques in more detail in the next chapter, as the present chapter is primarily concerned with multi-source retrieval.

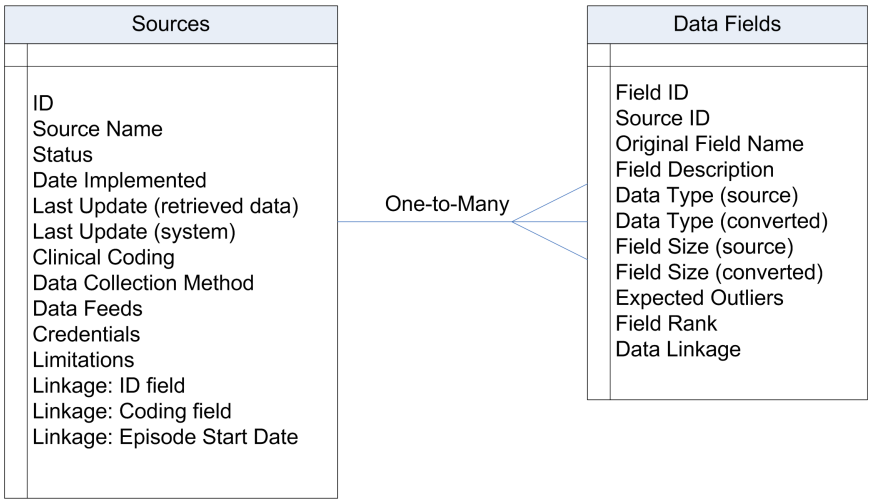


Figure 2.4: Simplified schema for a metadatabase containing multiple sources and their respective data fields. *Sources* table contains the list of sources retrieved together with their metadata, and the *data fields table* contains the attributes collected for each data source, with respective metadata.

2.5 Validating the Methodology: Stroke Study

The methodology presented in section 2.3 was later tested and evaluated using a second case study on stroke. The steps taken to collect the required stroke data and create a study data set are given in section 2.5.2 along with a summary of the key findings.

This exercise is a validation of the methodology, and resulted in the development of the Norwich Research Stroke Register at the Norfolk and Norwich University Hospital (NNUH). Further details of the techniques used to cleanse, link and integrate this data are given in the next chapter.

2.5.1 Background and Setting

Summary

The local stroke centre at the NNUH admits approximately 900 stroke patients per year. For every admission, specialist stroke nurses or doctors complete a stroke register form, which is then manually entered on the stroke register database by the stroke data team. The stroke database was developed and improved over time by the local stroke team. It stores data elements that are required at a national level for audits, planning and performance, as well as particular research studies. This database, partially due to its high volume of patients, has recently been gaining attention from clinical researchers who are keen on using it retrospectively.

History and Further Details

The Norfolk & Norwich Stroke Register, initially set up in 1996, has been described as routine collection of stroke services data [83]. It includes patients admitted to the hospital with a diagnosis or suspected stroke made known to the

stroke team. The clinical data in the register is routinely collected as a part of the local service development which enables health professionals and managers to evaluate services and conduct audits [83].

The register was initially set up as part of the European Basic Stroke Register and has been maintained by the stroke services team at the Norfolk and Norwich University Hospital.

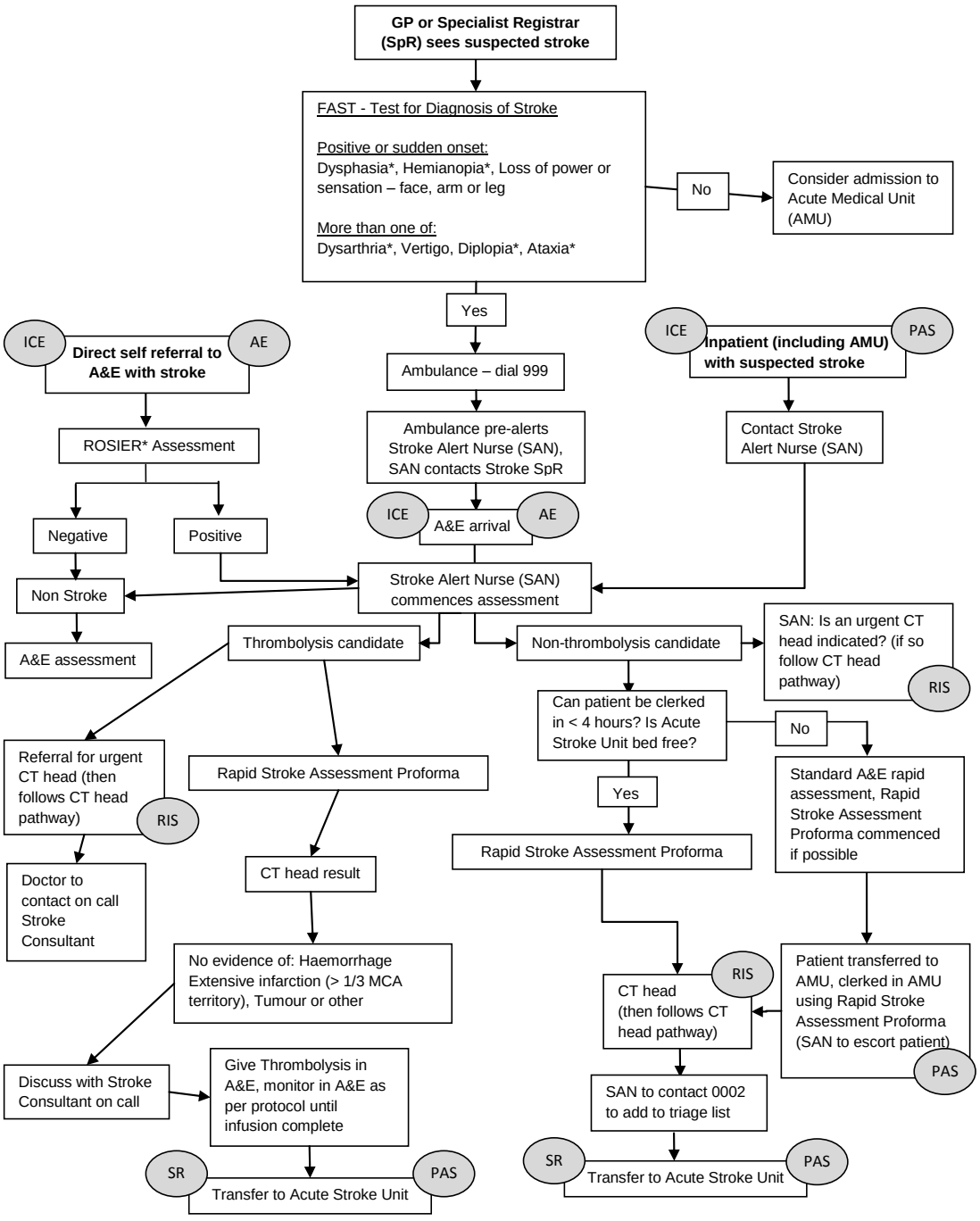
The stroke services team includes data collectors who manually enter the data from paper based forms onto the register as a part of a routine service level agreement. This data is then be used to produce statistics for management and it is also used for research [83].

The collected data elements changed over time as per service monitoring requirements and Appendix A shows the initial and currently collected attributes.

Stroke cases admitted to hospital are prospectively identified, however, the data may be collected both prospectively as well as retrospectively.

All clinical team members contribute to data collection. Data that can be obtained from paper and electronic records is reviewed by the data team and double checked with clinical team members for accuracy [83]. Data is collected until the patient is discharged from the acute hospital or dies as in-patient. No data after discharge is currently collected.

Figure 2.5 shows the pathway for the management of acute stroke in adults. This pathway has been adapted from the original [84] , published by the stroke clinical team at the NNUH and it used here for the purpose of understanding the patient flow and hence, the data. The points of entry of patient data into information systems are marked with greyed circles. A complete and up-to-date pathway describing the management of acute stroke is available in [84].



Glossary
Dysphasia – partial or complete impairment of speech.
Hemianopia - decreased vision or blindness takes place in half the visual field.
Dysarthria – abnormal or difficult articulating caused by problems with the muscles used in speech.
Diplopia – commonly referred to as double vision.
Ataxia – lack of voluntary coordination of muscle movements.
ROSIER – Recognition of stroke in the emergency room test used to establish the diagnosis of stroke
Thrombolysis – the breakdown of blood clots by pharmacological means.

Figure 2.5: Annotated stroke management pathway. Shaded circles reveal the information systems (details in Appendix A) where information pertaining to the activity is stored.

Objectives

Despite the retrospective and prospective nature of the collected data, issues have been reported regarding the inconsistency with which some of the data elements are recorded over time, as well as the lack of other information such as haematology or biochemistry readings. An added technical challenge is the existence of two distinct databases with different schemata, one which contains legacy data and a second operational database, which is currently in use. In this case it is important that each database is seen as a separate data source.

The purpose of this exercise is to collect data for the development of the Research Stroke Register, a data warehouse comprising information on stroke patients from multiple hospital sources. Details of the development of the data warehouse are given in the next chapter.

2.5.2 Application of the Methodology

- **Identifying Data Sources**

  Four main data sources were identified for this case study: the new stroke register (NSR) and the old stroke register (OSR) which together identify the complete cohort of patients (these already contain some information from other systems such as radiology, emergency department, and PAS); the biochemistry system (LAB) where blood tests are stored; the patient administration system (PAS) where comorbidities and follow-up information is present, and a national tracing system (NSTS) where it is possible to accurately retrieve patient demographic information on a national level, such as dates of death. Each of these systems were put through the process depicted in figure 2.3 and a combined summary of each major step is given below.

In addition to the above-mentioned systems and as noted in figure 2.5, other systems that are involved in the stroke pathway are the Accident and Emergency System (AE) and the Radiology Information System (RIS which for simplicity, it encompasses the Picture Archiving Communication System (PACS)). However, information from those systems is written onto the paper-based stroke register form, which is in turn entered onto the OSR/NSR. Information in the AE and RIS systems is then inspected to ensure the OSR/NSR database contain accurate information. For this reason, and also because of the nature of the research database to be developed, AE and RIS were not explored in detail in this case study.

One of the initial requirements set in the methodology is that the first data sources to be put through the process are those that can allow us to select the cohort of interest. In the stroke study, the cohort of interest comprises all records of stroke events in the OSR and NSR databases.

Biochemistry test results from the LAB source need to be carefully selected to match the respective stroke events. A selected number of blood tests was defined by the domain experts. Conversely, the data required from the PAS source was mainly demographics (such as date of birth, death, and ethnicity) and the electronic discharge letters (ICD coding).

The steps taken to extract data from the sources is summarised below.

- **System Understanding**

  The stroke registers (OSR and NSR) were by far the most challenging systems to understand due to the sheer volume of records and data attributes as well as the lack of a data dictionary and consistent formatting and syntax. The two distinct databases, OSR and NSR, are treated separately here as they need to go through a process of schema matching that can only be achieved if the sources and their data are understood (semantically and syntactically).

A series of consultations were carried out with the stroke data team in order to preview the system as well as gather necessary credentials to access the database. As a result, an honorary contract was needed in order to carry out the work.

The OSR and NSR are password-protected Microsoft Access databases. The PAS and LAB systems had already been previewed in the prostate cancer study yet credentials still had to account for access to these systems. The PAS system contains demographic information and coding, whereas the LAB system includes information on biochemistry tests carried out.

- **Data Understanding**

  Regarding the understanding of the data, the most time consuming sources were again the OSR and NSR mainly due to the lack of a data dictionary. Indeed, the paper based forms used to record patient-specific stroke data in the NSR and OSR databases was, together with consultation, most useful for the development of a data dictionary.

  The OSR and NSR databases needed to be merged later, and for that reason it was important at this stage to identify common attributes between them and gather enough metadata to produce a data dictionary.

  Challenges were identified in attributes' values including patient identifier fields. Data preprocessing was needed, for example, to split the identifier field which contained both the patient surname, initials and the hospital number, inconsistently recorded throughout the records. Furthermore, several fields' values changed overtime from a numeric coding into a descriptive text (e.g. discharge destination used to be coded as 1 when patients were discharged home, and this changed to 'home'). Regarding the data collection methodology, the initial concern is with the fields that allow for the linkage and retrieval to other sources.

  Once the identifier fields were formatted in the OSR and NSR, the other

Metadata - Understanding the System	Examples from the stroke study
Appropriateness and authoritativeness of system	All identified systems are authoritative and contain crucial information on stroke events.
Data flow chart	Data flows from PAS to LAB (so that records and hospital numbers are linked), but there is no systematic flow of information between OSR/NSR and any other systems. The stroke data team does, however, validate some of the data by manually querying PAS, the emergency system, or LAB.
Data extraction method (if known)	PAS - Business Intelligence software; LAB - Business Intelligence software and developed software; OSR - Back-end database query; NSR - Back-end database query; NSTS - Information Services Report.
Preview and system training	Previewed all systems, training only required for NSR an OSR.
System credentials	Read-only access to all systems.
Key data elements present (if known)	Hospital number, NHS number, Date of Birth and Date of Event present on all systems.
System Updates (impact)	The OSR is no longer updated however, the NSR is currently updated and records from previous months may be updated retrospectively.
System Limitations	The LAB system was introduced in 2003 and there is no electronically available data prior to this date.
System Live Date	OSR - 1996 to 2008; NSR - 2009 to present; LAB - 2003 to present; PAS - 1990 (and earlier) to present.
Last System Update Date	This information was available but changed throughout the application of the methodology.

Table 2.7: Summary of items gathered when understanding the system (stroke study).

sources (LAB, PAS) were then investigated. Both sources had been previously investigated in the prostate study, however, in the case of the LAB source, additional investigation was necessary to determine the appropriate biochemistry coding for each of the selected blood tests. Apart from semantic differences, the data types and formatting remained the same as in the prostate study.

The NSTS system was used in one occasion although it can only be accessed

by analysts in the Information Services Department at the hospital. The requirement for the retrieval of data from the national tracing service is a spreadsheet containing patients' NHS number and Date of Birth. The advantage of using the NSTS service is that it is thought to be most reliable for obtaining dates of death or current postal address, although causes of death are not stored here. Indeed, causes of death may be stored in electronic discharge letters (PAS/LAB) when patients die in hospital but otherwise they are not recorded in the HIS. Nevertheless it is expected that the PAS system is updated with patient details including dates of death and as such, it is a valuable source of information, particularly for those that die in hospital. A data linkage exercise carried out in the next chapter will study the quality of this data in more detail.

- **Extraction Preparation**

  The first data sources from where the cohort is extracted are the OSR and NSR. Because they are Microsoft Access databases, a simple query was designed, taking into account the issues identified in the previous step. The constraints used to query these sources were based on the attributes that allow linkage (identifiers).

  Subsequently, queries were built using an OLAP software to extract information from PAS on patients' episodes. This was done by deterministic linkage using the internal identifiers.

  The sheer size and number of queries to be run on the LAB source (21 blood tests for each patient in OSR/NSR), warranted an exploration of retrieval methods that cope with such volume of data. The information needed for every stroke episode was a summary for each blood test. The summary includes the blood readings closest to admission and discharge as well as a summary of the readings during the stay (maximum, minimum, average, standard deviation and total number of readings).

Queries were built using an OLAP software, a first approach used patient identifiers to extract all information on bloods for a particular time frame. This method was extremely time consuming not only due to the retrieval process, but also the fact that OLAP queries are limited in the number of clauses, meaning that several batches of identifiers needed to be run. A second method, need not split into batches yet it generated very large datasets for each blood test which then needed to be imported into a separate working database and queried using suboptimal software for large databases.

A third approach was to develop a script that runs a query for each stroke episode and automatically computes the required variables. This was the chosen approach due to a considerable improvement in querying time as well as reduced preprocessing. This is discussed again in more detail in the next chapter.

Table 2.8 gives a summary of the selected extraction methods for the stroke study.

Data Source	**Extraction method used in the stroke study**
OSR/NSR	Database query with constraints on identifiers.
PAS	OLAP software tool to retrieve patient episodes' records, in batches of 300 patients per query.
LAB	Sets of queries were developed to compute and retrieve biochemistry tests.
NSTS	Report run by Information Services analysts.

Table 2.8: Extraction methods for each data source identified in the stroke study.

- **Extraction & Evaluation**

  This step is crucial in assessing the quality of the retrieved data. Regarding the LAB source, simple queries were first built for individual patients using the OLAP software. Such queries allowed the manual inspection of the blood readings over time and the identification of potential sources of error. In the course of this inspection, some biochemistry codes changed over time which meant their readings were not consistent throughout. This prompted

further consultations with the pathology department in order to establish the correct codes using their coding table. Throughout this project and particularly during the stroke study, an OLAP query was developed to inspect the coding table, which includes the codes for all biochemistry and pathological tests carried out at the hospital. This was an invaluable source of information that should be made available.

In the OSR/NSR databases, used to produce the cohort of stroke patients, a small number of patient identifiers was erroneous and so, the extraction methods had to be re-run based on other identifiers such as NHS number and date of birth. This was a straightforward task but one which added further time in retrieval of data.

Because of the previous exercise on prostate cancer, the assessment of the quality of the other retrieval methods, such as the ones used with PAS, was greatly facilitated and no further issues were identified.

An analysis of the quality of the data is given later as here the concern is with the retrieval methods.

Using the data collection methodology a stroke ODS was created containing the data from the above-mentioned sources. The development of the research stroke register database is explained in detail in the next chapter along with summary statistics.

2.6 Conclusions

The collection of patient-centric data from secondary centres, such as the one described here, is a non-trivial task, and one which has received relatively little attention. No methodologies have been developed or tested for this purpose. In particular, data mining and knowledge discovery methodologies overlook the

process of multi-source data collection. Existing methodologies for data analysis, knowledge discovery, and data integration were reviewed and provided the framework to the patient-centric data collection exercise presented in this paper. The exercise was based on a case study on prostate cancer, validated using the stroke study, and presents the lengthy journey of retrieving and collating patient-centric multi-source data for research or other secondary uses. It is hoped that, by exposing the complexity of data retrieval in hospitals, this experience contributes to researchers and future data analysis studies, and that analysts are able to ensure the context of their data is fully understood.

The contextualisation of the retrieved data is critical in health and is important to understand the ways in which the hospital operates (including how data from inpatients and outpatients are recorded and coded, as well as other events that may not be classed as the first two, such as day cases in the NHS).

The work presented in this chapter has also identified that the collection of metadata and creation of a metadatabase in parallel with the study data pool are crucial to its understanding, linkage, and validation.

Furthermore, the practical obstacles of working in a heterogeneous multi-source environment, where data access is limited, are addressed and the key informational requirements for a successful data collection task in hospitals introduced. It has been identified that it is possible to obtain reliable research data from a secondary care centre. It is my belief that research studies should be encouraged to use such data, provided that hospitals and the NHS facilitate the process of retrieval, using a standardised process. It is expected that such process, when followed, will guide and reduce the time spent by researchers and hospital staff on data collection and related bureaucracy, whilst at the same time improving the validity of study data collected from hospitals. It should also be reassuring that de-identified patient data can be used for public good, and that, by creating a greater demand for data, their quality and accuracy are improved in hospitals.

Because each data source is invariably linked to a HIS, the work carried out in this chapter has also confirmed that the current HIS implemented in this particular secondary care centre, provide little help to extract or work with cohorts of patients. Indeed, most HIS are developed to work on a single-patient basis. This was particularly true with the PACS and LAB where an additional software had to be developed to extract the data.

2.6.1 Chronology of systems

Often overlooked, a full picture of the history of the implementation of hospital information systems is key to understanding hospital data as well as identification of potential issues arising from an ever changing environment. Figure 2.6 depicts the implementation of hospital systems over time and was compiled with information obtained during the case study in prostate cancer. However, since the time of writing, another PACS system was introduced, as was another version of the Somerset Cancer Registry (CRE) and the LAB systems.

Figure 2.6 illustrates the complex nature of information in hospitals from the point of view of systems' implementation. This shows critical changes to the way data is recorded and its availability over time and poses significant difficulties to the study of chronic diseases or longitudinal analyses using routinely collected data.

The case study on stroke was facilitated by the fact that the only sources prone to major changes were the stroke services databases. This, although increasing the complexity of schema matching, allowed for the core features of the target group to be locally available and for the staff to be aware of its changes. Indeed a major obstacle detected in the prostate study was the overlapping of systems and hospital departments, where neither information is consistently inspected nor staff were fully aware of its existence. Some of this has changed with the extension

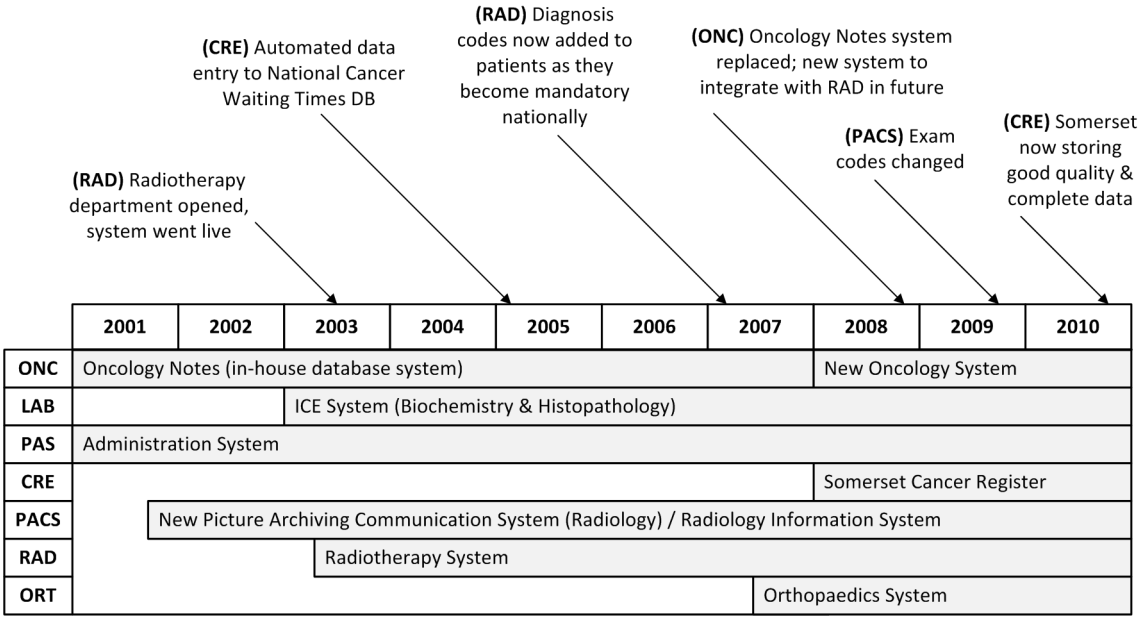


Figure 2.6: A chronology of hospital information systems relevant to prostate cancer.

of the national cancer waiting times data requirements and the introduction of adequate staffing, such as patient pathway coordinators, that oversee the data that is required nationally.

The stroke study was also facilitated by the acute and episodic nature of the disease, where the average patient length of stay is substantially shorter than for a cancer patient.

2.6.2 Further work

Recent work [85] carried out at the National Centre for Infection Prevention and Management, Imperial College, London, saw the development of a research database to facilitate hospital epidemiology and hospital syndromic surveillance. This work is similar in nature to the one carried out in this chapter. The authors reported issues and obstacles similar to those presented here and their work fur-

ther demonstrated the value of a data repository, data linkage and the importance of "more sophisticated uses of existing NHS data, and innovative collaborative approaches to support clinical care, quality improvement, surveillance, emergency planning and research"[85].

Likewise, it was identified that future work is needed to improve collaborative approaches between hospital departments, systems' owners and software vendors to enable the secondary uses of routinely collected hospital data. Given that technical solutions to data integration, storage and retrieval are possible, the major challenge will continue to be the top-level organisation of systems and their data models or ontologies. A sustainable, consistent and replicable approach for the implementation and organisation of health information systems across hospitals would greatly benefit their local management, particularly in a National Health Service environment. At present, the implementation of localised and isolated systems, each s a particular task or problem, will continue to add obstacles to the secondary uses of hospital data.

Given such a setting, it would be interesting to carry out further work on automated or pseuso-automated methods for ranking relevance of databases and data elements based on the collected metadata and informational needs. A pseudo search engine using such rankings, together with a retrieval system or OLAP software, would make useful the identification and selection of feasible, high quality data for research or other secondary uses.

Chapter 3

Preprocessing, Linkage and Data Warehousing

This chapter is divided into two main sections, the first on the prostate cancer study and the second on the stroke study. Throughout the two sections, selected methods and techniques are introduced and describe in detail the most relevant challenges in data preparation, linkage, integration and quality of routine hospital data. The section on prostate cancer addresses two problems: mining information from histopathology text reports (section 3.2.1); and data preparation and continuous value estimation (the age problem, section 3.2.2). The section on stroke addresses record linkage methods (section 3.3.1), and data warehousing and integration (section 3.3.2). The techniques presented here have all been developed in the course of this research.

3.1 Introduction

Data preprocessing can be broadly defined as a set of actions taken before a data analysis process starts [86]. Data preparation, data cleansing and data preprocessing can be considered analogous terms [67], although cleansing is often regarded as a preprocessing task. In the previous chapter, data mining and knowledge discovery methodologies were reviewed and preprocessing was identified as a key step leading to analyses. The methodology for data collection and development of a semi-integrated repository already introduced some degree of data preparation, particularly with respect to key identifiers. However, the main concern was the retrieval of patient-centric information and the development of a storage infrastructure that replicated the original hospital environment. This allowed some obstacles to be overcome (for example, integration), whilst keeping the data in its original form, to be further inspected and studied. This chapter is concerned with the inspection of the routinely collected hospital data available in the repository and the development and application of preprocessing techniques.

The latter are motivated by the need to [86]:

- solve data problems that may prevent analyses on the data;
- understand the nature of the data and perform more meaningful data analyses as a result;
- extract more meaningful knowledge from a given set of data.

Data preprocessing has also been formally defined in [86] as an operation that transforms raw real-world data vectors into a set of new, useful vectors where valuable information is preserved and issues with data are eliminated.

Depending on their purpose, preprocessing techniques may not always yield the results desired. It has been reported that data carefully prepared for warehousing

proved unsuitable for modelling whereas data that has been prepared specifically for modelling produces significant improvements in model accuracy [87]. It has also been reported that some preprocessing techniques may provide little improvement in the overall results of analyses, particularly when these represent summaries [88].

In the field of data mining and machine learning, data preprocessing has been identified as a key task in determining the outcome of knowledge discovery algorithms [5; 20; 86; 87; 89]. Erroneous, extraneous or inadequate data leads to less accurate or less understandable results, failure to discover any trends in the data, or poor generalisability of the developed models [86; 89]. Missing data, in particular, is a well established barrier to successful data mining and knowledge discovery [5]. As a result, preprocessing has received some attention from the computing sciences but still remains an *ad hoc* process and one often aimed predominantly at increasing the value and generalisability of data mining models.

However, in medical research, epidemiology and, to a degree, statistics, descriptions of preprocessing methods are lacking [7; 63; 86]. Regardless of robust study designs or reliable data collection methods, studies still have to deal with errors from various sources and their effects on the results [7]. The data preparation and cleaning process, as a whole, with its "conceptual, organizational, logistical, managerial, and statistical-epidemiological aspects", has not been described or studied comprehensively [7]. In the peer-reviewed literature, particularly in medicine and epidemiology, it is rare to find any statements about data-cleaning methods and respective error rates [7; 63], yet medical data is often described as most complex, voluminous and heterogeneous to analyse [20]. Study design, protocol compliance, and the integrity and experience of the investigators often receive more attention than any preprocessing methods used [7].

In addition, most medical statistics and epidemiology textbooks have scattered information on data preparation and cleansing techniques, and do not analyse

them in detail. As such, there are significant gaps in knowledge on data-handling methodologies and standards of data quality [7; 63; 90]. Furthermore, there is also a lack of standardised methods for assessing data quality in electronic health records and systems [18; 91] which further hinders the use of routinely collected data.

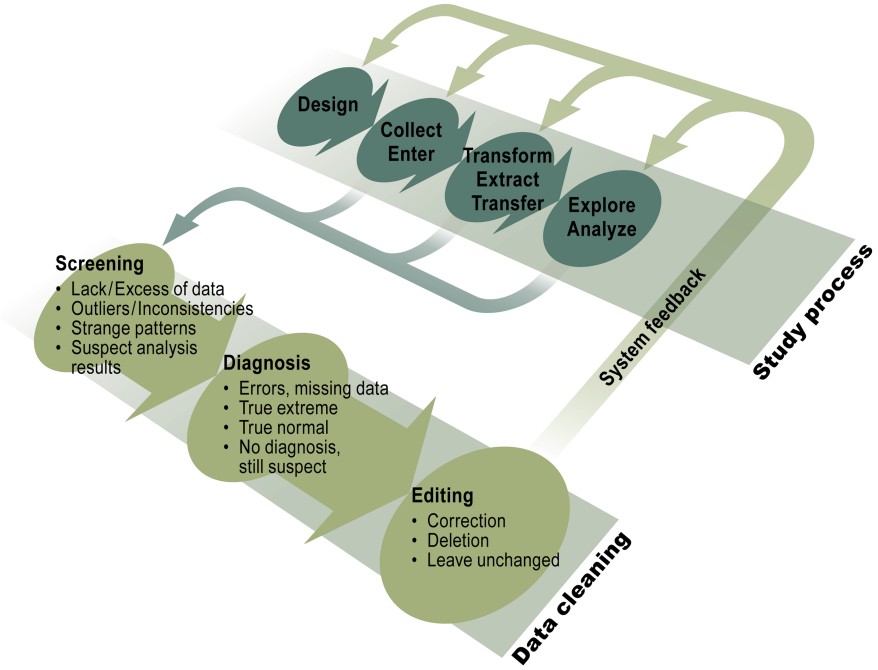


Figure 3.1: Data Cleansing Framework (from [7]).

A framework for the process of data cleaning has been proposed for medical research and it is depicted in figure 3.1. The framework consists of a three-stage process with repeated cycles of screening, diagnosing, and editing of suspected data abnormalities [7]. The screening phase aims at detecting missing or extreme values, often by inspecting overall summary statistics of the datasets. In the diagnostic phase, the purpose is to clarify the nature of the identified errors, for example, by perfoming integrity checks. Lastly, the editing, or treatment phase, is concerned with the implementation of a method to deal with erroneous data (either by deleting, correcting or leaving them unchanged). This framework helps

to illustrates how some preprocessing tasks (i.e. data cleansing) are carried out as part of a study. In the future, this framework could be generalised to include other preprocessing tasks, whilst keeping the three key phases.

In this thesis, preprocessing is defined as an activity encompassing all the necessary steps to obtain a dataset or data field for analysis. Examples of preprocessing tasks include general data cleansing, formatting, imputation and value estimation, information/data extraction or mining, and record linkage. The latter is often seen as a retrieval technique, however, when it is used for the purpose of compiling a more complete dataset or for validation, it can be seen as a preprocessing task. This chapter addresses the most significant preprocessing issues and techniques developed in the course of this research and explores how they impact the quality of the data. Further preprocessing, however, was still required in the analyses carried out in other chapters and it is summarised when appropriate.

Four key preprocessing techniques from the case studies on prostate cancer and stroke are introduced in the following sections. They are:

- a technique for extracting information from text reports (section 3.2.1),
- an algorithm for the estimation and imputation of continuous ordinal data (section 3.2.2),
- a set of deterministic record linkage rules (section 3.3.1), and
- a proactive approach for data warehousing and integration (section 3.3.2).

Throughout this chapter, m is used to identify the percentage of missing records or values (so when no data is available $m = 100\%$).

3.2 Prostate Cancer Study

3.2.1 Mining Histopathology Reports

It is well established in the literature that one of the major challenges in working with medical data are unstructured text reports [20; 92; 93] and that the discovery and extraction of new knowledge from unstructured text data is the primary goal of text mining techniques [93; 94].

The term text mining is often used loosely to describe automatic or semi-automatic techniques that use or analyse text and natural language. There are, however, significant differences between disciplines and techniques for the understanding, extraction or classification of unstructured text documents and these will be explored in detail in the next section.

Free-text histopathology reports, confirming diagnoses of prostate cancers, were collected electronically as part of the prostate cancer study. In their original state they are not usable for analyses because the narrative text, in its natural language, can not be easily interpreted by computers. It is therefore crucial to mine such text reports and extract useful information in a canonical form. This section introduces a technique developed to mine and extract useful information from the histopathology reports. The definition of useful information in the context of the prostate cancer case study is given later in section 3.2.1.2.

The following section (3.2.1.1) provides key background on the disciplines of text mining, information retrieval and natural language processing, which are relevant to contextualise the developed technique.

Section 3.2.1.2 introduces the histopathology of prostate cancer, and it is followed by a section on data familiarisation, and another on defining goals and useful information.

The algorithm design, testing and results are presented in section 3.2.1.5 followed by a discussion and evaluation of the developed technique.

3.2.1.1 Background on Natural Language Processing

There are arguably three major areas in computing research competing, in different ways, to gain understanding of unstructured texts: information retrieval, natural language processing and text data mining.

Classical information retrieval deals with the search for particular documents based on a predefined set of terms [95]. Examples of classical information retrieval techniques are those employed by internet search engines using algorithms such as PageRank [96] or HITS [97]. Such techniques inspect the frequency of the terms in documents and compare them against another set of terms (such as a user query), yielding a similarity value between the two. This value can then be used to rank documents against a particular query. Nevertheless, studies using such techniques in health informatics and medicine often mislabel their work as a data mining exercise [98] when it should be information retrieval.

Text data mining, however, is concerned with the classification or clustering of documents. An example of a text mining technique is the development of a multilayer neural network trained to classify medical documents in the area of cell biology [99]. Having achieved a satisfactory accuracy, the neural network developed in [99] was later used to automatically filter and classify documents to selected domains of cell physiology. Another example is the classification of lung cancer tumour stages from free-text reports [100]. The authors of this study used a support vector machine, trained for each stage category, and based on word occurrences from a corpus of histology reports. They reported an average stage classification accuracy between 69% and 75%, illustrating the difficulty of developing high accuracy text mining algorithms.

Natural language processing (NLP) is an overarching discipline concerning both text mining and information retrieval in that it acts as a means of turning text into data for analysis. It has also been reported that NLP research focuses on building computational models for understanding natural language [101]. NLP techniques vary from trivial syntactic analysis to complex semantic representation and interpretation systems using propositional logic or first order predicate logic [102]. Modelling context, however, has been described as the most difficult and least well understood aspect of NLP [102]. An example of a comprehensive NLP system is one that incorporates existing techniques such as the MMTx (MetaMap Transfer), a tool for mapping clinical text concepts to a standardized vocabulary, and NegX, a negation detection algorithm, to automatically extract and identify medical problems from a predefined list conditions [103].In fact, the negation detection algorithm can be seen as a NLP technique in its own right, and it is one that works primarily on syntax using regular expressions.

Having a general pattern notation, regular expressions allow the efficient description and parsing of text [104]. They have been used effectively, for example, in the identification and extraction of instances of documented blood pressure values and anti-hypertensive treatment from clinical reports [105]. In this example, regular expressions can be seen as a NLP technique for the purpose of information extraction (IE). In turn, IE has been described a specialised sub-domain of NLP concerned with the extraction of predefined types of information from text [101].

A review of the state of the art of information extraction from clinical texts highlighted five major characteristics of such texts, which are also seen as challenges [101]:

- Clinical texts are ungrammatical and composed of short, telegraphic phrases.
- Clinical narratives are repleted with shorthand (abbreviations, acronyms, and local dialectal shorthand phrases). It has been estimated that acronyms

are overloaded about 33% of the time and are often highly ambiguous even in context [106].

- Misspellings are common in clinical texts, particularly in notes without rich-text or spelling support.
- Clinical narratives can contain any characters that can be typed or pasted such as pasted sets of laboratory values or vital signs.
- Attempts to bring structure and consistency to otherwise unstructured clinical narratives, templates and pseudotables are common (for example,, plain text made to look tabular by the use of white space).

The same review pointed out the difficulty in evaluating NLP systems and the need for clinical texts to be annotated in order to help increase the effectiveness of IE or NLP techniques. Furthermore, the issue of portability of available techniques has also been identified. Most algorithms are domain-specific and hence not replicable with the same accuracy elsewhere.

The work carried out here pertains to the extraction of key information from prostate cancer histopathology reports. The techniques presented in the next sections may be described as information extraction techniques and the challenges and issues described above will be later discussed in regards to this case study.

3.2.1.2 Histopathology and Tumour Grading

The most definitive diagnosis confirmation of prostate cancer relies on a biopsy of the prostate tissue. The biopsy histology is analysed by histopathologists and the diagnosis and prognosis largely depend on the degree of differentiation of the neoplastic or malignant cells [107], for which a grading system is used. This histopathological grading should not be confused with the TNM Classification of

Malignant Tumours (TNM) staging [1], which is widely used for all cancers. The TNM staging for prostate cancer describes the extent of the cancer, whether it is localised *in situ*, extracapsular (beyond the prostate capsule), invading lymph nodes or metastatic. The TNM may also be assessed pathologically; however, this is not an assessment of the differentiation of the tumour's cells, but rather, its extent or invasion. The histologic grade, on the other hand, is centered on the tumour itself.

Several grading systems have been proposed in the last three-quarters of the 20th century yet the most commonly used worldwide [108] and at the NNUH is the Gleason system. The Gleason grading system is based exclusively on the histologic pattern of arrangement of the carcinoma cells [107; 108]. Such patterns have been consolidated into five grades, depicted in figure 3.2.

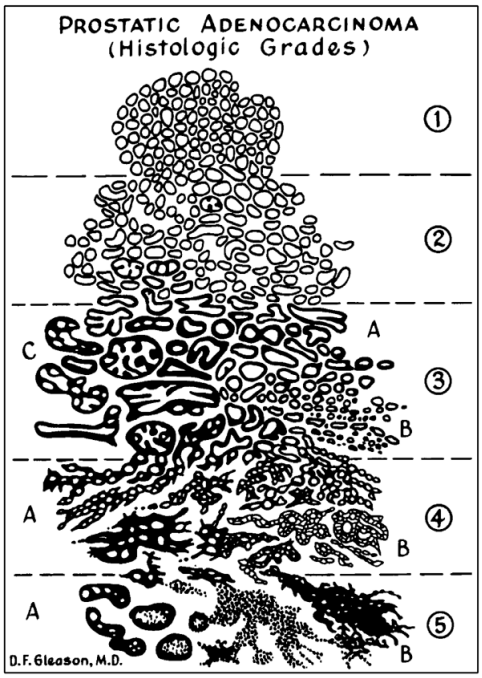


Figure 3.2: Standard Gleason grades [8]

A histologic score ranging from 2 to 10 is created based on the addition of the

two most common occurring patterns (i.e. the first most predominant in area by simple visual inspection, and a second most common pattern [108]). Histopathologists may also indicate a tertiary score, although this is less common than traditionally reporting two grades. When only one grade is present in the tissue sample, that grade is multiplied by two to give the final score [108]. The score can then be divided into low-grade (2-6) and high-grade (7-10). The higher the grade, the most differentiated the cells are and consequently the more aggressive the tumour is likely to be. The Gleason grading method is exclusive in human cancers as other malignancies are often based on the worst grading observed [108] rather than an addition of the most common grades.

Histologic tumour grading poses a degree of subjectivity. Indeed pathologist training and experience can influence the degree of interobserver agreement [107]. According to Allsbrook *et al.* [109], 'nonconsensus' cases in prostate cancer include low-grade tumours and those on the border between Gleason patterns. The same study concludes that although there is some variability in interobserver agreement among urologic pathologists, interobserver reproducibility of Gleason grading is in an acceptable range. These findings should reassure that the Gleason grading system provides somewhat consistent observations among different pathologists.

As previously mentioned, high Gleason grades are associated with aggressive tumours. In fact, increasing Gleason grades are directly related to a number of histopathological end points[107]: lymphvascular space invasion by carcinoma, tumour size, positive surgical margins (what should be a normal tissue margin at the edge of the specimen is also positive for carcinoma), pathological stage (including risk of extraprostatic extension and metastasis).

Neverthelss, patients with lower-grade (2-6) carcinomas are still at risk for having cancer spread beyond the prostate and not all patients with a high-grade carcinoma component (7-10) will have extra-capsular extension of the carcinoma

(extension beyond the prostate)[107].

The same author concludes:

> *'Prediction of pathologic stage by needle biopsy Gleason grade alone is possible but is not absolutely accurate for the individual patient'.*

As such, clinically, the Gleason grade is usually combined with other pretreatment factors, such as serum total PSA, % free PSA, local clinical tumour stage, and amount of tumor in needle biopsy, in order to predict pathologic stage [107].

Partin *et al.* [110] identified that PSA level, together with TNM clinical stage [1], and Gleason score, contributed significantly to the prediction of pathological stage ($p < .001$). As a result, the same team created the Partin tables (a nomogram) which estimates the risk for extraprostatic (extracapsular) extension, seminal vesicle invasion, and lymph node metastasis. Partin *et al.* continued collecting data and later revised the Partin tables in 2001 [111]. Further studies confirming the predictive ability of the Partin tables have been carried out in Wales [112] and other countries. However, a 2009 study in Canada [113] and another in Ireland [114] recommend caution in their clinical utilisation as the tables show poor predictability of pathological staging at radical prostatectomy. Later, in 2010, researchers from the Irish study presented two novel predictive models based on the Partin tables [115].

Despite the controversy surrounding the Partin tables, the key informational elements conveyed in them are highly relevant for the prognosis of prostate cancer. Indeed, the National Institute for Health and Clinical Excellence (NICE) guidelines states that Partin tables are the most commonly used clinical nomograms for determining the risk of nodal spread (i.e spread to lymph nodes) and uses the same informational elements to compute risk stratification for men with localised prostate cancer [116]. As such, tumour TNM staging and PSA value

are important elements to use together with the Gleason grade and should be extracted from the histopathology text reports when possible. This defines the informational elements needed from histopathology reports in the case study.

After a first inspection of the histopathology reports retrieved as part of the prostate cancer case study (in section 3.2.1.3), section 3.2.1.4 will discuss the feasibility of extracting such data based on concrete examples from the case study.

3.2.1.3 Data Familiarisation

The histopathology dataset, as retrieved from the ODS, contains a total of 5,083 reports with the following data attributes:

- $HospNo$, the patient identifier;
- $DateOfEntry$, the date when the histopathology report was produced;
- Age, the age of the patient at the date of entry;
- $FullReport$, the free-text histological report.

From the above attributes, only Age had missing data (m=28.5%). The most important attribute to investigate is $FullReport$, containing the full semi-structured histopathology text report in natural language.

The following is an anonymised example of a typical histopathology text report in its original format:

ADDRESS FOR REPORT: XXXX, Urology,Copy To: Report to Cancer Registry,, HISTOPATHOLOGY REPORT,LAB No: XXXX,,CASE HISTORY:,PSA = 6.4 on 5 February XXXX. Known carcinoma of prostate,(Gleason 3+5 = 8 in right lobe; XXXXXXX). Post DXT. Abnormal,hypoechoic left peripheral zone on scanning.,, MACROSCOPIC:, 1. Right: four cores of pale tissue, 1.8 cm, plus fragment., 2. Left: four cores of pale tissue, 1.8 cm. YG.,, MICROSCOPY:, 1. Two cores of prostatic tissue contain Gleason 4+5 = 9,adenocarcinoma. There is focal perineural invasion., 2. Cores of prostatic tissue, which show no evidence of high-,grade PIN or carcinoma.,, DIAGNOSIS: ,PROSTATIC BIOPSY: ADENOCARCINOMA (GLEASON SUM = 9),, REPORTED BY: Dr X Consultant Histopathologist, Dr X Consultant Histopathologist,,REPORT DATE: XX.XX.XXXX

Upon a first inspection, the text reports presented an overall average length of 972 ± 493 characters and 129 ± 73 words. From the above example report, it is possible to observe an overall general structure is present. The reports appear to be structured; divided in sections, each providing short descriptive sentences on the patient and the tissue sample(s) observation(s). However, a preliminary inspection revealed that not all reports have the same sections. An algorithm using regular expressions was developed to identify and quantify the sections in all histopathology reports. The following sections were identified, with corresponding missing values:

- Address for Report / Copy To (m=0.01%)
  The name and address where the report is sent to.

- Case History (m=5.76%)
  The patient history, any symptoms, results of other examinations or procedures. This section needs to be ignored by mining algorithms since it includes information that is not directly related to the findings of the report and can produce false positive results.

- Macroscopic / Macroscopy (m=0.06%)
  The macroscopic observation (for example, enlarged size of prostate).

- Microscopic / Microscopy (m=0.16%)
  The microscopic observation of the tissue sample.

- Diagnosis (m=0.06%)
  The diagnosis of adenocarcinoma, benign hyperplasia, prostatitis or other.

- Reported by (m=0.01%)
  The consultant histopathologist who analysed the tissue sample and others who reviewed the specimen and/or contributed to writing the report.

5.8% (n=293) of all reports did not have a *Case History* section and all other sections had a marginal number of missing values. It is possible to say that, apart from *Case History*, which is not a field required for the analyses, over 98% of reports have all remaining sections and that the overall structure of reports is somewhat homogeneous.

3.2.1.4 Defining Useful Information

When observing the sentences within the reports, the presence of certain terms such as adenocarcinoma or hyperplasia does not, on its own, reveal a positive or negative observation of that term. In order to understand the observations, the complete sentence and text report need to be understood. Sentences describing

prostate tissue with no malignancy (see example report, line 10) and others with adenocarcinoma (example report, line 8) are common. This is misleading for naive mining algorithms that do not attempt to understand the document yet extract information on a sentence-level context. Therefore, prior to any analyses it is essential to determine the most relevant information to extract from the reports.

The first most authoritative source of conclusive information from histopathological reports on prostate cancers are the Gleason grades. Indeed, in a given report, there may be more than one grading, for different specimens, and so the most relevant would be the highest grade reported. A preliminary regular expression search revealed that all reports had at least one occurrence of the term *Gleason*. However, Gleason grades are present in reports in different natural language structures and not all reveal a clear staging (typos were also found, for example, "Gleeson"). Therefore a successful algorithm to mine such information needs to first focus on the correct extraction of different structures throughout reports (section 3.2.1.5). In some reports, the only reported grades are the total Gleason sum and there is no discrimination between the first and second values. In such cases it is not possible to assume whether a total Gleason of 7 is the result of 4+3 or 3+4, which are histologically different. In order to cover all reported grades, the total Gleason sum should be computed and when possible, the two (or three) individual values should be extracted.

As previously discussed, other informational elements such as TNM staging and PSA are also important. Using regular expressions, it was possible to identify that tumour staging is present in 21% of histopathology reports in the form of $'pT1a'$ or simply $'T1a'$ (i.e. $T1a$ denotes the tumour was incidentally found in less than 5% of prostate tissue resected, accoring to TNM [1]). This notation uses the TNM parameters with a prefix modifier, p, denoting the stage was given by a pathologic examination of a surgical specimen. However, because in the

databases it was not possible to consistently retrieve TNM staging, any exercise to retrieve the staging, whether clinical or pathological, would be worthwhile. Nevertheless, the number of reports including TNM in the above format is low and so a careful inspection of the tumour grades available in the reports is carried out later in section 3.2.1.6.

The PSA readings are a biochemistry test and, as such, are present in a more canonical form in a different database. PSAs are "easily" retrievable in a canonical form from the ODS and can be linked using the patient identifier. Therefore, no algorithms were developed to identify or extract PSA values from the histopathology reports.

The following sections introduce the systematic approach to developing the techniques to extract the Gleason grades and potentially TNM staging from the reports.

3.2.1.5 Algorithm Design

Designing algorithms to retrieve information from unstructured text is a challenging task in two major areas: the accurate extraction of information, and the validation or evaluation metrics. The extraction of information relies on a thorough understanding of the natural language and its structure. Validating the accuracy of the algorithm, particularly in NLP exercises, often relies on a manual clerical review of a selected number of reports [117]. The first objective of the information extraction algorithm developed and presented below is to retrieve the Gleason sum and, when possible, the two (or three) values used to compute it. A second objective is to extract TNM staging when available.

A first, feasibility algorithm was built to analyse and understand the structure of the text adjacent to the keyword Gleason and is described below.

	Symbols			Numeric Values		
Type	=	+	neither	0 (none)	1	2 or 3
Coverage	63.11%	89.47%	10.17%	0.28%	6.63%	93.09%

Table 3.1: Text Mining Feasibility algorithm: Total number of symbols and numbers in relevant text segments and their respective coverage.

Sentence breaking or sentence boundary disambiguation (SBD) was used such that $'.'$ and $' :'$ are the only symbols that unambiguously denote the end of sentences, given the general text report structure (as seen in the example report). This approach will miss sentence boundaries when the reports' sections are segmented by commas and have no other punctuation. Nevertheless this was considered sufficient to highlight the most common structures of the text adjacent to the keyword of interest:

1 "GLEASON 4+3=7"
2 "Gleason pattern 3"
3 "GLEASON SCORE 4+3"
4 "Gleason score 7, (3+4)"
5 "Gleason patterns 4 + 5 (and some 3)"

The above led to an investigation of the total number of symbols (addition and equal signs) and numeric values often present in the *Gleason* text segments. This is an important step in shaping an algorithm that can accurately retrieve the Gleason grades.

Table 3.1 shows that 90% of the *Gleason* text segments contain the addition sign, 63%, the equal sign and 93% contain two or three numeric values. This feasibility analysis was carried out on the first occurrence of the word *Gleason* in 5,083 reports, and leads to the definition of concrete rules to extract Gleason grades from the histopathology reports.

Based on the feasibility analysis, three baseline rules to retrieve the Gleason grades from identified text segments were defined:

- **Baseline Rule 1**
  IF symbol "=" exists AND is followed by a numeric value THEN
  Gleason grade ← numeric value
- **Baseline Rule 2**
  IF "+" exists AND is surrounded by two numeric values THEN
  Gleason grade ← $\sum$ of the two numeric values
- **Baseline Rule 3**
  Gleason grade ← the first occurring numeric value

The baseline rules, together, generate three different possible Gleason scores. An algorithm was developed to implement the above rules on all sentences (text segments) where the word *Gleason* occurrs in a single document. The maximum grade found in a document, for each rule, is used. The developed algorithm first strips off the sections that are not relevant and could bias the results (i.e. case history, address for report, reported by), then identifies every segment of text where word Gleason occurrs and applies the baseline rules to that segment, yielding a result for each baseline rule.

The results from the baseline rules can be combined together to maximise the accuracy of the retrieved Gleason grades. A hierarchical selection of rules (or combination of baseline rules), allowing the selection of most relevant and accurate information was identified:

$R0$ If Baseline Rule 1 and Rule 2 retrieve the same number, then this number is the most authoritative Gleason grade. This is because there is evidence

that the total Gleason grade was indeed computed as a sum of the other two observed grades. E.g: "Gleason patterns 4+3=7".

$R1$ The result of a valid Baseline Rule 1 is the second most authoritative. The equals sign is the second most reliable source of evidence of the final (computed) Gleason grade as it reiterates the given score and it is unlikely to be entered by accident. E.g: "Gleason score=8".

$R2$ Results from Baseline Rule 2 are the third most authoritative. The clear structure of two integers and a addition sign allows for the computation of the total Gleason grade. E.g: "Gleason (4+3)".

$R3$ A valid Baseline Rule 3 is the least authoritative of the rules, only taking into account the very first occurrence of a numeric value in the identified text segments. E.g: "Gleason grade 7".

A discussion and evaluation of the application of the developed rules is given in the following section. Regarding the technical aspects of this work, the dataset was stored as a Microsoft Excel document and the algorithms were written in Visual Basic for Applications (VBA) programming language with a Regular Expressions library. VBA is an event-driven programming language and in this case was used with Microsoft Excel allowing full control of spreadsheets and their data. Although the running time for the developed algorithms was slow, automatically populating the Excel spreadsheets with their results proved fruitful; it allowed a thorough debugging, fast inspection and analysis.

3.2.1.6 Algorithm Results and Evaluation

The results of the application of the above hierarchical rules ($R0$ through to $R3$) are shown in figure 3.3. The figure shows the individual and cumulative percentage of reports classified as having satisfied each hierarchical rule. Records

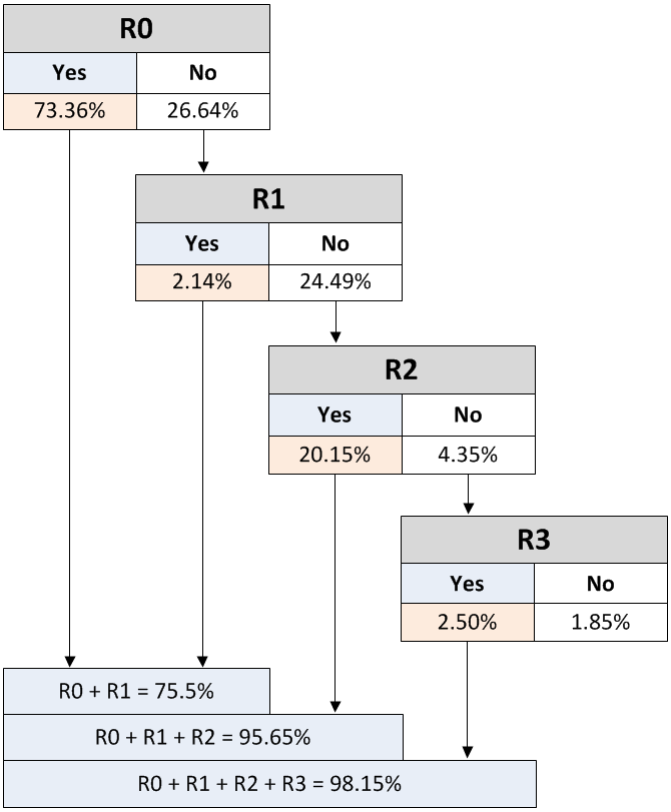


Figure 3.3: Results of the application of hierarchical rules for information extraction. When R0 = "Yes", the R0 rule was satisfied, and when R0 = "No" it was not possible to find a Gleason with that rule.

that do not satisfy a rule are further evaluated with the next most authoritative rule. The order in which the rules are applied is relevant because of the initial feasibility analysis and understanding of the rules' authoritativeness.

The overall coverage for each rule, and cumulative coverage are illustrated in figure 3.3. When rule $R0$ is used alone, it allows the extraction of Gleason grades for 73.3% of reports. However, when rules $R0$, $R1$ and $R2$ are used together, a total of 95.7% reports are covered. There is a 20% gain in coverage by using the first three rules as opposed to the first two alone. A further gain of 2.5% in coverage is achieved if all rules are applied.

Nevertheless an evaluation of the true accuracy of the rules is needed to ensure that they are indeed retrieving the correct Gleason grade. This is discussed below.

Evaluation of the Algorithm

In order to evaluate the results, a number of reports were selected at random and manually inspected. Table 3.2 presents the overall results by hierarchical rule.

	Rule				
	$R0$	$R1$	$R2$	$R3$	Total
Valid	50	14	22	12	98%
Invalid				2	2%
Total	50%	14%	22%	14%	100%

Table 3.2: Results of the evaluation of the hierarchical rules.

A randomised set of 100 unique reports was selected: 50 from the subset identified with $R0$, 14 from the subset $R1$, 22 from $R2$ and 14 from $R3$. The number of selected reports is based on an even distribution of the first 48% for each rule (baseline of 12 reports), and the remaining 52% are distributed based on the overall percentage of reports identified by each hierarchical rule (for example, overall $R0$ has 73% coverage on the original dataset which translates into 38 reports, then adding the baseline number 12 giving a total of 50 reports for $R0$). There is also an even distribution of the years in which the reports were produced, between 2003 and 2010.

The above results show two incorrectly classified reports for rule $R3$ and no incorrect reports on any of the other rules. This should reassure that $R0$, $R1$ and $R2$ provide the most reliable results, yet $R3$, also because of the naive nature of this rule, should be used with caution.

Classification as Benign or Malignant

It is established that a Gleason grade positively identifies malignancy. However, the absence of such a grade does not confirm or deny this. It is therefore essential to investigate the reports where no Gleason grade was found, or where a degree

of uncertainty exists (i.e. reports not covered by rules $R0$, $R1$ or $R2$).

Because of the relatively small number of reports, this inspection was performed manually. Indeed the vast majority of reports identified by $R3 = 0$ are benign prostate hyperplasias and were classified as "benign". There are also reports where histopathological grading was not possible due to the small size of the focus.

The manual classification into "benign" and "cancerous" together with the extracted Gleason grade resulted in two new fields, appended to the Histopathology dataset. The first field indicates whether the biopsy result is a malignancy, and the second, discretises the Gleason grade sum into low-grade, high-grade or inconclusive.

The number of reports classified a high-grade is 4,654 (91.5%), whereas 335 (7%) reports were classified with a low-grade and 94 (1.9%) were not graded. Further results will be given in the next sections.

Examples of reports classified as malignant include: diagnoses of adenocarcinoma but no Gleason given, focal adenocarcinoma, metastases, lymph nodes, or difficult to grade but cancerous cells detected. Examples of benign reports include: no evidence of malignancy in the prostate, hyperplasia, post-treatment biopsy showing no malignancy.

There are also reports where a biopsy of bladder tissue is also present. This does not affect most (malignant) reports as the Gleason grade is only present for prostate cancer tissue. Reports affected by this were accounted for during the manual inspection. Such cases are also identifiable by inspecting the TNM staging, when present, as depicted in the next section.

Extraction of Tumour Staging

The developed algorithm was enhanced to extract TNM staging from histopathol-

Parameter	**T**	**N**	**M**
Description	Primary tumour	Regional lymph nodes	Distant metastasis
Values	TX, T0, T1, T1(a,b,c), T2, T2(a,b,c), T3, T3(a,b), T4	NX, N0, N1	M0, M1, M1(a,b,c)

Table 3.3: Simplified TNM Staging parameters and values, adapted from [1].

ogy reports. A simplified list of all possible TNM staging parameters, its values and a description are given in table 3.3.

This algorithm retrieves 1) any occurrences of "pT" and 2) any of the possible "T" stages, as listed in table 3.3. When an instance is found the algorithm iterates through the subsequent characters until it finds a break (defined here as a non alphanumeric character). The algorithm will also ignore the "history" section and also a "scan" subsection which includes information previous to the histopathological analysis. This approach resulted in the positive identification of TNM staging for 1,090 reports (21%).

Outliers were found when the algorithm failed to identify when to break, i.e. a suffix of alphanumeric characters continues beyond the TNM stage (for example, "pt3aaatlleast" or "pt2ciisaappropriate"). Such outliers were inspected and removed using filtering. An alternative aproach would be to limit the length of the TNM string to its maximum (three characters and the pathological prefix "p") but this approach would misclassify cases, such as "pt2and" for "pT2a" rather than the correct form "pT2".

Overall, the algorithm identified the following TNM stages: "T1" - 358 reports, "T2" - 377 reports, "T3" - 333 reports, "T4" - 21 reports, "NX" - 516 reports, and "MX" - 632 reports some of which are cumulative (i.e. reports may be of the form "pT2b, NX, MX").

Interpretation and Evaluation of TNM staging

When more than two TNM stages are identified, only the one with the poorest prognosis is used as the final tumour staging. In some cases, a histopathology report may contain two specimens, one from the bladder and a second from the prostate. In this situation, should the bladder have a higher TNM stage, the latter will be used as the final TNM stage. Bladder tumours should be identified or removed after linking with other data sources, which will confirm the primary tumour. The maximum number of potential bladder biopsies identified is 11% of the total number of records identified with a TNM staging. This was determined by checking for at least one occurrence of "bladder" in the report using regular expressions.

Overall the algorithm was able to extract the TNM staging correctly and highlighted outliers (extreme cases, n=18), which were manually fixed. Despite the encouraging results using the developed algorithms, the overall number of reports with a pathological TNM staging was low and therefore its use is limited. Should histopathologists consistently report TNM staging, under the same or a similar report structure, the approach presented here would allow the retrieval of the vast majority of cases.

3.2.1.7 Intraobserver and Interobserver Analysis

The Intraobserver and Interobserver Analysis is an interesting analysis that not only inspects the validity or accuracy of the histopathological findings, but also, the completeness and structure of the reports, and possibly, the validity of NLP algorithms, which has not been explored. Any conclusions from this analysis are based on the results of the algorithms presented and the cohort of known prostate cancers.

This analysis consists of identifying and assigning unique IDs to the histopathologists who wrote the reports (using regular expressions) and then grouping reports by ID of the first author (note that reports are often written and/or double checked by more than one author). In this task, the authors anonymity is ensured by the unique ID number, created for this purpose.

A total of eight author groups were created:

- six groups of individual histopathologists with a relatively high number of reviewed reports over time: groups $C1$ (n=929), $C2$ (n=191), $C3$ (n=448), $C4$ (n=420), $C5$ (n=864), and $C6$ (n=686);
- one group of reports whose histopathologists were in advanced training, i.e. specialist registrars: SpR (n=431);
- and a final group, containing all remaining reports: $Others$ (n=1114).

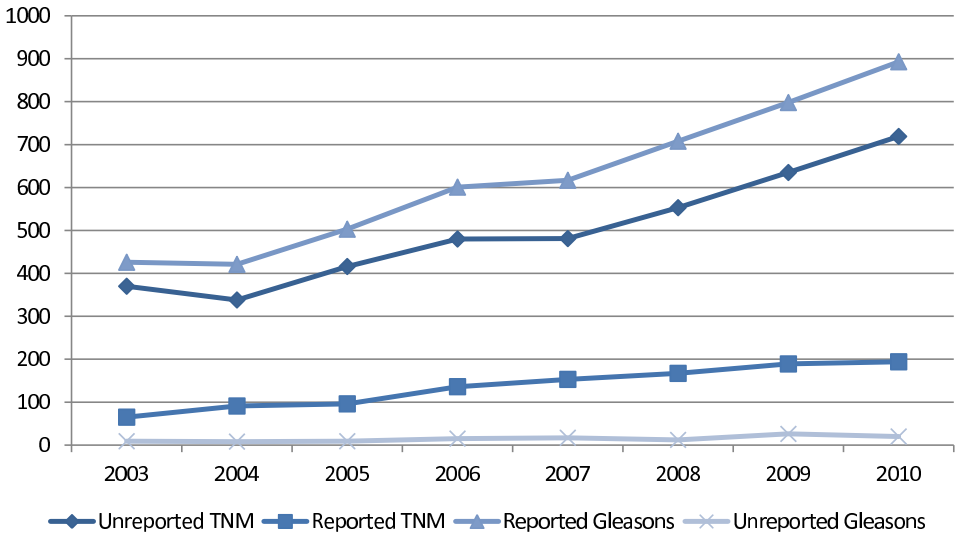


Figure 3.4: Frequencies of reported and unreported TNM staging and Gleason grades from 2003 to 2010.

Figure 3.4 shows the underlying trends of histopathology reports on prostate cases (frequencies of reports against year); it reveals the steady increases, particularly in recent years, of the reported Gleason grades and the unreported TNM stages. The number of unreported Gleason grades is mostly constant at $\sim$ 20 reports a year. There is an increase in the total number of cancerous biopsies analysed by the histopathology department of the NNUH but it is interesting to observe that although the number of reported pathological TNM staging shows a slow increase, a much steeper increase is observed in the number of unreported TNMs. It is not clear, at the time of writing, why histopathologists show increasingly little interest in reporting pathological staging.

Another interesting set of analyses are those where a relation is made between the different histopathologists' groups, the reported Gleason grades, and TNM staging, over time. Figure 3.5 shows the percentage of high-graded Gleason grades (7-10, n=4654) reported among the eight histopathologists' groups; it shows an overall increasing trend of high-grades being reported. It is interesting to note that a particular consultant, $C3$, has consistently graded the least number of high-grade Gleasons (from 2006-2010) and that the SpR (specialist registrars) group has graded the most high-grade Gleasons in the last two years including 100% (n=50) in 2009. Apart from the SpR group, only one other, $C2$, reported 100% high-grades in 2006 (n=2) and 2008 (n=59).

When observing low-graded Gleason grades (2-6, n=337), figure 3.6 shows the percentages among the eight histopathologists' groups. The low-graded Gleasons show a seemingly symmetrical trend to the previous figure, where the total number of reported low-grades is mostly on the decline until 2009.

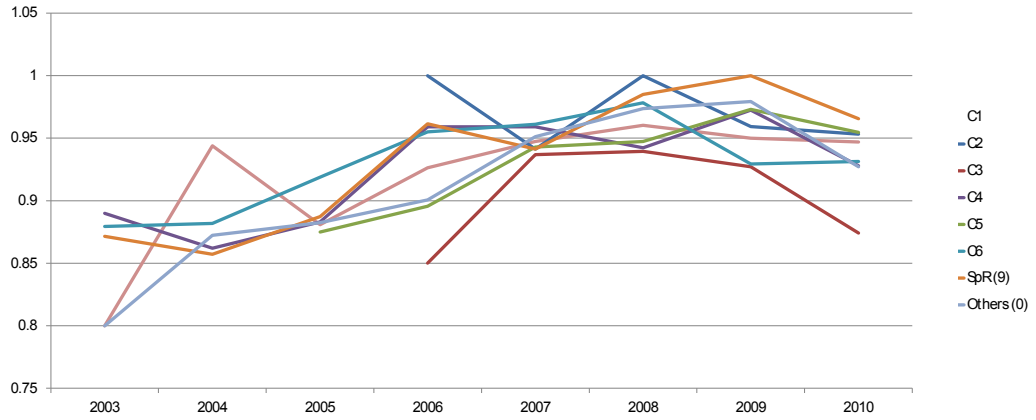


Figure 3.5: Percentage of high-grade Gleasons reported among the different histopathologists' groups.

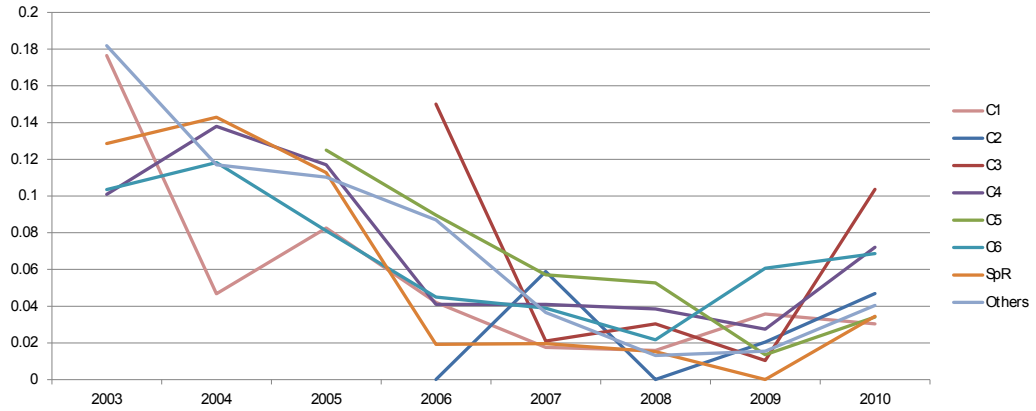


Figure 3.6: Percentage of low-grade Gleasons reported among the different histopathologists' groups.

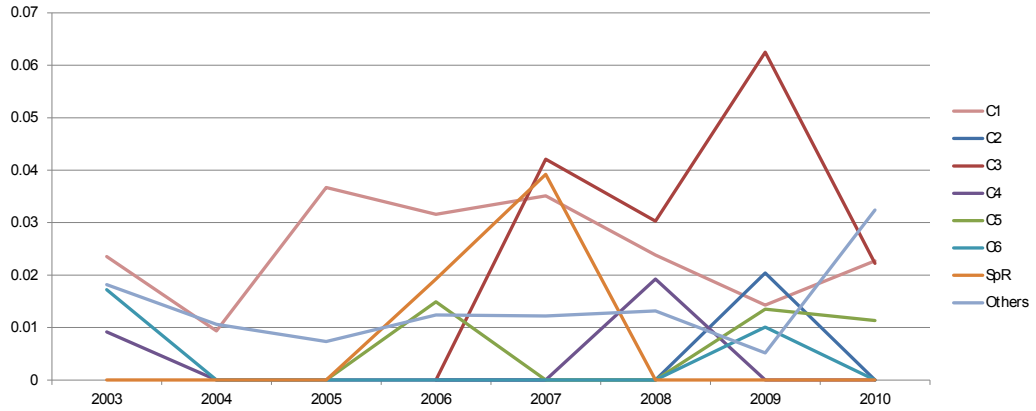


Figure 3.7: Percentage of non-reported Gleasons among the different histopathologists' groups.

Figure 3.7 shows the percentages of non-reported Gleason grades among the eight histopathologists' groups. Because of mostly low numbers, the lines are more dispersed, but it is still interesting to note the differences in non-reporters, particularly among the SpR group, only non-reporting by 3.9% in 2007. Overall it is not possible to say that non-reporting is either on the increase or decline over time. It is also important to note that there are other reasons, unknown here, as to why particular histopathologists have a higher rate of reports with no Gleason. One reason may be a tendency of a a particular pathologist to analyse and report on more difficult specimens.

Lastly, when observing the number of reported TNM staging, figure 3.8 shows the percentages among the eight histopathologists' groups. A steady decline in TNM reporting in the SpR group since 2007 is observed. It is also possible to observe that two consultants, $C1$ and $C4$, were consistently better at reporting TNM than any of the other groups. However, it is not possible to say whether these consultants report TNM because they have received different training, have a different reporting technique or style, or whether they were given specific cases where TNM reporting was needed.

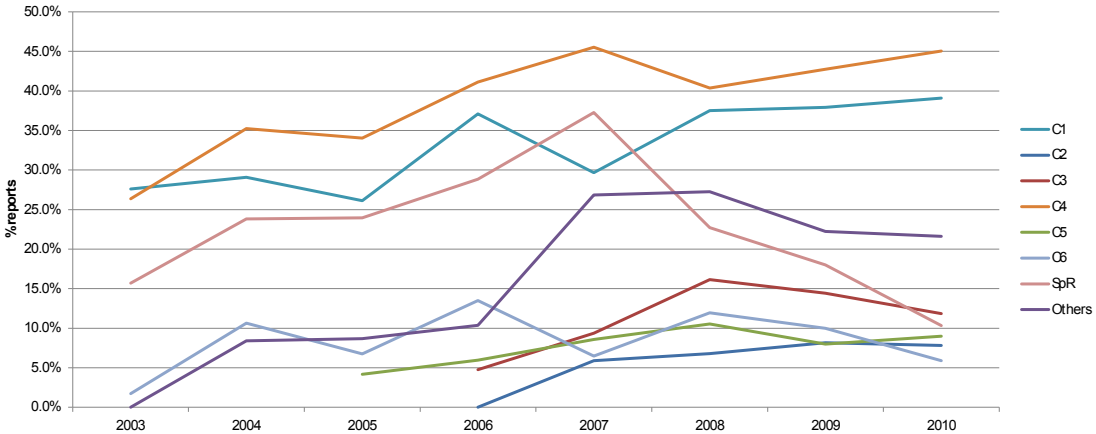


Figure 3.8: Percentage of TNM tumour stages reported among the different histopathologists' groups.

Nevertheless, the intra/interobserver analysis points out that the differences in reporting authors should not be ignored. Indeed, should these numbers be made available more easily and in near real-time, they would inform histopathologists how their reporting compares with their peers, and alert them when significant changes or deviations ocurr.

Another interesting observation of the intra/interobserver analysis was that there was no particularly signigicant relation between the way in which histopathologists write the reports and the accuracy of the algorithms.

When observing typographical errors in reports, the typo "Gleeson" mentioned earlier, was found in a total of 13 reports, all belonging to the same first author, $C1$. It is important to note that the reports containing this typo all had different second authors. When inspecting all reports for their total number of words and characters, author $C1$ had written the most words (163 ± 94), i.e. 34 words above the mean and 18 words above all other peer groups. It could be argued that authors writing the most extensive reports would be more prone to typos than their peers, but on the other hand they may produce more comprehensive reports. A more detailed analysis of typographical errors, grammar and style would be required for conclusive assertions to be made. This was, however, an interesting finding that illustrates how certain systematic errors, otherwise overlooked, can be identified using the proposed methods.

Intra/interobserver analysis can be a useful resource for the evaluation of the NLP algorithms as it could highlight a particularly difficult writing style for algorithms to interpret, or even systematic errors and outliers. With the exception of the "Gleeson" typo, there was no particular relation between structures, styles or errors found in the above exercise.

3.2.1.8 Discussion and Conclusions

As described earlier in this chapter, and commonly reported in the literature [118], there are two major challenges for text mining and natural language processing techniques: standardisation and structure, and measuring success.

Standardisation and Structure

The lack of a report structure and/or standardisation is a problem that pervasively affects medical data and, in particular, narrative reports. This challenge was encountered in this chapter but did not pose a significant barrier to extracting the Gleason grades from prostate cancer histopathology reports. Although there were different ways in which reports were written, the nature of the problem domain (particularly the existence of a Gleason grading system) greatly facilitated this exercise. This reassures the importance of a thorough understanding of the problem domain, and a detailed investigation of the ways in which histopathologists report cases.

Overall it is possible to say that, despite the multitude of ways in which natural language was used in the histopathology reports analysed, an underlying structure was present, and natural language processing techniques can be used to successfully extract information from the reports.

Retrieving tumour stages (TNM), however, was more challenging in terms of contextualisation. TNM staging is common not only on prostate cancers but other cancers and therefore bladder tumours, which may be present in the cohort of prostate cancers, may be incorrectly identified. This issue can be dealt with at a later stage when further information from other data sources is linked in and other cancers and comorbidities are easily identified.

Another important part of the exercise presented in this section was the classification of reports as either malignant, benign or inconclusive (i.e. the overall

conclusion of the report). This classification enabled the highlight and correction of issues where a TNM value is present but there is no malignancy. Indeed this can partially solve the problem of multiple tumours.

Measuring Success

Measuring the success of a text mining or natural language system is one of the most difficult challenges. Current evaluation methods are based on manual clerical review and are highly dependent on the reviewers' knowledge [102; 118].

Nevertheless, as it was identified in this chapter, it is possible to extract information from the reports, which help in the validation of the accuracy of a classification rule or other information retrieved. These were still complemented by a manual clerical inspection to ensure validity.

Furthermore, a resulting dataset built with the proposed or similar techniques, should only be used in combination with other data sources, so as to increase its accuracy as well as to ensure the correct contextualisation. Even then, data sources where coded information is entered by trained staff or migrated from other databases may not always be reliable. A particular study on postoperative complications showed that a natural language system can outperform standard coding [118]. Another study on NLP analysis of free text medical records [118] also revealed better performance (higher sensitivity and lower specificity) for NLP techniques than a coded patient indicator. It is therefore important to be aware of such issues when linking data across sources.

Another analysis that may help to validate and highlight issues in NLP algorithms is one of the writing style and trends in which histopathologists write reports.

Conclusions

Overall the work carried out in this section reassures that natural language processing techniques remain *ad hoc* exercises, heavily bound to the context of their

production. It is arguable that heuristic rule-based models, tailored to a specific domain, are reliable and accurately extract information from text. It would also be interesting to evaluate and compare different NLP approaches and models. However, this proved difficult to achieve and it remains a current topic of research and further work [117].

The work presented here pointed out some additional benefits of extracting information from text reports. In particular, it enabled the intra/interobserver analysis in which interesting results were found. Further work is still required to better interpret and make sense of those analyses. The extraction techniques here presented would be facilitated, however, if reports were annotated or coded. Indeed it would be of interest to develop natural language writing systems where users can easily highlight key terms (or terms are suggested to them by the computer) to facilitate identification of positive or negative findings. Indeed annotations have already been pointed out as one of the ways of improving the quality of reporting as well as the accuracy of natural language processing techniques [106].

The objectives set for the information extraction algorithms were achieved and all patients in the prostate cancer cohort now have Gleason grades, and when available, pathological TNM staging, assigned to them in a cannonical form.

3.2.2 Data Editing and Imputation

Another commonly reported challenge when using medical databases is handling errors, outliers and missing data [20; 80; 92; 119]. In large medical databases, almost every patient-record is missing values for some feature, and almost every feature is missing values for some patient-record [20]. In some cases, erroneous or missing values may be estimated and automatic imputation methods or other techniques used. Common approaches to dealing with missing or erroneous data include substituting them with most likely values, or replacing their values with all possible values for that attribute [20]. Other approaches, such as training a neural network to predict missing values, are possible, yet dangerous [80]. The problem of missing data together with methods to handle them have been studied by statisticians, particularly in regards to survey data. *Data editing* and *imputation* are well established terms in statistics and are discussed in detail below (in section 3.2.2.1).

This section discusses the issues of data preparation with examples from the case study, and it introduces a method for estimating continuous, ordinal data, such as age. Data preprocessing methods and techniques presented here are based on the prostate cancer biochemistry datasets, namely the Prostate Specific Antigen (PSA) dataset. They illustrate the typical challenges encountered, and techniques used in datasets with attributes of a similar nature.

3.2.2.1 Background on Data Editing and Imputation

Official statistics institutes and other similar entities providing high-quality statistical information need to be as up-to-date and as accurate as possible [88]. Traditionally, such entities collect data in the form of surveys (such as a census) and administrative data, and these inevitably contain errors [88]. *Statistical data*

editing emerged in the 1950s as a process of improving or correcting the effects of erroneous data. Techniques were developed and have been extensively applied since the mid-1950s [88; 120]. Over the years, several studies showed that it is not necessary to correct all data in every detail and that such, "over-editing" often leads to little improvement in the overall results. In fact, errors caused by incorrect data are acceptable as long as they are small in comparison to the sampling error [88].

Missing data can be seen as a form of erroneous data and one that is easier to identify yet often more difficult to correct [88]. Examples of sources of missing data encountered in early statistical exercises are questions not answered by respondents either because of a lack of understanding or refusal [88]. However, missing data has a more complex set of reasons for existing in medical databases. A brief description of such reasons was given as "value(s) accidentally not entered, or purposely not obtained for technical, economic, or ethical reasons"[20]. Also, the integration and use of retrospective clinical data can introduce additional sources of missing data, often difficult to understand. Reasons for collecting information often change over time and are poorly documented in legacy, routinely collected, database systems (as seen in the previous chapter). An understanding of the sources of erroneous or missing data would be benefitial to the development of data editing techniques to cope with them.

The process of replacing missing data with substituted values, which can be used in combination with other editing techniques, is called imputation [88]. Some authors, however, see imputation as the process of replacing values, regardless of whether they are true missing values or identified measurement errors [88; 121]. In this thesis the latter definition of imputation is used.

Data editing has been described as techniques by which statistical data are checked and made correct with respect to both individual values and "mutual compatibility between the values for different variables"[122]. Similar to other

preprocessing steps and techniques (for example, those in a DMKD process), data editing can consume up to nearly half of the total resources spent on a project [123].

The purpose of data editing, as defined by Granquist, is threefold [124]:

- it creates the foundation for the improvement of statistical data in the future,
- it produces information about the quality of statistical data,
- it cleans the data errors (they should be analysed for their overall importance).

Five editing categories have also been defined [122]: *Completeness edits* (for example, missing data), *Validity and range* edits (where only certain codes or ranges of values are permissible), *Consistency edits* (comparison of different answers from the same record to check logical consistency), *Historical edits* (for example comparison of response for a survey with a previous response), and *Statistical edits* (checks based on statistical analysis of data where suspicious values are identified, this could include historical data). Certain datasets and problems may require the use of more than one category.

Regarding imputation, three main categories have been defined by statisticians [122]:

- Logical (or deductive) imputation, used when a reliable, explicit solution exists given appropriate assumptions (deterministic imputation);
- Model based imputation, where a model is fitted to the data (including probabilistic approaches);

- Real donor imputation, where the imputed observation value is "donated" from another respondent.

Furthermore, imputation rules should: be derived from the corresponding edit rules, satisfy all edits by changing the fewest possible items of data; and maintain data structure and frequency [125].

Another issue that often needs to be tackled as part of edits or imputation methods are outliers. An outlying observation (outlier) is "one that appears to deviate markedly from other members of the sample in which it occurs" [126]. Measurement errors or heavy-tailed distributions often result in outliers. Methods to minimize the effects of outliers include trimming and Winsorisation [122]. Such techniques can work alongside imputation and editing techniques, but have also been described as imputation rules [127]. Trimming consists of identifying and stripping the extremes, whilst Winsorisation (sometimes Winsorising) is the process of moving extreme values towards the centre of the distribution [122].

The next section introduces the PSA dataset, its attributes and values, and the motivations for the development of a data editing technique.

3.2.2.2 The PSA Dataset

Prostatic Specific Antigen (PSA) is a widely used blood test for the early detection of prostate cancer. Men with abnormal PSA concentrations usually undergo further examinations such as the digital rectal examination (DRE), transrectal ultrasonography (TRUS) and prostate (needle) biopsy for a diagnosis. Further discussion and background on the PSA is given later, in Chapter 4, where results are analysed. This section is concerned with the preprocessing methods of potential erroneous values.

The Prostatic Specific Antigen (PSA) dataset is similar to other Biochemistry

datasets in respect to attributes, their syntax and, to some degree, semantics. This dataset was originally retrieved from the Biochemistry data warehouse at the NNUH, where other biochemistry readings are stored in a similar fashion. Upon consultation with the head of pathology at the NNUH, it was understood that the local laboratory is responsible for most, if not all, PSA tests carried out in Norfolk. However, there may be a small number of instances where tests are carried out in other laboratories. This makes the PSA dataset more valuable in that it allows the inspection of PSAs not only for the cohort of cancers, but for all others, including potential "screening" patients.

The complete PSA dataset collected in this research contains over 150 thousand readings, yet here, a subset was used and should be considered sufficient to highlight and generalise missing and erroneous values. The dataset used here covers a five year period (from 2003 to 2007), contains 80,738 records (33,013 unique hospital numbers), and the following key attributes:

- Hospital Number ($m = 3.1\%$) - Patient identifier;
- Clinical History ($m = 4\%$) - Free text field describing clinical history;
- Date of Entry ($m = 0\%$) - Measurement date;
- Time of Entry ($m = 0\%$) - Measurement time;
- Test Data ($m = 5.3\%$) - The biochemistry test value (eg. PSA reading);
- Age at Episode Date ($m = 0\%$) - Patient age when blood sample was taken.

A small percentage of hospital numbers was missing. It could be possible to recover hospital numbers using a patient's other attributes, including history and dates, but this process may still not provide accurate results and would likely result in over-editing. It would also be important to link the records to other

clinical data in the future, which is why other methods would not work here. Records with missing hospital number were therefore removed from the dataset.

The *Clinical History* attribute contains 4% missing data, calculated as values with less than two characters. The remaining 96% of records, however, contain varied information: regarding drugs the patients were taking (for example "pt on Zoladex"); diagnoses (for example "Advanced prostate Ca"); symptoms (for example "Painful feet plus problems with Nocturia"); and references to previous tests or procedures (for example "PSA retest", "pre-op"). These examples illustrate how the attribute is not used consistently with regards to semantics, which make it difficult to understand and derive information. Upon visual inspection, syntax and typographical errors seem common and systematic. This field was not used in this section but kept in the ODS should further investigations be required.

The date and time of entry attributes are automatically filled in by the biochemistry software and so there is no missing data in them and the values, upon inspection of descriptive statistics, are not expected to be erroneous.

The *Test Data* attribute, however, shows that 5.3% of its values are missing. This attribute stores the PSA, a numeric value with one decimal point. The data type and format are not consistent and valid, non-numeric values such as "< 1", are common. It is not possible nor trivial to recover PSA values because of their broad range and sparseness (in this dataset the range is from 0 to 23,750 with mean 30 (SD 284)). Probabilistic methods for automatic imputation would be difficult, but may be possible upon normalisation. Because future analyses in this research will not always group and summarise data, missing or erroneous fields should be identified and removed. Becasue accepted PSA values include 0, the values identified for removal were: missing values, text values, and negative values. Values indicating an approximation (such as "> 100") were corrected as follows:

- value "$< .1$" modified to 0.05 (in n=8,532 cases)
- value "$< .2$" modified to 0.1 (n=7)
- value "$< .4$" modified to 0.1 (n=1)
- value "< 1" modified to 0.5 (n=2)
- value "> 100" modified to 100 (n=5)
- value "$> 10,000$" modified to 10,000 (n=12)

Regarding the *Age at Episode Date* (age) attribute, no missing data was found. However, upon an observation of its range, $[-25; 127]$, erroneous values and outliers were suspected. This field is indeed the result of clerical data entry in a busy laboratory and it is not consistently retrieved nor validated against hospital administration systems. In fact, the laboratory often takes samples from patients who are not registered with the hospital or even treated there. Given that there is no date of birth field available in this dataset, it is not possible to accurately determine whether a patients age at episode date is correct. It is possible, however, to link this dataset to other hospital sources on patient number, but that will restrict the working dataset to the number of matching links between the two. Indeed for mismatches it would not be possible to determine the correctness of the values, and hence, the quality of the age attribute. Because the interest in the PSA dataset is greater than just the subset of the prostate cancers identified in the ODS, it is important to have reliable key demographic attributes such as age. This led to further investigations and the development of an algorithm to solve the *Age problem*, described below.

3.2.2.3 The Age Problem

Upon deleting records with identified errors in *Test Data* and *Hospital Number* the total number of records in the dataset was 73,902.

A preliminary assessment of the age values together with an understanding of their origins revealed that the *Age at Episode Date* attribute had poor reliability. It is essential to ensure the age values are correct for the validity of analyses using biochemistry data. It is possible to retrieve the correct patients age by linking to other datasets where a date of birth attribute exists. This is, however, unfeasible for patients not seen at the hospital (i.e. without any further records available to us). Yet a subset of linked, positive matches may be used to test the accuracy of any new methods developed.

When plotting a frequency distribution of the age attribute, missing and possible erroneous values were detected in the data. This is illustrated in figure 3.9. The latter shows erroneous, outlier, values that do not seem to follow a normal distribution. This is particularly obvious for age 0 (0.3%) and any negative age values ($< 0.1\%$). A well-known normality test, the Kolmogorov-Smirnov, failed for the age attribute. Indeed it is expected that, given its age range, a prostate cancer dataset is negatively skewed. This is confirmed by the skewness test of asymmetry that scores -1.048. Further descriptive statistics are given in table 3.4.

Erroneous outlier values (i.e. age ≤ 0) were identified and their impact on the normal distribution assessed. Table 3.4 shows the potential impact of cleaning outliers from the age attribute on its normal distribution. Records with other erroneous attributes had already been removed.

This analysis, although sufficient to highlight the extent of outliers, does not help to determine whether the values are indeed correct. However, the PSA dataset

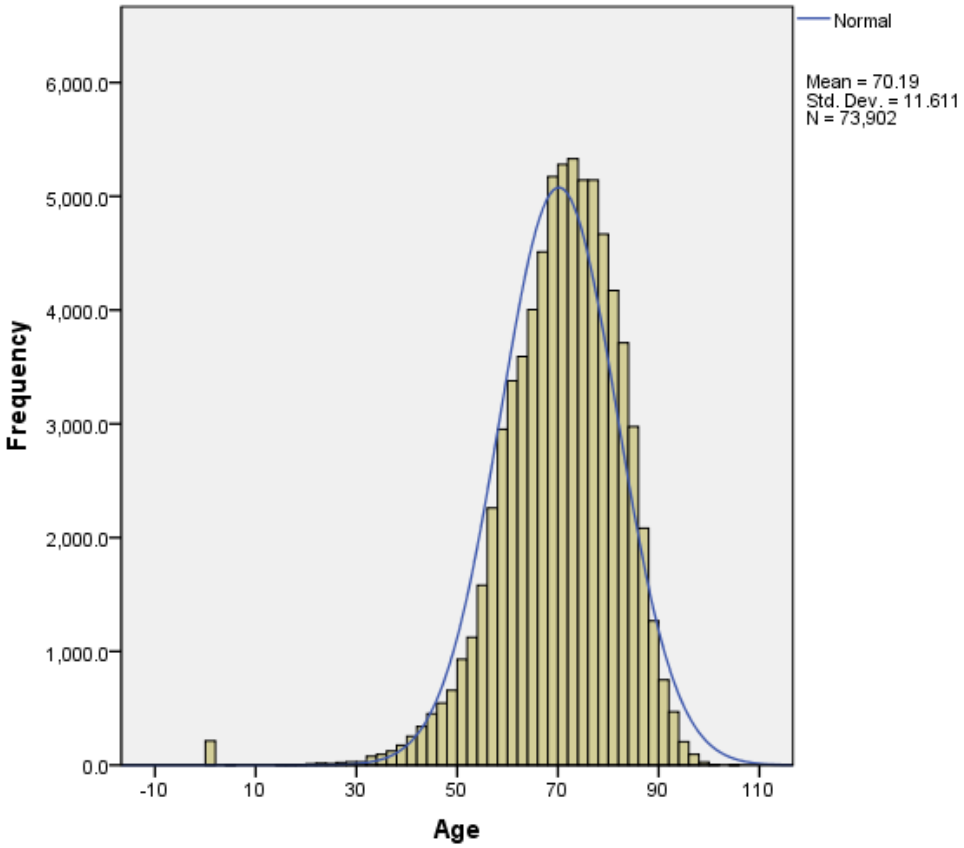


Figure 3.9: Histogram showing frequency distribution of age in the PSA dataset.

Statistic	Age (raw)		Age (clean)	
	Value	Std. Error	Value	Std. Error
Records	73,902		73,804	
Range	152		96	
Minimum	-25		5	
Maximum	127		101	
Mean	70.19	.043	69.57	.040
Median	71		70	
Mode	72		72	
Std. Deviation	11.611		10.805	
Variance	134.813		116.758	
Skewness	-1.048	0.009	-.469	.009
Kurtosis	3.846	0.018	.369	.018

Table 3.4: Statistics showing the impact of cleaning the age values.

contains age values recorded each time a patient carried out a PSA examination and on average, there are 2.4 (SD 1.1) records per patient. Working with other records from the same patient may help determining whether age readings are correct. The *Age Problem* is then defined as the evaluation of integrity among multiple age readings over time and any subsequent corrections. A solution algorithm could be seen as consistency and historical edits resulting in the imputation of correct, best estimate, values.

3.2.2.4 Age Integrity Check

The first analysis carried out relied on the integrity of pairs of patient episodes (i.e. a chronological sequence of measurements). Pairs of episodes are ordered chronologically by date of entry. For each pair it was possible to compute the difference between expected and observed chronological events using the date of record entry and age at episode. The results of this analysis are listed in table 3.2.2.5. Two methods for computing the integrity were investigated. A first, less accurate, was calculated using year and age as integers and a second, with higher sensitivity, was calculated in days.

Let r_1, r_2 be a pair of age values and v_1, v_2 a corresponding pair of date entry values where $v_1 \leq v_2$. The integrity of a sequential pair of records $(r_1, v_1; r_2, v_2)$ is satisfied in the integer space (in years) iff

$$(v_1 - v_2) - (r_1 - r_2) = 0. \tag{3.1}$$

The above can be improved to increase accuracy by measuring days instead of years. Then the difference between the age values and time interval between two episodes is valid iff

$$0 \leq (v_2 - v_1) - (r_2 - r_1) \leq 365.25. \tag{3.2}$$

The average year length accounting for leap years is 365.25. This is important for accuracy over long periods of time. The results of the above computations are given in table 3.5 and represent an initial estimate of the number of possible erroneous values in the data. The most reliable results, i.e. most consistent, are given in of course days.

	Years		**Days**	
Integrity Check Result	Frequency	Percent	Frequency	Percent
Possible Inconsistent Record	14,333	19.4%	237	.3%
Consistent Record	59,569	80.6%	73,665	99.7%
Total	73,902	100%	73,902	100%

Table 3.5: Results of the age integrity check for consistency.

3.2.2.5 Assessment using Linked Data

A data quality assessment using date-of-birth (DOB) in the PSA cohort was carried out to quantify the magnitude of erroneous records. This was only feasible for those patients that link to other hospital sources. It was possible to link 21,966 patients from the PSA dataset with the administration dataset. The date of birth was calculated in the PSA dataset using *Date of Entry* and *Age at Episode Date*. The difference between the two was computed in years. The linkage exercise revealed that 0.5% (n=104) were erroneous.

This exercise showed that records from patients that are simultaneously on the administration database (i.e that have at least one inpatient record with prostate cancer diagnosis) and on the PSA cohort are correct 99.5% of the time. The latter may not be true for records that do not link to the administration system,

as the laboratory would not be able to double-check their age.

Table 3.6 shows the number of invalid records after linkage to the administration system (n=104) and from those, the number of cases highlighted by the integrity check (n=83, 79.8%).

		Frequency	**Percent**
Validation	Invalid	104	.5%
	Valid	21,862	99.5%
	Total	21,966	100%
Cases highlighted by the Integrity Check	Highlighted	83	79.8%
	Missed	21	20.2%
	Total	104	100%

Table 3.6: Results of assessment using linked data. True error rate for age fields.

3.2.2.6 The Age Problem Algorithm

The above results led to the development of an algorithm to estimate patients' age. This is also important because future studies need to be carried out using other biochemistry tests (datasets with the same specifications and issues) for which we cannot recover patients age at episode in days.

Let x be the number of days since a patient was last seen, and y the corresponding date at episode in days. Given a sequence of n pairs $(x_1, y_1), \ldots, (x_n, y_n)$ where $0 < x_1 < \ldots < x_n, x_i \in Z$ and $y_1, y_2, \ldots, y_n \in Z$ is there a base number z (the age at which patients were first seen) such that

$$\lfloor \frac{z + x_i}{k} \rfloor = y_i \tag{3.3}$$

$$\text{for } i = 1, 2, \ldots, n.$$

If the above is true, the set $\{(x_i y_i) \mid 1 \leq i \leq n\}$ is k consistent. A simple algorithm to determine this is based on the observation that if

$$\lfloor \frac{z + x_i}{k} \rfloor = y_i \tag{3.4}$$

$$\text{then } ky_i \leq z + x_i < k(y_i + 1).$$

$$\text{that is } ky_i - x_i \leq z < k(y_i + 1) - x_i.$$

.: There is a base number z iff

$$\cap \ [ky_i - x_i, \ k(y_i + 1) - x_i[\ \neq 0 \tag{3.5}$$

The purpose here is to find the largest k consistent sequence where $k = 365.25$. Let z_0 be the lowest value, and z_1 be the highest value such that

$$\lfloor \frac{z_0 + x_i}{k} \rfloor = y_i \qquad \lfloor \frac{z_1 + x_i}{k} \rfloor = y_i \tag{3.6}$$

$$\text{for some } i$$

We then seek $z_0 \leq z \leq z_1$ such that

$$N_i \ = \ \mid \{i : \lfloor \frac{z + x_i}{k} \rfloor = y_i\} \mid \tag{3.7}$$

is maximised.

This is found by iterating through $m = (z_1 - z_o)$ values of Z and evaluating N_i for each one. This is a pseudo-polynomial algorithm $O(n * m)$.

The algorithm was written in Python programming language and integrated with SPSS Statistics version 16 programmability extensions. The performance of the algorithm is evaluated in the next section.

3.2.2.7 Evaluation and Performance of the Algorithm

In order to assess the algorithm's performance the age values were computed on the set of linked data (test set, n=21,966 total records) where the true patients' age is available from another data source.

The algorithm computed the correct (exact) value 53.2% of the time and was one year off 46.5% of the time. Therefore, the algorithm computed the correct age within one year 99.7% of the time, in a set of 6,076 patients (test set). This is illustrated in figure 3.10. The remaining 0.3% of records (not visible in the figure) were incorrectly computed due to outliers or missing values and can be manually corrected.

However, the algorithm's performance was poorer when working with individual episodes (rather than by patient). Only 77.2% were correct within one year and 99.9% were correct within two years. The frequencies are illustrated in figure 3.12. Likewise, the remaining 0.1% of records (not visible in the figure) were manually corrected.

Other methods to estimate age were investigated using the mode, mean and mean of computed year of birth as possible estimations for sets of patient records. These approaches achieved low accuracies (between 5% and 6.1%) on the test set of linked data.

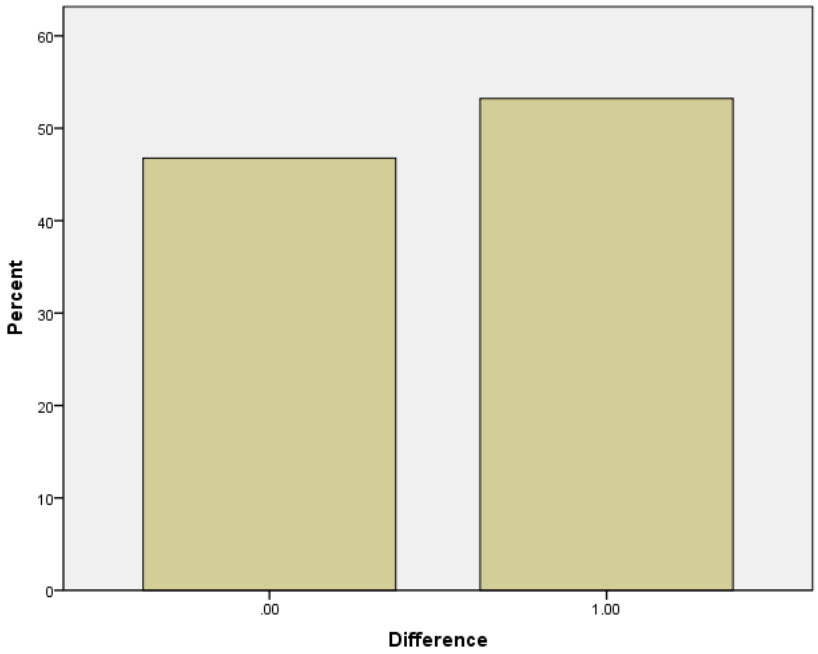


Figure 3.10: Frequency distribution of the accuracy of the algorithm (difference in years) for individual patients on the test set.

3.2.2.8 Discussion and Conclusions

The primary objective of the developed algorithm was to compute the correct patient's age at first presentation. Subsequent values can be estimated by adding the elapsed time intervals. Therefore, the AP algorithm introduced here can compute the patient's age correctly (within a year) with an accuracy of 99.7% which is considered sufficient for the purpose of future work.

The AP algorithm was applied to the main PSA dataset (n=73,902) and extremes were removed. Figure 3.12 shows the frequency distribution of the age attribute after erroneous values were fixed by the algorithm. Table 3.7 compares key distribution statistics of the age attribute against values corrected by the algorithm and the same attribute when a simpler cleansing was performed (same as Table 3.4). Although the mean and median are now closer to the dataset's initial pre-

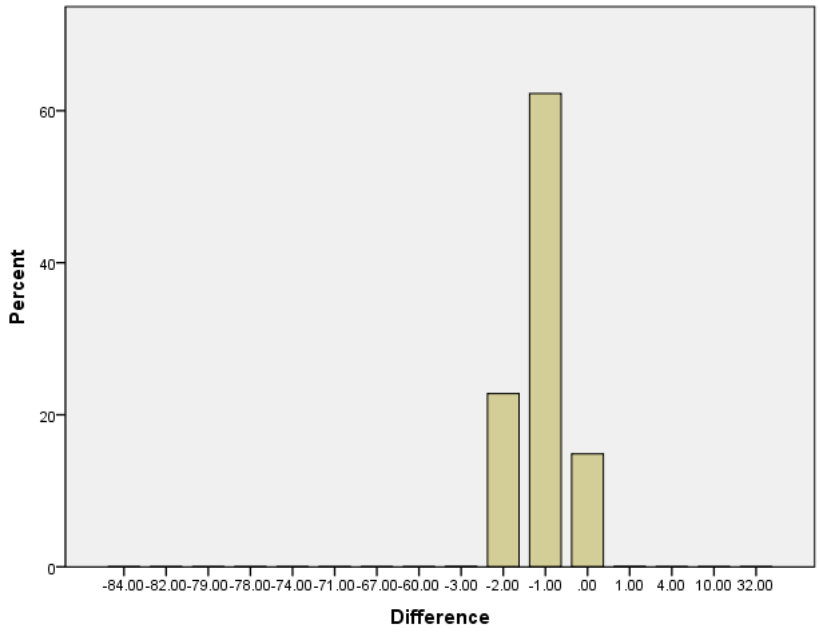


Figure 3.11: Frequency distribution of the accuracy of the algorithm (difference in years) for all records on the test set.

sentation (i.e. without any transformations), there are no significant changes in any of the statistics.

This would confirm the literature's findings [88], in that efforts to correct individual values may introduce "over-editing" and provide little overall improvement. Simpler methods to identify and remove outliers, such as the cleansing described in 3.2.2.3 should suffice when the data is to be summarised. However, when accuracy of individual records is required, an approach such as the AP algorithm can provide most reliable results.

All cleansing and editing operations were carried out using IBM's SPSS Statistics Software (version 16). The SPSS programming language (SPSS Syntax) was used to compute statistics, the age integrity checks as well as remove outliers and other cleansing operations. The linkage exercise involved running database operations and SPSS syntax scripts. The AP algorithm was written in the Python

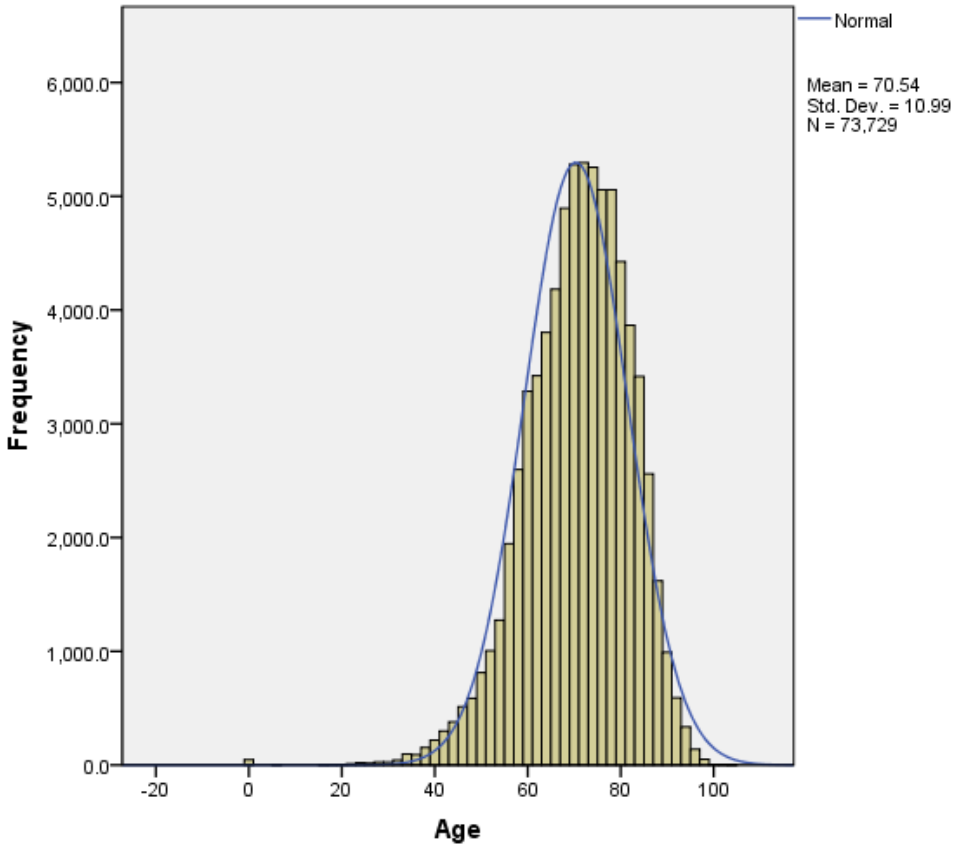


Figure 3.12: Histogram showing the distribution of age values after corrections were made by the AP algorithm.

programming language together with SPSS 16 Programmability Extension, an API allowing full control of the SPSS data and functions. Despite the implementation in Python, the extensive use of SPSS's API significantly slowed down the algorithm's run time. When handling large datasets such as the ones used in this section, the SPSS API would not be recommended.

Statistic	Age (corrected)		Age (clean)	
	Value	Std. Error	Value	Std. Error
Records	73,729		73,804	
Range	96		96	
Minimum	5		5	
Maximum	101		101	
Mean	70.54	0.040	69.57	0.040
Median	71		70	
Mode	72		72	
Std. Deviation	10.99		10.805	
Variance	120.880		116.758	
Skewness	-0.505	0.009	-0.469	0.009
Kurtosis	0.388	0.018	0.369	0.018

Table 3.7: Statistics showing the impact of the algorithm (age corrected) against simple outlier cleansing (age clean) in the distribution of the age attribute.

3.3 Stroke Study

3.3.1 Record Linkage

Data or record linkage is a common preprocessing step in data mining or data analysis projects. It allows the aggregation of data to create information that would not be available otherwise, and the identification and removal of duplicates (deduplication) [128]. A typical linkage exercise consists of identifying records on two or more sources belonging to the same entity or event (i.e. a patient or episode). An example of a simple record linkage was given in the previous section (3.2.2.5), where, for the age problem, a biochemistry dataset was linked to an administrative dataset.

Data linkage exercises are common in the support of projects in clinical, health services or public health research, with the purpose of associating records of exposures with those of outcomes [9]. These are known as *ad hoc* linkage exercises. Alternatively, linkage of health data may be undertaken in a proactive and sys-

tematic fashion, with a view to support future research and analyses [9]. The latter involves the development of a data linkage system. In the context of the work carried out in this thesis, data linkage exercises should be perceived as *ad hoc* exercises, essential for the aggregation of information that is scattered across hospital databases. Nevertheless, the work presented here introduces methods that can be used proactively to feed information into integrated repositories.

In this section, two deterministic record linkage exercises are described:

- The first exercise aims to select the correct blood tests from the biochemistry system with respect to a cohort of stroke admissions from the stroke register. Record linkage is necessary to complement the stroke datasets with accurate blood readings from biochemistry databases.
- The second exercise involves linking the stroke register with the administration system and an external, national demographics database, in order to validate the patients' dates of death over a period of 15 years. At present, it is not well established how reliable hospital administration databases are at providing accurate dates of death, in particular for those that die after they are discharged from hospital.

The next sections introduce key background information on record linkage (3.3.1.1), a formal definition of data matching (3.3.1.2), and the linkage process (3.3.1.3).

The section on the first deterministic linkage exercise (3.3.1.4) presents the technique used to match stroke admissions to their corresponding sets of blood tests (in this case, only haemoglobin is used). The second exercise on the accuracy of mortality dates using hospital databases is given in section 3.3.1.6.

Results and discussion for each exercise are given in their respective sections and general conclusions are given in section 3.3.1.8.

3.3.1.1 Background on Matching

Matching is the process by which record linkage occurs and it involves the use of key common variables or identifiers. The uniqueness or commonness of identifiers such as names, addresses or national insurance numbers is an important factor contributing to a successful matching process. However, even the most reliable identifiers yield incorrect matches [129]. Traditional data linkage is, hence, probabilistic in nature and weights should be assigned to matching or mismatching pairs accordingly [128; 129].

Computer assisted data linkage has its first *ad hoc* heuristic applications in the 1960s and the idea of probabilistic matching was first introduced by Newcombe and Kennedy in 1962 [130]. The theoretical foundation of probabilistic matching is later given by Fellegi and Sunter in 1969 [131]. As seen in the previous section, Fellegi also worked on data editing and imputation techniques at the time. Indeed linkage often produces datasets where editing techniques can then be applied.

Fellegi and Sunter's classification process relies on the computation of a vector with matching weights for each compared record pair. The summation of the vector weights yields a final composite score (likelihood ratio) for a given pair or records. An upper and lower threshold are subsequently assigned to classify pairs as matching, non-matching and possible matches. The thresholds are selected based on desired error rate bonds [132].

Probabilistic methods are particularly effective in the absence of unique identifiers such as a national insurance number. In contrast, should reliable unique identifiers be available, deterministic methods may be preferred due to the simplicity of their implementation and accuracy.

Deterministic methods differ from probabilistic ones in that linkage is traditionally performed based on a matches for a set of identifiers. An example would be

a database query, using structured query language (SQL) to join two tables on a primary key (the identifier).

Probabilistic methods are, therefore, commonly used on inter-organisation linkage exercises [133; 134; 135] and deterministic ones on intra-organisation exercises. The latter, however, often do not reveal their methods and techniques in detail in the literature [136]. This is mostly due to the specific and *ad hoc* nature of such exercises within the organisations. However, sharing methods and techniques would be beneficial to other researchers as the challenges are often similar.

Large organisations, where several departments have a high level of independence in the way they store and manage their data, may fall in the category of inter-organisation linkage However, these issues may be addressed by introducing further constraints to the matching process. This can be regarded as a rule-based deterministic method and it is the technique used in this section.

3.3.1.2 Formal Definition

A formal definition of data matching is given in detail by Gomatam [136] and summarised here.

Let A and B be two data sources with n_a and n_b records, respectively. Each of the n_b records in source B is a potential candidate match for any of the n_a records in A. Hence, there are $n_a \times n_b$ record pairs to be assigned as match or non-match [136]. Disjoint sets M and U can be defined from the cross-product set $A \times B$. Therefore a record pair is a member of set M if that pair represents a match, or U if it is a non-match. A typical record linkage process attempts to classify each record pair as belonging to either M or U [136; 137].

Matching problems, however, often need further constraints. For instance, if each record in data source B refers to a distinct entity, a record in data source A

cannot be matched to two records at the same time in data source B. This has been described in the literature as the constrained matching problem [138]. It is more generally referred in databases as 1-1 (one-to-one) linkage and contrasts with an alternative 1-many (one-to-many) linkage. 1-1 linkage, since it has more constraints, is a more difficult problem [138].

3.3.1.3 The Linkage Process

Holman [9] defined five practical steps in the process of data linkage, similar to those also reviewed in [137]. The process can accommodate *ad hoc* as well as systematic (proactive) linkage exercises; it is described below and in figure 3.13.

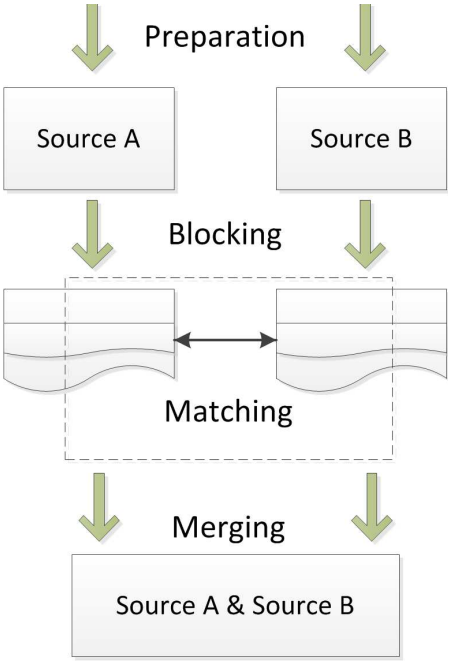


Figure 3.13: The Linkage Process, adapted from [9].

1. **Preparation**

   As with most data analysis methodologies, data preparation is a key step

where data transformations and cleansing occur. When working with personal identifiers' attributes, preparation techniques include the use of phonetic compression (i.e. transformation of names into their phonetic codes, similar to the use of stemmers in text mining), equivalences dictionary (i.e. compression of forename aliases), or geocoding (i.e. identification and categorisation of location). In linkage projects, it is also important to identify the file (or source) types. They have been defined [9] as:

- **Type 1** : files have one valid record per individual (for example, death registration)
- **Type 2** : mostly one record per individual (for example, cancer notifications)
- **Type 3** : mostly multiple and variable records per individual (for example, hospital episodes)

The above has also led to the definition of simple and complex file groups. In a simple group there is only one type 3 file whereas complex groups contain two type 2 or 3 files. The latter is also true if none of the files contains a full census of all the individuals in the analysis.

2. **Blocking**

   In this step, records are ordered and/or filtered to improve the efficiency of searching of matches. The most reliable fields with the least missing data are best suited for blocking. This is particularly important when probabilistic matching is used. The term blocking can also be referred to as a block of records pertaining to the same individual, however, this is a different meaning to the one used in this linkage process.

3. **Matching**

   This is the step where potential linkable pairs of records are systemati-

cally compared against all candidate records. This may be probabilistic or deterministic, as detailed above.

4. **Storage**
   This is exclusive to data linkage systems, and it may be skipped in *ad hoc* data linkage exercises. The storage of matching results and weights enables future studies to be carried out without having to repeat all steps of the process. In *ad hoc* linkage exercises the matching results may be discarded and the overall linked dataset is kept.

5. **Merging**
   The data from the different sources is amalgamated in a way that they can be analysed as a set of composite records for each individual. This step also includes validation checks, error detection and correction of the linkage process.

3.3.1.4 Rule-Based Linkage for Biochemistry Values

The exercise carried out here pertains to the matching of records from the stroke register to biochemistry records using relational databases. The purpose is to find, if any, the most relevant reading of a blood test for each record in the stroke register. This can be seen as a complex file group exercise since the main table of the stroke register is a type 2 file and the bloods table is a type 3 file.

The preparation and data transformations of the identifiers from the stroke register are omitted here but more detail is given in section 3.3.2 and a full list of the original stroke register attributes is given in Appendix A. A subset of biochemistry readings was used to design and test the deterministic rules presented in this section.

Let A be a relation containing the stroke register and B a relation containing the

biochemistry database with n_a and n_b records, respectively. A_{ID} and B_{ID} are the identifiers for each source, and A_{Date} and B_{Date} are the event (or episode) dates for A and B respectively.

In the first instance, we want S to be the resulting set of the equijoin of A with B where $A_{ID} = B_{ID}$. This can be defined in relational algebra as the selection:

$$S = \sigma_{A_{ID}=B_{ID}}(A \times B). \tag{3.8}$$

We then compute, t, the difference in days, $A_{Date} - B_{Date}$, for each record in S. Upon this transformation, the purpose is to select a set M containing all minimum absolute values of t in S, for each record:

$$M = \sigma_{min(|t|)}(S). \tag{3.9}$$

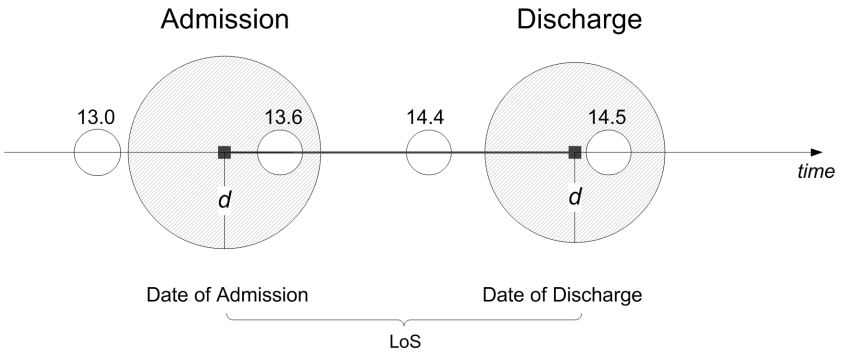
The first operation (3.8) results in a set where each record in the stroke register is linked with a record from the biochemistry dataset, and has an associated distance, t, between the date of the stroke and the date of the biochemistry test. The second operation (3.9) selects the records with the minimum distance from the first set.

However, the two operations above may not always yield the correct matches as there are no constraints in regards to time, $min(|t|)$. Hence, in some cases, a further selection query is needed to restrict records to the time period in question. Selecting $min(|t|) \leq d$, for instance, would ensure the matching records are within an absolute limit of d days. The value assigned for d is determined by the clinical time dependency between records in source A and B, i.e. determined by domain experts with a particular clinical interest.

In this study, the threshold d was set to 7 days by the clinical team. Hence, blood records closest to admission with a limit of 7 days were matched against the stroke episodes. The same exercise was carried out for discharge events.

Figure 3.14 shows a graphical representation of two possible deterministic approaches based on an example from the study. The horizontal lines represent a patient's timeline. The smaller white circles intersecting time illustrate biochemistry tests, with example readings above them. The two large circles in the first approach (I in figure 3.14) are centered in the admission and discharge dates respectively; their radius, d, is the threshold set above, in days. Any biochemistry readings within d would be considered as potential candidate values for admission or discharge. The distance between admission and discharge is the length of stay (LoS). In the second approach, the larger circle shows the LoS, and at its upper and lower bounds, admission and discharge circles were placed.

I.



II.

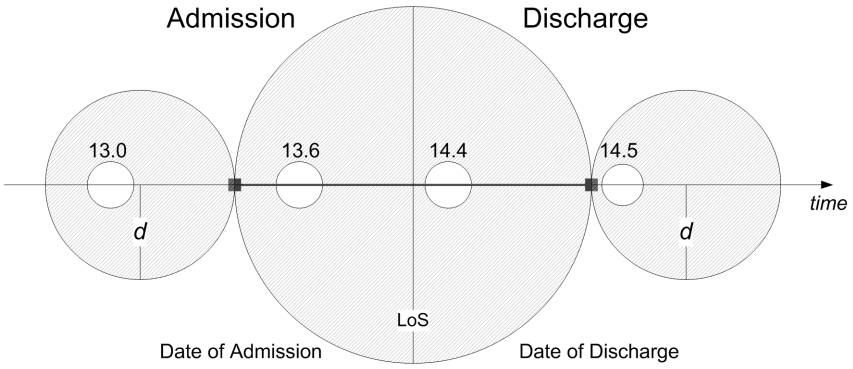


Figure 3.14: Deterministic methods (I and II) to find the appropriate biochemistry records for linkage. Circles were used to create a figurative radius (d) of individual blood tests (in white) pre-admission, post-discharge and the distance between the latter two (length of stay, LoS).

The figure illustrates, with an example from the study, that the first approach would retrieve the biochemistry readings closest to admission and discharge but ignore the period of time between them, the LoS. This is a typical approach used when running database queries and indeed the result of the abovementioned operations. The second approach, however, shows how other values can be included instead. When using the first approach, a biochemistry reading with value 13.6

is chosen as the admission value. The second approach, however, would choose the biochemistry test with value 13 instead.

This is important because of the potential effects that admission (and the time between admission and the first reading) may have had on the values of the biochemistry reading. Hence, the most correct reading for admission, in the example, would be 13. This can be achieved by introducing a constraint on $t \geq 0$ meaning $A_D \geq B_D$ for admission. Regarding discharge, however, the absolute t closest to discharge date would still be most correct.

As such, the operations involved in this exercise are:

1. Perform selection, $S = \sigma_{A_{ID}=B_{ID}}(A \times B)$ to match all records in A with B where $A_{ID} = B_{ID}$, and append $t = A_{Date} - B_{Date}$.
2. Obtaining the value for Discharge:
   (a) Select from S the records where $|t| \leq 7$, resulting in set $M = \sigma_{|t|\leq 7}(S)$.
   (b) The resulting value for discharge, A_{ID}, is the minimum value in M.
3. Obtaining the value for Admission:
   (a) Select from S the records where $t \geq 0$ and $t \leq 7$ resulting in set $M = \sigma_{t\leq 7 \text{ and } t\geq 0}(S)$.
   (b) The resulting value for admission, A_{ID}, is the minimum value in M.
4. When no value for Discharge is found, the same value at Admission may be used.

The above operations were carried out using a subset of biochemistry tests in order to evaluate the results, described in the next section.

3.3.1.5 Biochemistry Linkage Results and Evaluation

A set of 7,208 stroke episodes (2003-2011) from the stroke register was used as the main strokes table. A second table, containing all haemoglobin (Hb) blood tests from the same time period was retrieved from the NNUH biochemistry data warehouse. The method for data collection was the same used in the previous chapter. The Hb table contains 3.5 million haemoglobin readings and this volume poses an additional challenge to the current and future linkage exercises.

The purpose of this record linkage exercise was to match the correct blood reading, in this case haemoglobin is used , to two stroke events, admission and discharge. At least one haemoglobin reading is expected as part of a routine full blood count near the time of admission, but the extent in which this is consistently performed across all stroke patients is unknown. As seen in the previous section, different rules were used to obtain the Hb values at discharge and at admission. The results of this exercise are presented below and were evaluated by an inspection of the descriptive statistics and a clerical review of patient notes. The work presented here uses Hb as an example to demonstrate the linkage and data transformations, however, the same was techniques were applied to all blood tests identified by the clinical team.

Admission

Deterministic record linkage was applied to the dataset and 94.3% (n=6,795) of stroke admissions had a matching haemoglobin reading within 7 days. However, when observing the distance, t, in days, between the blood records and admission, only 79.3% matched on the same day ($t = 0$) yet 91.4% matched within a day of admission. The frequencies for all t are given in table 3.8. Those patients without a reading of Hb at admission (n=113) did not appear to belong to a specific year group or have any other characteristic that would indicate a reason for not having the blood reading.

Admission t	**Frequency**	**Percent**
0	5,713	79.3%
1	875	12.1%
2	95	1.3%
3	53	0.7%
4	19	0.3%
5	19	0.3%
6	11	0.2%
7	10	0.1%
> 7	304	4.2%
No Match	113	1.5%
Total	7,208	100%

Table 3.8: Frequency of matching records for admissions. t indicates the number of days before admission where a blood reading (Hb) was found.

Discharge t	**Frequency**	**Percent**
0	1,052	14.6%
1	1,479	20.5%
2	1,011	14%
3	764	10.6%
4	532	7.4%
5	497	6.9%
6	348	4.8%
7	222	3.1%
> 7	1,190	16.5%
No Match	113	1.6%
Total	7,208	100%

Table 3.9: Frequency of matching records for discharges. t indicates the number of days before or after discharge where a blood reading (Hb) was found.

Discharge

Regarding discharges, only 14.6% of patients had a Hb reading on the same day they were discharged and 34.6% had a day before. According to the deterministic rules, 81.9% of stroke patients have a relevant reading at discharge (i.e. within 7 days). Table 3.9 shows the distribution of t for discharges.

The number of non-matching records is the same as for admissions (n=113). This happens because the given rules for discharge can overlap with the admission ones and so the same number of non-matching records, as measured here, is expected. Indeed when the length of stay is less than or equal to 7 days and only one blood reading from the patient was taken at admission, the reading at discharge is the same as the one at admission.

Discussion, Evaluation and Challenges

The descriptive statistics resulting from the matching exercise were used to evaluate the total number of exact matches, partial matches and mismatches. Partial matches were determined by t, the distance between the admission/discharge and the corresponding blood reading. A further evaluative part of this exercise was carried out by a manual inspection of the matching results against patient notes.

Overall no errors were identified and the deterministic linkage provided the correct values for all records. When no Hb values were found they were also correctly classified as non-matches. Furthermore, this exercise was carried out for a single blood test, Hb. The deterministic linkage rules presented here should still be used carefully in other exercises and manual inspection of the results is also recommended. Further details on other biochemistry tests are given later in this thesis, together with a discussion on other issues affecting linkage using deterministic methods.

Regarding technical challenges, this exercise was first attempted using Microsoft

Access databases as they are installed on hospital computers. However, Microsoft Access was not able to handle the linkage exercise both due to the sheer size of the tables and the difficulty of writing and running complex Structured Query Language (SQL) queries. A second, most successful approach was to use a portable database management system, SQLite. The latter is compliant with the properties that guarantee database transactions' reliability (Atomicity, Consistency, Isolation, Durability) and it implements the SQL standards required for the exercise. Even though SQLite is remarkably faster than Microsoft Access in running queries, the queries above took $\sim$ 16 minutes to run for admissions and $\sim$ 10 minutes for discharges. The exercise was carried out using an average computer at the time of writing (4Gb RAM, Dual Core 2.4 Ghz).

3.3.1.6 Quality Assessment of Linked Mortality Data

A second linkage exercise, critical to ensure the quality of the stroke dataset and in particular, mortality data, was undertaken. This exercise attempted to evaluate the quality of the date of death field present in hospital systems and can be regarded as an inter- and intra-organisational linkage exercise. This exercise emphasizes the value of linked health data in the understanding of underlying processes and data quality.

Upon a patient's death in hospital, the attending clinician or a coroner complete a medical certificate of cause of death. The patient is then discharged and the death is registered (registrations are discussed in more detail below). In hospital, the administration system (PAS) is updated with the patient's date of death and, in the stroke database, the patient is discharged with the respective status. For the purpose of hospital service data the patient's records for a deceased patient can be archived and the quality of the data is thus assured. When the patient is discharged alive, whether to a rehabilitation center, care home, or home, the pathway for data collection is also considered complete from a service perspective but from a research perspective this may not hold true, as further events may follow that are not recorded in HIS.

Given the negative skewness of the age distribution in stroke patients, this cohort of patients would naturally see an increased mortality when compared to younger cohorts, even upon a successful treatment. Any research studying longevity would need additional follow-up information to assert whether the patient is alive at a particular time-point after discharge. However, it is unclear how death information is reported back to the hospital. Upon spot-checking individual records in the hospital administration system and consultation with clinicians, it was observed that deaths were being reported to the hospital by GP practices or by the families of the deceased.

Further consultation with the hospital's data quality department revealed that a weekly process is in place to ensure that information on deaths, among other details, is fed to the hospital records from the Summary Care Records (SCR) database. The SCR is a "centrally stored summary of key medical details that is created from a persons existing NHS record (...) and made available to NHS staff in emergency and unscheduled care situations" [139]. The SCR is fed data from GP practices and was first rolled out in 2006 (2007-2009 in around Norfolk). At the time of writing, only one GP practice with a Norwich postcode was still planning to commence using SCR. In order to send clinical information to the SCR, GP practices must synchronise their system with the Personal Demographics Service (formerly NSTS - National Tracing Service) to identify the correct patients. It is expected that the quality of the records, including date of death, has largely improved with the SCR system.

As reported in the previous chapter, the NSTS system was a national database where services such as NHS Trusts submitted queries to find accurate demographic information about anyone registered with the NHS. From May 2013 until the time of writing, the NSTS was unavailable and being replaced by the new Personal Demographics Service (PDS). Prior to SCR and the PDS, hospitals used reports from NSTS to gather information on their patients. However, upon local consultation with the clinical team and administration services, it was not clear how and when the hospital was feeding this data into its PAS system. The overall process and reporting times remained unclear.

Constraints to the death registration process could pose additional obstacles. Indeed, when inquests and post mortems take place, the death is only registered after the inquest. The overall figure from the 2012 coroners statistics bulletin [140] states that post mortems were carried out in 30% to 50% of all deaths in Norfolk and in 42% in England and Wales. The same report states that the average time to process inquests is between 20 and 30 weeks in Norfolk (27 weeks

in England and Wales), however, post mortems are usually carried out within a few days of the patient's death. Nevertheless, NHS institutions only keep the date of death, and the NSTS/PDS hold the date of death and a separate indicator of whether it has been registered or just recorded by an NHS institution. It is not expected that inquests and post mortems are likely to affect the quality of the reported deaths. However, with regards to timeliness, in cases where deceased patients were found some time after they died, post mortem investigations might reveal an earlier date of death.

The following sections detail an investigation on the quality of the dates of death present in the hospital's patient administration system, resulting from a record linkage exercise between the cohort of stroke patients, the hospital PAS system and the National Tracing Service (NSTS). The data collected from the NSTS, late in 2012, was requested by the hospital's information services department based on a list of NHS numbers suffering a stroke. This time consuming step resulted in a matching set containing the most recent and accurate demographic information for that cohort.

Cohort, Linkage and Validating Identities

The record linkage step consisted of linking the NSTS dataset with the stroke register and a PAS dataset based on NHS number. This exercise uses a simple file group with no type 3 files and only one type 2 file (the stroke register). Only patients with a valid NHS number could be matched accurately to the national database. In this cohort, all selected patients have a valid date of death (DOD) in the national system (between 1997 and 2011) and from the stroke register database, only the patients' last admission is considered. Some of the records in the national system, however, may have been previously restricted at the patient's request and this may result in a mismatch.

A total of 5,092 records from the NNUH with valid NHS numbers and a date of

death in NSTS were selected. From this set, 2.2% (n=116) of records were not accurately matched with NSTS either due to incorrect or inconsistent identifiers (including date of birth or surname) or because they were restricted or not present in the NSTS database. It was not possible to ascertain the true reason for the mismatches, however, the vast majority of those (71%, n=82) died before 2001. A manual inspection of some of the records revealed that the issues could lie with NHS numbers but that further inspection and access to the NSTS system would be needed to understand whether the mismatching pairs were indeed from the same person.

The mismatches were removed from the dataset and a total of 4,976 patients (56% female) with average of 331 patients per year (SD 73) and matching NHS number and hospital numbers was found. The average age of the patients at stroke episode is 80 years (SD 10) and the average age at death is 82 years (SD 9).

In a first instance, it was important to further validate the identities of the mismatching pairs by looking at any mismatches of the identifiers (other than NHS numbers). Gender and date of birth were used and considered sufficient to validate the links. As shown in table 3.10, no mismatches were found in gender or NHS number, yet a small number was found in date of birth (0.4%). A larger number of mismatches was found in date of death (10.4%) and, upon validation of the links with the identifiers, this is the focus of the work carried out here.

Data Element	**Frequency**
NHS Number	0
Gender	0
Date of Birth (DOB)	0.4% (n=21)
Date of Death (DOD)	10.4% (n=519)

Table 3.10: Overall number of mismatching records for the three patient identifier elements (three first rows) and date of death.

An analysis of the distribution of the mismatches in the identifiers (in this case, only dates of birth), allowed an understanding of the extent of the error (i.e. how far apart the two dates of birth from the different systems were). All records had a valid date of birth (no missing values): 38% (n=8) were a day apart, 19% (n=4) were between 5-61 days apart, and 33% (n=7) were either 365 or 366 days apart. This leaves two values, one 730 days apart, and another 336 days.

When investigating whether the mismatches would have a potential effect on the quantification of mismatching dates of death, it was found that only five records would be affected That is, of the 21 mismatching dates of birth, five also had a mismatching date of death. Four of those had a mismatching date of birth within a day and one within 365 days. These could be both due to typographical errors and not true identity mismatches. Because all other records (n=16) had a matching date of death and their names were manually inspected, they were also not considered identity mismatches.

It is possible to state that, from the cohort of 4,976 stroke patients, 0.4% had an incorrect date of birth, most likely due to typographical error (same day and month but different year, or a day before or after) or a small delay in the reporting process. However, it is not possible to ascertain here which data source holds the true date of birth.

Nevertheless, sufficient evidence was gathered to validate that the linkage and an analysis of mismatching dates of death should not be affected by incorrect identities. The validation of identities could be considered a pseudo-probabilistic exercise in the sense that only matches with a probability of 1 were considered valid using the available identifiers. The computation of the probability of mismatching records belonging to the same person based on the other identifiers was not necessary. It is now pertinent to inspect the distribution of the 10.4% (n=519) mismatching dates of death.

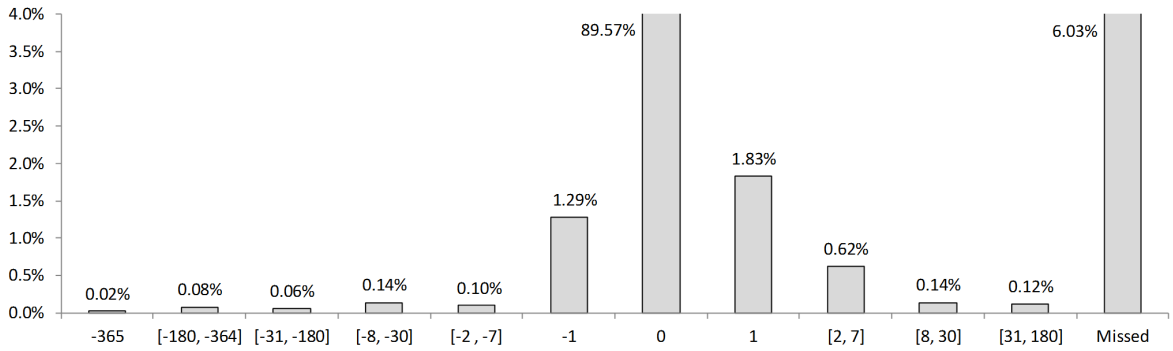
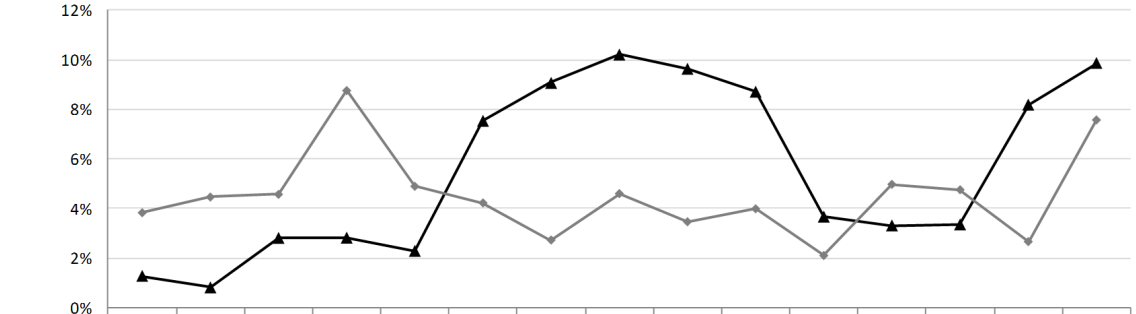


Figure 3.15: Distribution of Date of Death Matches and Mismatches. Difference in days between the date in PAS and date in NSTS, and percentage of missing data from PAS. The figure is scaled to show the distribution of missing data across all time intervals.

3.3.1.7 Mortality Linkage Results and Evaluation

Figure 3.15 shows the distribution of the matching of all records. The negative values indicate that the date of death in PAS was before the date in NSTS. A first observation is that there are more positive values (n=135) than negative ones (n=84), perhaps indicating a natural delay in processing or reporting the information from the local to the national system. Another important observation is that 6% (n=300) of mismatches are indeed missing from the PAS system. It is also important to note that the remaining 4.4% of mismatches do not exceed a 365 days difference, and hence, when using the PAS source alone to determine dates of death, the worst case error would be for the dates to be at most within a year.

Given a distribution of the dates of death (DOD), on average, 20 patient deaths (SD 14) are missing from the PAS system every year and 15 (SD 5.6) are recorded incorrectly (within a year). However, this number isn't consistent throughout the years. Figure 3.16 shows the yearly rates of missing and erroneous DODs. Despite the volatile numbers, the proportion of missing DODs has increased over



	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Missing	1.27%	0.81%	2.81%	2.81%	2.28%	7.53%	9.06%	10.20%	9.63%	8.71%	3.66%	3.31%	3.35%	8.19%	9.87%
Error	3.82%	4.45%	4.56%	8.75%	4.89%	4.22%	2.72%	4.59%	3.46%	3.98%	2.09%	4.97%	4.75%	2.65%	7.57%

Figure 3.16: Distribution of Date of Death Mismatches (erroneous and missing) by year, from 1997 to 2011. The yearly average is 10.1% (SD 3.6) of which 5.6% (SD 3.4) is missing and 4.5% (SD 1.7) is recorded incorrectly (within a year).

this time period, having a maximum of 10.2% in 2004 and a minimum of 1.27% in 1997. The computed average (15 years) for missing deaths is 5.57% (SD 3.37) and for errors is 4.5% (SD 1.68). When observing the last five years (2007-2011), the missing average is 5.68% (SD 2.79) and error average is 4.41% (SD 1.94) suggesting little difference before and after the implementation of the SCR. However, the 2002-2006 and 2010-2011 periods are noticeably high for missing deaths, both averaging 9% (SD 1 and 0.008 respectively). It was not possible, at the time of writing, to explain the reasons behind this, even after consultation with hospital staff.

When observing additional features of those with missing dates of death in PAS (n=300), a larger number of females was present (66%, n=197). The proportion of females with erroneous (typographical) errors was found to be 53% (n=118) and 55% (n=2468) for matching records. When grouping errors and matches together and comparing them against those with missing dates of death, a statistically significant relationship was found. The odds of having missing dates of death in females is 1.83 times higher than in males (Confidence Interval [1.423, 2.357], p-value < 0.001).

It is not well understood why females have proportionally more missing dates of death, however, an increased life expectancy is a possible driver. Overall, females in this cohort died with an average age of 83 (SD 9) whereas males died at 80 (SD 9). Within those with missing information, females died on average at 85 (SD 8) and males at 81 (SD 9). Further information, not available here, could help to explain whether the missing data exists due to families (or partners) not reporting the deaths back to hospitals or GP. However, when using other available data fields such as discharge destination, no significant difference was found between those who were discharged home alone or home with family.

Table 3.11 shows a distribution of the discharge destination field, discriminating between OSR and NSR and showing the original values. This exemplifies the semantic heterogeneity between the systems. Table 3.11 is also in ascending order of matching percentage. As expected, from those that died in hospital, only one record was missing a date of death. Indeed deaths in hospital have a significant effect in lowering the overall number of missing dates of death. When these are excluded from the dataset, the overall (adjusted) number of missing dates of death becomes 10.39% and errors, 4.86%. Nevertheless, the above correlation between females and missing data still holds true when excluding those that died in hospital (Odds Ratio 1.6, Confidence Interval [1.22, 2.05], p-value < 0.001).

From table 3.11, those in the OSR who had a "long stay hospital" also had a percentage of missing records below the adjusted average at 9.87%. This is particularly significant because "long stay hospital" patients make up a quarter of the total number of patients in the dataset. Those discharged "home with family [no care pkg]" from the OSR are the third largest group (10%) and also under the adjusted average at 9.16%. Lastly, the fourth highest (8%) is OSR's "Rehabilitation unit" with an above average value of 11.42% and the overall highest percentage of missing values (22%) pertains to those discharged to another hospital or department.

Source	Discharge Destination	Total	Match (N)	Missing (N)	Error (N)
NSR	CARE	0.04%	100% (2)	0	0
NSR	home alone	0.22%	100% (11)	0	0
NSR	DEATH	6.33%	98.10% (309)	0	1.90% (6)
OSR	DEATH	35.83%	95.85% (1709)	0.06% (1)	4.09% (73)
NSR	Rehab - gerneric community bed	0.80%	95% (38)	0	5% (2)
NSR	stroke inpatient rehab	1.19%	93.22% (55)	6.78% (4)	0
NSR	Early Supported Discharge	0.52%	92.31% (24)	0	7.69% (2)
NSR	home with family [no care pkg]	1.15%	91.23% (52)	5.26% (3)	3.51% (2)
NSR	home with care package	0.58%	89.66% (26)	6.90% (2)	3.45% (1)
OSR	home with family [no care pkg]	10.09%	86.25% (433)	9.16% (46)	4.58% (23)
NSR	Residential home	0.56%	85.71% (24)	10.71% (3)	3.57% (1)
OSR	long stay hospital	23.41%	85.49% (996)	9.87% (115)	4.64% (54)
OSR	Rehabilitation unit	7.92%	84.01% (331)	11.42% (45)	4.57% (18)
NSR	transfer other acute hospital	0.24%	83.33% (10)	16.67% (2)	0
OSR	Unknown	0.24%	83.33% (10)	16.67% (2)	0
OSR	home alone	3.94%	82.65% (162)	11.22% (22)	6.12% (12)
NSR	HOME	0.10%	80% (4)	0	20% (1)
NSR	Nursing home	0.90%	80% (36)	15.56% (7)	4.44% (2)
OSR	residential home	1.41%	80% (56)	12.86% (9)	7.14% (5)
OSR	nursing home	2.03%	77.23% (78)	12.87% (13)	9.90% (10)
NSR	EMI nursing home	0.08%	75% (3)	25% (1)	0
NSR	Unknown	0.08%	75% (3)	25% (1)	0
OSR	other hospital or department	2.09%	74.04% (77)	22.12% (23)	3.85% (4)
NSR	REHAB	0.14%	71.43% (5)	14.29% (1)	14.29% (1)
NSR	community bed [not rehab]	0.10%	60% (3)	0	40% (2)
-	Total	N=4,976	89.57% (4,457)	6.03% (300)	4.40% (219)

Table 3.11: Original discharge destinations from the OSR and NSR and given coverage, percent of matches (in descending order), missing and typographical errors for dates of death.

When grouping those that were discharged to other institutions (n=437, community hospital, nursing/care home, other hospital) with those discharged home (n=800, home with and without family, early supported discharge) a statistically significant relationship was found. Those discharged to other institutions were found to be 1.52 times more likely to have missing dates of death when compared to those who were discharged home (Confidence Interval [1.04, 2.23], p-value = 0.023). The remaining group (n=3,739) includes those that died in hospital, those discharged to another hospital location, and unknowns.

In this linkage exercise, the identity of each patient was validated against other identifiers. As such, even though it is a inter-organisational linkage exercise, it was not necessary to compute the probability of a particular record referring to the same identity. A simpler, deterministic linkage approach was undertaken.

Key Findings

This exercise allowed important questions that may be useful to other studies, or in other areas using hospital data, to be answered. In particular, this exercise has helped with the understanding of how the information flows back to the hospital after a patient dies. Four major key points, previously unknown, were identified based on a cohort of 4,976 patients who had a stroke between 1997 and 2011:

- The total number of potentially erroneous or mismatching dates of death when using PAS data was identified at 10.4%. However, given the available data, it is expected that erroneous dates, when available (4.4%), lie within a year of the true date of death.
- The overall number of missing dates of death in PAS for patients who have indeed died is 6%. From those that died in hospital only one record was found to be missing (<0.005%). When removing those that died in hospital, the overall adjusted number of missing dates of death rose to 10.39%.

- Overall, females were more likely to have a missing DOD than males, and those discharged to other institutions were also more likely to have a missing DOD when compared to those who were discharged home.

- A year-on-year distribution of missing DOD over the 15 year period would be best fitted by a higher order polynomial trend line; it shows periods of high values followed by, or preceded by, periods of low values. The maximum recorded missing value was 10.2% in 2004 and the lowest, 1.27% in 1997. The introduction of the SCR showed no significant or sustained improvement in the number of missing DODs, although more data is required to confirm this assertion.

The limitations of this exercise include the fact that only stroke patients were considered, which might reduce its generalisability to other domains, despite a large cohort of mostly elderly patients being used. The lack of consistency in field values between OSR and NSR made it more difficult to group discharge destinations and, more generally, to prepare the dataset, as other field values also had to be mapped. When the cohort was first retrieved from PAS, additional important data could have aided the interpretation. In particular, GP post codes, marital status, and free-text notes regarding the date of death would have been beneficial.

Nevertheless this exercise provided valuable insights into a previously unknown and unquantified topic. This would not be possible without the linkage of three distinct data sources. It is now known that, should these numbers be generalised to other populations, a conservative margin of 10.5% missing values for those who did not die in hospital should be taken into account. It is expected that, when SCR are fully implemented across GP practices and records are systematically synchronised with the PDS system, hospitals can begin to rely on SCR reports to accurately feed their systems. Further work is still needed and it would be bene-

ficial to analyse missing data by GP postcode for a longer time period. Despite the limitations, this exercise allowed the stroke register database to become more complete as missing values were replaced with dates of death from the national database.

3.3.1.8 Discussion and Conclusions

This section explored record linkage by undertaking two exercises. A first exercise examined how rule-based linkage was used to join a set of stroke episodes with a respective set of blood readings. The second exercise explored how stroke patients were linked with a national demographics database to assess their follow-up status. Both exercises revealed that record linkage is feasible and can be used to enrich the value of the data present in hospitals.

Regarding the first exercise with a complex file group, the purpose was to find, if any, the most relevant reading of a blood test for each record in the stroke register. The exercise provided insights into the the availability of electronic biochemistry records for stroke admissions and discharges. The majority of patients had a haemoglobin reading at admission (91.4% within a day of admission) yet only 35.1% had a reading within a day of discharge. The distribution of readings at discharge is scattered, and depends on each patient's length of stay. Nevertheless, the number of mismatches was consistent for both admission (1.5%) and discharge (1.6%). In this exercise, haemoglobin was sucessfully used to develop and test the linkage technique. The same technique was later applied to 21 other biochemistry tests, as depicted in section 3.3.2.

The second exercise, aimed at linking the stroke data with mortality data from PAS and the national demographics system was a simpler linkage exercise from a technical standpoint. However, this exercises demonstrated the feasibility and value of linking patient hospital data to national databases. In particular, given

the observed time-period, it was found that 10.39% of all patients who had a stroke and died after discharge were missing a date of death (DOD) in the hospital systems. By linking other data sources and obtaining additional features about those with missing DODs, females were found to be more likely to have a missing DOD than males, and those discharged to other institutions were also more likely to have a missing DOD when compared to those who were discharged home. Furthermore, changes to hospital services, data quality processes and national systems are likely to have contributed to a volatile distribution of missing data over time. This exercise also served to provide correct mortality information to the stroke database.

The evaluation of data linkage exercises, that is, assessing whether matching records have indeed the same identity, is often a difficult task, only possible with a test set of previously known matching and mismatching pairs. This is particularly true in probabilistic exercises [141] whereas in deterministic exercises, additional features can be used to inform on the accuracy and reliability of the matches. In the first exercise described in this section, evaluation was performed by a manual inspection of patient notes. In the second exercise, other key identifiers were used to ascertain the true matches. In the latter, 2.2% of records in the initial dataset were not accurately matched with the national database due to inconsistencies in other key identifiers. This shows that deterministic linkage, despite relying on unique identifiers, is also prone to error.

Overall in this section, the data linkage exercises were performed to provide additional data for research. However, data linkage systems as a research infrastructure have been described [9] and in the next section, the proactive use of record linkage is described as part of efforts to build a research data warehouse.

3.3.2 Data Warehousing and Integration

In the previous chapter, a methodology for multi-source data collection was introduced and used to create a stroke operational data store (ODS). The previous chapter also briefly discussed how an ODS could be used to develop integrated repositories or data marts. This section describes this process in more detail and with regards to the development of a sustainable infrastructure for linkage, storage and reporting of the stroke data. In particular, this section describes how the stroke data in the ODS was integrated and the development of the Norwich Research Stroke & TIA Register data warehouse.

Data warehousing is a concept that emerged in the 1980s to tackle the challenges of using traditional transactional systems for reporting. Upon a thorough background review, the works of Bill Inmon and Ralph Kimball dominate the literature [2] but are mostly aimed at IT practitioners. It is not until recently that academics began to take an interest in the subject, and thus the field has been mostly driven by the market, rather than by the research community [142]. As such, there is little consistency in the literature regarding the historical factors that led to the development of data warehousing techniques, concepts and technologies.

Section 3.3.2.1 introduces a historical perspective on data warehousing; it introduces key ideas, concepts and models and the state of the art developments in the field. Later, in section 3.3.2.2, this is explored with regards to health informatics and key background information on similar developments tailored to this area is introduced. In section 3.3.2.3, a thorough understanding of data warehousing architectures allowed the adoption of key concepts and methods for the development of the stroke register data warehouse. In particular, section 3.3.2.4 introduces a methodology to develop the warehouse. This is a novel hybrid approach, applicable to other domains, with the purpose of supporting and building

research capacity infrastructure whilst allowing a better understanding of stroke care over time and new research questions to be answered.

3.3.2.1 A Historical Perspective

The creator and principal architect of the first commercially available database management system was Charles W. Bachman in 1963 [143; 144]. The Integrated Data Store, also known as the navigational database, was a revolutionary system based on a hierarchical model of arranging data. In this model, a tree-like structure represents information via relationships, for example, a parent node having one or more child nodes (a one-to-many relationship) [145]. Paths and pointers were used to navigate and query the database, creating an emphasis on the navigation itself rather than "declaratively" selecting the required data [143; 145].

During the 1960s other systems and data models continued to appear[146; 147]. Of particular relevance is the Conference on Data Systems Languages (CODASYL) model which helped to standardise database interfaces. Also known as the network model, the CODASYL model allowed each record to have multiple parent or child records, forming a graph-like structure rather than a tree-like one, enabling the representation of more complex relationships [147; 148].

By the end of the 1960s the early database systems were arguably home grown, *ad hoc* collections of ideas formulated into systems, originally designed to solve a particular problem and later extended to become a more general purpose solution [149; 150]. The shortcomings that then existed in the fields of database management and application development (in particular, querying and relating data) were addressed in 1969 when Edgar F. Codd introduced the relational model [149; 150]. Codd's work on the relational model was preceded by David Childs' feasibility work of a set-theoretic data structure [151], which introduced the idea of relations in data systems.

In brief, the relational model is a database model based on first order predicate logic where data is represented in terms of lists of elements (tuples) grouped into relations (between a tuple and an attribute) [150]. The concept of a relational database with a tabular organisation where each record (row) has a series of attributes (columns) was then introduced, and it became one of the most fundamental concepts in computing sciences. Using a relational database system, users were able to query the data tables resulting in a specific set of data (called a view or resulting set). Developing the relational model, in the following years, Codd [150] introduced a set of normalisation rules to minimise redundancy (dividing large tables into smaller ones) and dependency (utilising relations whereby modifications to the table could be propagated throughout the database).

Before Codd's ideas were introduced, businesses had already begun implementing operational database systems for storage and retrieval of day-to-day activities. Indeed the concept of On-line Transaction Processing (OLTP) technology emerged soon after Bachman's navigational database [143; 145]. An OLTP system focuses on tasks that are structured and repetitive, consisting of short, atomic, isolated transactions [152]. Such systems work with detailed, up-to-date data, and read or update database records [152]. With Codd's relational model introduced, most operational systems began to rely on database normalisation and a data model (such as the entity-relationship model) to ensure data integrity and the timeliness of recording of their transactions. This model was ideal for businesses' day-to-day operations.

Upon sucessful implementations of operational database systems [146], businesses became interested in harnessing the statistical power of their ever growing data for reporting, accounting, planning or forecasting. Operational systems, however, were not at the time built for this purpose. As a result, the increasing demand for complex analyses and summaries of transactions added pressure to the existing operational systems, affecting their primary uses. A solution that enabled the

secondary uses of such systems and their data was much needed.

Data warehousing as a concept dates back to the late 1980s when IBM researchers Barry Devlin and Paul Murphy developed the "business data warehouse" [153]. At the time, the purpose of a data warehouse was to feed data from operational systems to decision support environments for analysis [154]. Without a data warehouse in place this process would be technically difficult for the operational system and would introduce a significant delay in both reporting and the current operations. Three main technical difficulties were given by Delvin and Murphy [153]:

- Ensuring that the performance of the production (operational) systems is not disrupted by *ad hoc* queries or analyses.
- Requiring that information needed by end-users is not changing as they use it (point-in-time data).
- Operational systems designed for high volume processing are not often suitable for answering unpredictable queries from end-users.

The concept of a data warehouse was emerging as a way of separating and organising the data stores according to their intended use. Operational systems would primarily focus on day-to-day activities (for example, order entry, distribution, billing) while one or more informational systems concentrated all aspects of reporting and analysis. As such, Delvin and Murphy worked on extending the basic data architecture to support this environment. Based on the existing relational database environment, they proposed the "Europe, Middle East and Africa Business Information System" (EBIS) architecture for informational systems [153].

The core component of this architecture is the business data warehouse (BDW), regarded as the single logical storehouse of all the information used to report on

the business [153]. End-users interacting with the BDW are presented with a particular view, determined by the user's requirements and containing accessed data. The BDW connects to the operational system via a data interface (that feeds data in an agreed-upon format) and to the end-users' workstations. A business data directory is also connected to the BDW where information from a data dictionary (first introduced in 1975 by Uhrowczik [155]) and business process definitions are made available. The data in the BDW can be retrieved raw or it can be enhanced (by, for example, aggregation or summarisation), depending on the level of abstraction needed by the end-user.

The EBIS achitecture fuelled the discussion among practitioners and researchers on data warehousing improvements and business models adaptations, and is arguably one the foundations of today's modern data warehousing architectures, discussed in detail in the next section.

Shortly after EBIS, in 1993, William Inmon introduced the concept of a data warehouse as a "subject oriented, integrated, non-volatile and time-variant collection of data in support of managements decisions"[156]. Inmon's work [156] is widely cited and he is often referred to, in the literature, as the "father of data warehousing", having published the first book on the subject. Inmon's work became prominent in the 1990s and established four key characteristics of data warehouses:

- **Subject Oriented**
  In a DW, data should be organised by major topic or subject area. This is an important characteristic that determines the usability and performance of the data warehouse.

- **Integrated**
  Data from different operational systems needs to be integrated into a data warehouse. It is often the case that information across operational systems

lacks consistency in encoding, naming conventions or physical attributes, as seen in the previous chapter.

- **Non-volatility**
  Data in the operational system is highly volatile, i.e. it may change over-time. In data warehouses, data is non-volatile in that only a snapshot of the data (cross-sectional data) is available. Data warehouses may keep several snapshots of the data over time.

- **Time variancy**
  Every record in a warehouse must have an associated time (date of transaction or time stamp) indicating the moment in time during which the record is accurate.

Along with these characteristics, Inmon advocated a top-down approach to developing data warehouses that adapts traditional relational database tools to the needs of an enterprise-wide data warehouse [2]. Inmon's view of a data warehouse is of a centralised system with all, or almost all, departmental systems in an organisation feeding data into it.

Also in 1993, Codd introduced the concept of an On-Line Analytical Processing (OLAP) system to oppose the On-Line Transaction Processing (OLTP) [149]. An OLAP system can be regarded as a informational system (such as a data warehouse) and a OLTP system would be an operational system. However, there are also differences between a data warehouse (as defined by Inmon) and OLAP. Inmon describes OLAP as a technology while data warehousing is an architectural infrastructure, he also notes that a "symbiotic relationship exists between the two" [156]. When it comes to organising data, OLAP uses a multidimensional view of aggregate data to provide quick access to strategic information for further analysis [157] while a data warehouse is usually based on the relational model.

The multidimensional model organises data in multidimensional tables (called data cubes) and it may be implemented alongside a data warehouse [158]. Normalised data from relational databases would typically be denormalised for an OLAP environment so as to improve performance and facilitate reporting. In an OLAP system, dimensions include data from a single domain. An example from a sales environment would be to define the following dimensions: time, product and geographic area. An OLAP report would then look at sales by specific dimensions, where similar information is lined up. Common capabilities of OLAP also include aggregation, drill-down and roll-up, and slicing and dicing [157; 158]. These operations provide additional ways of selecting data from the cubes and their dimensions and have been particularly fruitful in answering business questions [142; 158].

Following OLAP developments in the early 1990s, in 1996, Ralph Kimball introduced a novel model for data warehousing [159]. Kimball's model opposed Inmon's in that building a data warehouse was by him considered a bottom-up approach, heavily based on the dimensional model. Kimball's model suggests creating one database (or data mart, a domain-specific subset of a data warehouse) per major business process first and then utilising a "data bus" to enable interoperability and integration between them [2]. Inmon's model, in contrast, relies heavily on a single centralised data warehouse. The differences between Kimball's model and Inmon's model are given in more detail the next section.

In addition to the idea of a centralised data warehouse, in 1998, Inmon worked on the development of the Operational Data Store (ODS). The purpose of this database, as seen in the previous chapter and the works of Brazhnik and Jones [6], is to act as a facilitator for the integration of data from multiple sources. An ODS would then allow transformations and the Extract-Transform-Load (ETL) process to upload information to the data warehouse. Indeed, an archaic version of an ODS and ETL had previously been envisaged by Devlin and Murphy whereby an

interface to operational systems and a transmission manager moved information from operational systems to the data warehouse [153].

Both Inmon and Kimball continued to perfect their models in the 2000s and implementations of both together with hybrid models of data warehousing began to appear. Inmon's approach tends to be more costly than Kimball's but it may suit the sustainable needs of large enterprises. Inmon's approach may, in addition, be better suited for developing warehouses where data queries are not repetitive or recurrent (for example, monthly reports).

Also in the early 2000s, Dan Linstedt [160] introduced the concept of a data vault. This is a hybrid approach, perhaps the most prominent to date, and it encompasses "the best of breed between 3rd normal form (3NF) and the star schema"[160]. In this context, 3NF represents the normalisation form achieved with the relational model and the star schema is the simplest style of a dimensional data mart where a central main table (fact table) is surrounded by dimensional tables that allow operations such as slicing or dicing. In brief, the data vault model attempts to separate informational elements that are prone to change from more static ones, and to represent relationships dynamically using link structures[160]. As a result a vault model is more scalable, flexible, has a dynamic structure, reduces data redundancy, and copes with data's rates of change [160]. Such a model is particularly useful in coping with change in the environments and would suit large business implementations.

Later, in 2010, Inmon introduced the second generation of data warehousing (DW 2.0) [161], concentrating efforts on the life cycle of data within the warehouse, the new types of data (for example, multimedia) including unstructured data, metadata, and the "hunger" for integrated corporate data [161]. The data vault model was also revisited in 2013 but, at the time of writing, little research was found in the literature and, to our knowledge, no significant implementations were reported in research papers.

Despite the efforts to standardise data warehousing methods, models and architectures, building a data warehouse continues to be an *ad hoc* exercise, tailored to individual problems and domains. The following section discusses key implementations of data warehousing technologies in healthcare, which has traditionally been one of the most challenging environments [21].

3.3.2.2 Summary of Data Warehousing in Health Care

In 1966 the Multi-User Multi-Programming System (MUMPS) system was developed at the Massachusetts General Hospital, Boston by Greenes and Pappalardo [162]. MUMPS was an interpreted programming language incorporated into a hierarchical database file system [163; 164] and it worked as an application program for clinical data management. This OLTP-like system was optimised for high-throughput transaction processing and, at its core, lay a database that later, in 1989, functioned as a clinical-data repository and an on-line data warehouse, with a querying software called ClinQuery [164; 165].

Another early system had been developed in the US in 1975 at the Latter-day Saints (LDS) hospital in Salt Lake City, Utah. The HELP system, today, is one of the oldest active systems [166] and it was arguably the first to integrate clinical data with a decision support system (DSS) [166; 167]. This system was the first in healthcare to employ the Entity-Attribute-Value (EAV), a data model available before commercial relational databases based on Codd's model. The EAV represents data vertically, where each row contains three columns describing the entity, the type of information, and its value [22; 168].

Clinical data management systems continued to be developed [164; 169; 170; 171] and Codd's description of systems as *ad hoc* and home grown also held true in healthcare. Most database systems and data warehousing solutions, even implementing published data models [163], were mostly in-house solutions for

particular problems. It would become apparent later that the individual problems and solutions had similarities throughout institutions [164].

Another example of a home-grown system was made known later, in 1997, after a report of a successful clinical data warehouse was published [172]. In this study, datasets were extracted from the production system at Duke's University Medical Center, North Carolina, cleaned and mined using exploratory factor analysis. This is also one of the early publications in medical data mining. Similar reports continued to follow [164; 173].

It is then, in 1998, that a first survey on clinical data warehousing emerged in the literature [142; 174]. The authors acknowledged that the use of data warehousing and other computing technology in healtcare had been hindered by a lack of understanding of what benefits this technology might bring. It was natural that only the institutions that had close links with computing scientists or engineers were able to implement their own solutions.

As reported in the survey, one of the few European efforts in implementing clinical data warehousing systems at the time was at the Turku University Hospital in Finland [174; 175]. The reported system would integrate data from several hospital and laboratory systems to provide a broad view of clinical data suitable for research [175]. Their first studies successfully combined blood readings with drug prescriptions and it resulted in novel clinical results showing how certain drugs interfered with particular blood tests [174; 175].

Since then, and as predicted by the authors of the survey, clinical data warehousing have continued to emerge [142; 174] and the development of technology, data standards, ontologies and other works in health informatics have contributed to enrich the literature and advance implementations. However, in the early 2000s, Sujansky reported that the "decentralized nature of our scientific communities and healthcare systems has created a sea of valuable but incompatible electronic

databases" [21]. Sujansky's work highlighted the perennial problem of heterogeneous database integration in healthcare, and as discussed in the previous chapter, this problem continues to exist.

Efforts to address this issue, in the early 2010s and up to date, have focused on developments of biomedical ontologies, semantic web technologies and also frameworks for integration of clinical data with other research data. Initiatives such as the health level-seven, the Clinical Data Interchange Standards Consortium [176], and the recent DW4TR [177] and i2b2 [178] have seen fuitful applications. The latter is a system designed in the US for cohort identification of clinical and research biology data, allowing users to perform queries on integrated health and research (biological) data. The system was later extended to include reporting and visualisation tools and its software is now available as open-source, allowing other institutions to implement and test it. At the time of writing, i2b2 is being tested at a single UK site, at the University of Leicester, in order to help with data integration and selection of cohorts. In 2011, an evaluation of four different implementations was carried out in Germany [179]. Although issues were reported with regards to runtime, querying restrictions and setup process (in particular where manual interventions were needed to construct valid ontology metadata), the authors reported that the i2b2 system was a valuable platform for integration and data query tasks in four different use cases [179]. The i2b2 continues to improve its software with the focus of integrating clinical and research data (such as genomics data) but it also continues to depend on existing implementations of OLTP (such as electronic health records) and OLAP (such as local data warehouses) systems. Furthermore it has been noted [180] that the i2b2 was designed for querying and processing purposes and it does not provide powerful means for the loading of data into its own database.

In 2013, Inmon reviewed data warehousing in healthcare and highlighted two overriding architectural differences when compared to corporate environments

[181]. The first difference is that in healthcare environments, unlike corporate environments, the information can also be generated outside of the confines of the provider. The second difference is the high volume of non-numeric, non-transactional data with which commercial database systems are not designed to work. There are also, similarities: the high volume of data (some of it transactional), the need to look at data holistically (and "patient-centrically"), the need for data transformation, and the need to support day-to-day activities as well as reporting or research.

The use of frameworks such as the i2b2 or DW4TR should help to overcome the integration of health data and the development of databases or data warehouses for research. However, there is still little evidence as to how such systems and implementations can co-exist with hospital operational systems in the long term.

The following section discusses the current approaches on data warehousing architectures and methodologies further and introduces the technical background and foundations of the work presented later.

3.3.2.3 Architectures and Methodologies

Before delving into a particular data model, architecture or methodology to build a data warehouse, it is customary to adhere to an overall design philosophy guided by either Inmon or Kimball's views [182].

Inmon's model sets out to deliver a sound technical solution based on proven database methods and technologies while Kimball's focuses on delivering a solution that makes it easy for end-users to directly query the data and still get reasonable response times [2]. Table 3.12 shows a summary the differences between the two models.

In this chapter, the purpose is to build a storage infrastructure that holds the

Characteristic	Kimball	Inmon
Nature of the organization's decision support requirements	Tactical	Strategic
Data integration requirements	Individual business areas	Enterprisewide integration
Structure of data	Business metrics, performance measures, and scorecards	Non-metric data and for data that will be applied to meet multiple and varied information needs
Scalability	Need to adapt to highly volatile needs within a limited scope	Growing scope and changing requirements are critical
Persistency of data	Source systems are relatively stable	High rate of change from source systems
Staffing and skills requirements	Small teams of generalists	Larger team(s) of specialists
Time to delivery	Need for the first data warehouse application is urgent	Organization's requirements allow for longer start-up time
Cost to deploy	Lower start-up costs, with each subsequent project costing about the same	Higher start-up costs, with lower subsequent project development costs

Table 3.12: Specific characteristics favouring Kimball or Inmon's model [2].

Norwich Research Stroke Register and is capable of answering diverse research-led queries. Clinical data on strokes already exists in operational databases or legacy systems but additional data needs to be fed from other systems (such as the laboratory or administration system).

While Kimball's solution seems appropriate from the point of view of "business size" by concentrating on an individual business area (strokes), it is limiting from the point of view of data structure as the dimensional model favours recurrent or repetitive queries. Inmon's "relational-friendly" approach satisfies a much wider range of data queries which is indeed what is expected from the stroke register

warehouse for research.

The above general design philosophies are discussed in detail and with respect to their methodological and architectural implications in the next sections.

Selecting an Appropriate Methodology

Inmon and Kimball also propose two distinct yet complementary approaches to data warehouse development. Inmon proposes a top-down methodology where the primary interest is that of ensuring that the technical solution works [2]; it is based on the spiral software development methodology [183], is goal-driven and supports an organisation long-term in turning strategy into action [184]. Boehm's spiral methodology can be succinctly summarised by its quadrants [183]: Determine objectives, alternatives and constraints; Evaluate alternatives, identify and resolve risks; Develop, verify next-level product (includes implementation and design validation); and Plan next phases (includes evaluation).

In contrast, Kimball's bottom-up approach, where multiple data marts are built first and then integrated, has a short-term focus. It is not as technically demanding and, as a result, engages end-users [2]. Furthermore, the bottom-up approach "exploits the database and is suited for *tayloristic* measurement" [184]. Kimball's methodology comprises, in its highest level, four planning and design steps [185]: Requirements & Realities, Architectures, System Implementation, and Test & Release.

A further distinction has been made between data-driven and user-driven approaches [184]. Inmon argues that warehousing environments are data-driven as opposed to being built purposely for users' needs [156; 184]. A data-driven development methodology has also been recommended for data mining and data exploration purposes [184] and would seem appropriate for the work carried out in this thesis.

Overall, methodologies share a common set of tasks: requirements analysis, data design, architecture design, implementation, and deployment [182]. Other research papers on the methodological aspects of data warehouse design were reviewed [2; 182; 184; 186; 187], yet for the purpose of the work carried out here, the work of Szirbik *et al.*[3] stood out as most relevant.

Szirbik *et al.* noted that most methodologies typically only cover partial aspects and too many gaps are left; however, attempts to fill these gaps would also reduce the comprehensibility and usability of the framework [182]. The authors introduced six methodological steps specifically tailored for the development of medical data warehouses [3]. The methodology is an extension of the Rational Unified Process (RUP), an iterative software development process framework[188], but it may be used on its own as a streamlined version of RUP [3]. The six-steps of the methodology are summarised in table 3.13.

The methodology used to develop the stroke data warehouse is introduced later and is based on the guidelines set out by Szirbik *et al.* and the above-mentioned common tasks across data warehousing methodologies.

Selecting an Appropriate Architecture

As previously discussed, the main difference between Inmon and Kimball's architectural approaches is the "centralisation" of the data warehouse. In addition, the data vault architecture is similar to that of Inmon's but would introduce a more complex data model which would not bring any particular benefits to a smaller scale implementation such as the stroke register. As a result, the data vault architecture and model were not explored further. Other architectures were considered but only those relevant to the work carried out with the stroke register are presented here.

A centralised approach would be more appropriate for a system such as the stroke

Step	Description
1. Recognise and resolve the most critical issue(s)	Identify and prove that the hardest task can be solved; decide whether to go ahead with the project.
2. Identifying the scope and granularity of data sources	Identify and analyse the data sources their data and complementarities; identify inconsistencies; validate data collection.
3. Identify the required data	Determine data requirements to achieve stakeholders' results; reduce the scope of the data and simplify schemas.
4. Ontological alignment	Build an ontology that maps relevant terms in the scoped universe of discourse; semantic alignment between data sources schemata and the repository.
5. Determining update policies	Establish update policies for each local source and estimate the costs involved; procedures for data transformations.
6. Validating and fine-tuning	Real-life data gathering; test the correct information is deposited in the repository; collected data is analysed and compared with the results of the analysis of the sources; correct any problems and fine-tune the repository.

Table 3.13: A summary of Szirbik *et al.* six methodological steps [3].

register because only one main source of data is expected and hence, a main data table would contain most information, i.e. stroke admissions. The additional data sources are then expected to provide limited amounts of information (blood readings, follow-up or comorbidities) to individual records in the main table. Furthermore, the work carried out in the previous chapter already resulted in an operational data store with some degree of integration. As a result, a structure where multiple data marts are created for each data source would not, in this case, facilitate the development of a warehouse.

Figure 3.17 shows two possible architectures, adapted from [189], where the extract-transform-load (ETL) process is responsible for propagating information

from the operational sources into an integrated or pseudo-integrated collective. In the first architecture (figure 3.17 at the top) that collective is the data warehouse itself whereas in the second architecture it is the operational data store (ODS) that acts as an intermediate step to reach the data warehouse.

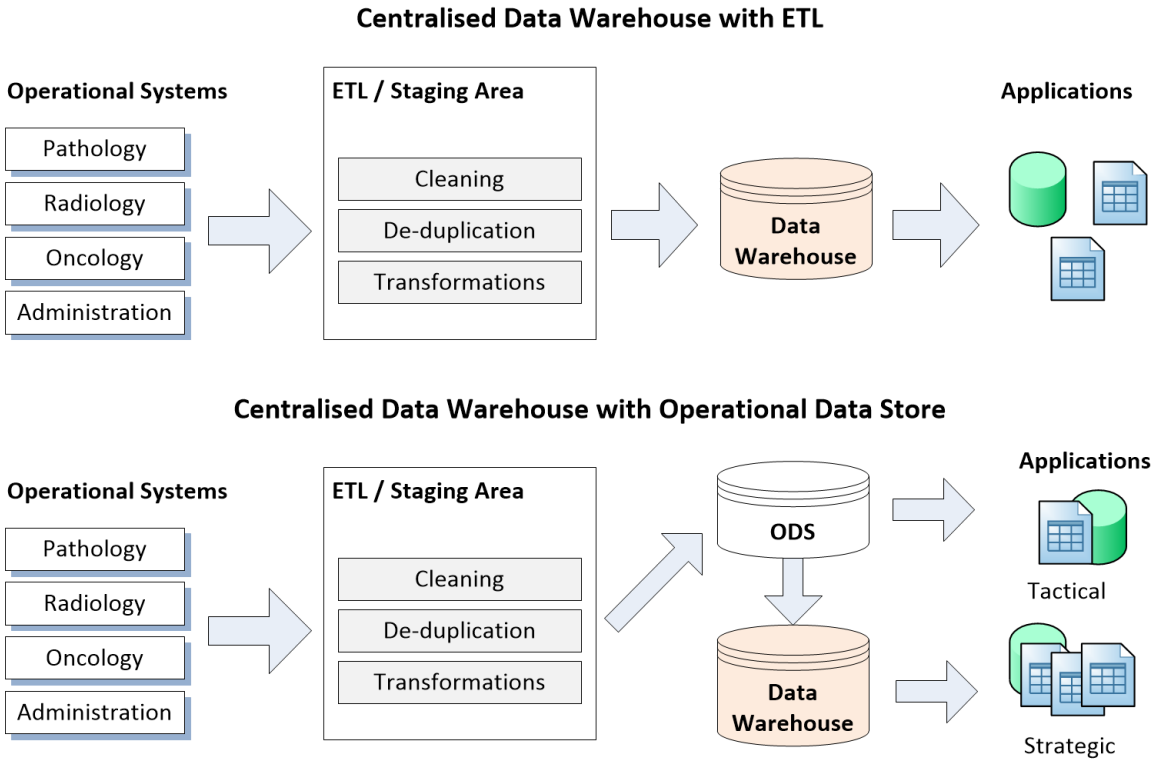


Figure 3.17: Data warehousing architectures with a centralised warehouse.

One of the benefits of using the latter architecture is that it enables further applications, such as reporting, directly from the ODS. As seen in the previous chapter, such applications are of particular interest and an ODS had already been developed. The centralised data warehouse with operational data store would therefore be most appropriate.

A further extension of the above architecture is to create data marts directly from the ODS and/or central data warehouse[189]. This would add benefits with regards to data control and manipulation. Here, data marts are considered to be

a possible application of the data warehouse and this is discussed again later.

Selected Model, Architecture and Methodology

In summary, a hybrid approach was chosen to build the stroke warehouse where:

- the data model should be fundamentally relational yet pseudo-dimensions or additional normalisation could be added should they prove beneficial for query optimisation;
- the architecture is based on a centralised warehouse (with an operational data store) rather than on multiple data marts and an integration bus;
- a methodology to build and maintain the data warehouse should be based on the common of tasks reported in the literature (requirements analysis, data design, architecture design, implementation, and deployment) and guided by the methodological steps tailored to clinical environments set out by Szirbik *et al.*

3.3.2.4 Methodological Steps

In the previous chapter, a framework for data collection from multiple sources was introduced. The output was a study dataset and metadata for each identified data source. An Operational Data Store (ODS) was fed both metadata and study data from several sources and from it it is possible to link, transform and derive domain-specific datasets for particular studies.

This section describes the methodology used for the development of a domain-specific clinical data warehouse with data from several hospital information systems. The methodology was based on previous literature and reworked with the experience from the case study. Previous work pointed out that data-driven and simple methodologies produce better results and maximise users adherence. The

common steps in data warehousing and software development methodologies were used as a foundation. The guidelines from Szirbik *et al.*[3] and experience from the case study were used to further develop each step and the overall structure.

The work carried out here acknowledges that the issues of time, resources and project management are important and will be discussed when appropriate, but the methodology concentrates primarily on the delivery of a sound technical solution and on mapping the key steps to achieve it.

The methodology comprises five incremental steps (each achieving a conceptual milestone that builds upon the previous step) forming an overall iterative cycle describing the life cycle of the clinical data warehouse. Each step contains three states within itself: design, test and completion. The design state indicates that work is being carried out to achieve a particular milestone, the test state indicates validations of that work are being carried out, and the completion state indicates successful achievement of a milestone. Upon achieving a milestone (i.e. completing a step) the next step can begin. However, in steps where backtracking is allowed, the design state of the previous step would be revisited. In such cases, the completion state of the original step is not achieved.

Figure 3.18 illustrates the five steps and their interactions with the ODS (dashed lines) and the data warehouse (dotted lines). The staging area comprises the ODS, an "adjuvant" database for testing, a metadatabase, a transformation process (Extract-Transform-Load), and the external data sources. In the case study, the stroke data collection had already happened as described in chapter 2, and hence, a first version of the ODS was already available. The separation between the ODS and the "adjuvant" database is important as they serve different purposes. The role of the first is to store the data retrieved from the sources whilst the latter deals with building and testing a workable solution for the data warehouse (data model, schema, and architecture). Overall, the methodology emphasises the supportive role of the staging area and, in particular, the ODS, in the devel-

opment and deployment of a data warehouse. Each step of the methodology is described in detail below.

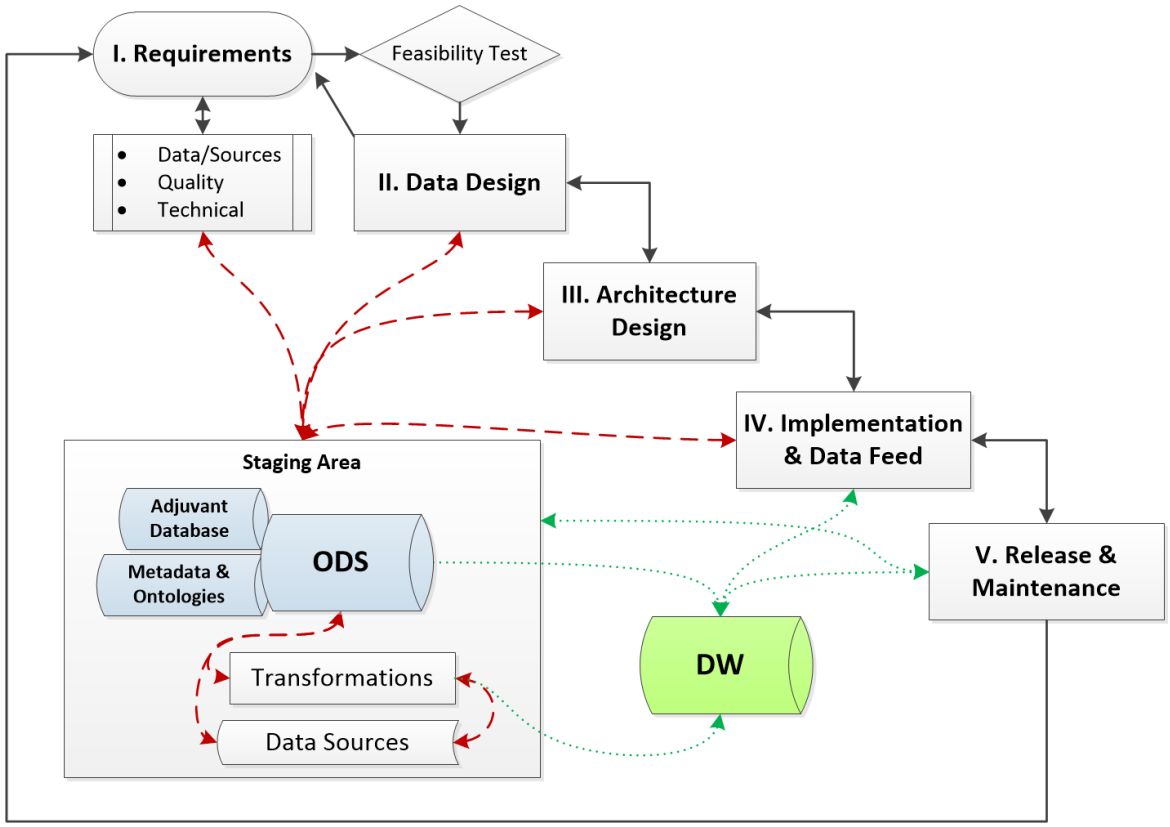


Figure 3.18: Methodology to build and maintain a clinical data warehouse.

I. Requirements

The requirements step is regarded as the phase where investigations take place to understand the needs for the system development, constraints, and intended purpose.

The intended purpose is to build an infrastructure (a data warehouse) for the storage and extraction of data marts for research. The data warehouse should be a centralised and integrated repository of stroke and TIA cases

with additional laboratory and comorbid information. Details of the source systems, data and other technical constraints were already given in the previous chapter when following the methodology that created the operational data store.

Previous work from Szirbik [3] suggested that the first step would be to identify and solve the most critical requirement. When developing the stroke DW, this was indeed found to be a pertinent step and one that should, together with data requirements, data collection and quality comprise the larger requirements step.

Identify the most critical requirement

The most critical requirement in the stroke study was to prove that it was possible to extract and link data from an existing biochemistry system to a set of stroke events. Without this, the development of the stroke data warehouse would not be possible. This requirement, however, only had a workable solution after the next sub-steps were carried out.

Identify the required data and sources

The identification of operational systems where relevant information is present is one of the first considerations. Meetings with the stakeholders were held to determine in a first instance, the required data, the systems where this data is likely to be present, the technical feasibility of accessing the system and data, and the likely support and collaboration from the system's owners.

Five data sources were identified: the patient administration system (PAS), the biochemistry system (LAB), a legacy system containing strokes from 1996-2008, the current stroke database system, and the transient ischaemic attack (TIA) database, all having different schemata. The national tracing service system (NSTS) reported in the previous chapter became obsolete by the time this work was carried out and, for that reason, it is not included here as another data source. The information on this system (follow-up, death

registration) was replaced by the information available from the linkage exercise carried out in the previous section and, from 2012, PAS was used.

Quality and scope of the data and sources

Much of the information regarding the sources had already been gathered as part of the data collection methodology in the previous chapter. As such, metadata and data flow diagrams were produced based on four steps defined in Chapter 2: system understanding, data understanding, extraction preparation, and extraction & evaluation. This resulted in a sound understanding of the systems and their data, and in turn, the solution for the extraction of biochemistry data. At this stage, only a sample of the data from each system is needed to ensure feasibility and inspect data quality.

Data quality issues were found with some of the identifiers in some of the sources and a systematic approach to correct them was devised. For example, a sequence of queries would identify anomalies in the hospital number and attempt to correct them. Key identifiers across the sources included hospital number, date of birth, and date of event. The first two were sufficient to correctly ensure a patient's identity and the latter ensured the time variancy aspect needed for any data warehouse (i.e. the moment in time when a particular event or transaction took place).

The Requirements step ends with a report and a "go ahead" decision from the stakeholders. Their decision should take into account the most critical requirement but also the feasibility report on the above-mentioned points.

II. **Data Design**

This step is concerned with the development or implementation of an existing data model and ontology. The data model represents the structure and integrity of the data elements and sets of elements and can be based on relational, dimensional or other such as the Entity-Attribute-Value (EAV)

model. It is the data model that determines the expressive power of the data queries through data engineering languages such as SQL [190] and it is also often represented by database schemata.

Ontologies, however, can be seen as a set of formulae intended to be always true according to a specific conceptualisation [191]. An important application of ontologies is interoperability and, indeed, an ontology may not only facilitate the integration of data but also improve communication with the stakeholders [3].

Both data models and ontologies must consider the structure and the rules of the domain that one needs to model [190]. Aligning ontologies with a data model is therefore a critical milestone to produce a data warehouse. Recent efforts in the development of biomedical ontologies is relying on semantic web approaches with technologies such as the Web Ontology Language (OWL) and Resource Description Framework (RDF) [192].

In the stroke case study, the previously gathered metadata allowed the creation of a simple ontological table which can be regarded as an extended data dictionary. Because this is a domain specific database, efforts to model an elaborate ontology were not deemed necessary and a descriptive dictionary was considered sufficient at this stage. For each data source, a data dictionary was created. A description of the tables and how they relate to each other is given in figure 3.19.

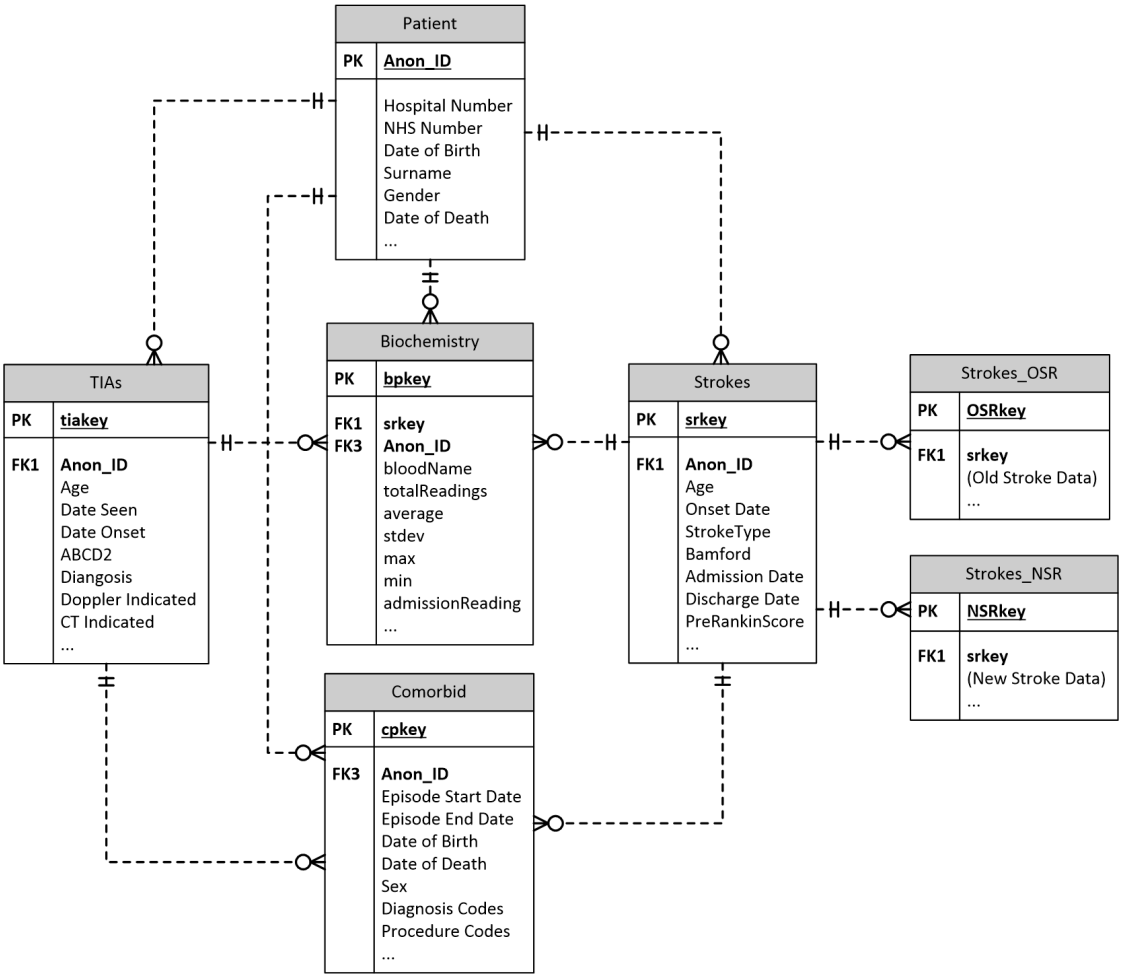


Figure 3.19: Stroke data warehouse schema.

Schema Development

The overall purpose of the data warehouse is to gather information on stroke and TIA episodes and as such, the core tables contains all stroke and TIA admissions. The staging area and ODS (as seen in Figure 3.18) served as development and testing grounds.

The first challenge was to map the legacy stroke database to the current one, where a laborious consultation phase resulted in additional metadata.

50 common attributes were also identified as present in both the legacy (OSR) and the current system (NSR). In turn, the intersection of these two data sources, upon transformations to match formatting, resulted in the main Strokes table, central to the architecture of the warehouse. The remaining attributes from each of the two sources were accounted for in two additional tables (Strokes_OSR and Strokes_NSR), linked to the main strokes table. The Patient table was another important table created. It contains patient sensitive information (name, date of birth, hospital and national health numbers) that allows the correct identification of a patient. The exercise carried out in the previous section allowed validation and for further information to be introduced regarding dates of death.

Two other tables were created, one for biochemistry results and another for comorbidities. These are row-modeled tables containing transactional data where each row represents biochemical readings events or hospital admissions and their respective comorbidities' coding. The linkage exercise carried out in the previous section was crucial in identifying appropriate biochemistry readings for a stroke episode; however, further transformations were needed to create additional "dimensions" and reduce granularity.

A first approach for the biochemistry data relied on the retrieval and storage of every transaction for the cohort of stroke patients. However, due to the sheer size of transactional data and computational effort needed for querying, the table was modeled differently to reduce granularity. A more elegant solution relied on the summarisation of a stroke or TIA episode into a single row, for each biochemistry test. The summarised elements (see table 3.14) were carefully selected with the clinical team so that they are transversal across a wide range of biochemistry tests. This could be seen as a pseudo-dimensional table, effectively enabling OLAP features. Nevertheless, as discussed earlier, any dimensional approach has limiting effects on analyses. For example, a time series analysis of a particular blood test

Attribute	Description	Data Type
bloodID	Primary key, unique for each row.	Natural number
patientID	Patient ID, foreign key.	Natural number
bloodName	The name and internal code for a particular blood test.	String
totalReadings	Total number of blood readings during the stroke admission.	Natural number
average	Computed average of all readings.	Real number
stDev	Computed standard deviation for the above average.	Real number
max	Maximum value found within the stroke admission.	Real number
min	Minimum value found within the stroke admission.	Real number
admReading	Blood reading closest to admission.	Real number
admReadDate	Date of blood reading closest to admission.	Date
disReading	Blood reading closest to discharge.	Real number
disReadDate	Date of blood reading closest to discharge.	Date

Table 3.14: The biochemistry data table attributes.

within a patient's episode is not possible. However, including attributes such as average, standard deviations and range for blood readings enables further analyses and was considered sufficient.

The relations between tables along with some of their attributes are given in figure 3.19 and further details of the attributes are given in Appendix A.

III. Architecture Design

Data design and its subsequent step, architecture design, are closely linked in that the first defines the representational model of the data and the latter focuses on the design and build of an infrastructure that is compatible with, and optimises the use of, such a data model and schema. A number of iterations between architecture and data design steps can be expected until both the data model and architecture are in agreement.

The architecture affects hardware, software, security, coding practices, personnel, and operations. Issues with these could compromise data quality and preserving or improving the quality of the data is paramount at all stages of data warehousing development.

With regards to hardware and personnel, the preferential database management system was Microsoft Access due to its availability on-site, the already in-use services database operated with the same RDBMS software, and the fact that no significant additional training was required as staff members are aware of the software. This system would also speed the delivery of the first versions of the data warehouse as well as minimise hardware requirements. Nevertheless, in the long term, it is expected that the DBMS changes to a SQLite, a portable and powerful DBMS (as seen in the previous section) or an i2b2-compatible DBMS such as Oracle, PostgreSQL or SQL server. However, the first implementations will be carried out using MS Access and software to access the DBMS should use Open Database Connectivity (ODBC) or a Microsoft Access driver. A split MS Access database architecture is often used to improve performance and simplify maintenance; it contains two MS access databases, one with the data tables and another with a front-end and queries. In the case study, the stroke register database contains the data, query templates and an integration screen, but a separate querying interface software was developed in Python to run detailed back-end queries and automatically produce and export CSV files with results. Relevant operations in the querying software were loosely determined at first but, over time, and based on study needs, this became more specific and constrained. It is during the architecture step that the design and improvements of the software are carried out. Likewise, the hardware infrastructure should also be thought of, but in this case study given the technologies used, no particular hardware was needed.

Perhaps the most laborious part of this step is concerned with building

Extract-Transform-Load (ETL) systems and processes that feed data from the data sources into the ODS. The methodology developed in the previous chapter dealt with most of these issues and as a result, standard ETL processes from the required hospital sources were readily available. Nevertheless it was necessary to agree on coding practices, such as those previously identified in the data model and schema, and on data updates. It was determined that each version of the data warehouse would represent a cross-section of the strokes and TIAs at the hospital and have a determined follow-up date for all the collected data. In effect, in every version of the database, new data is added and the follow-up dates are extended. It was not possible to completely automate updates but processes were developed and some degree of automation was provided in the extraction and transformation tasks. Indeed it is argued that it is not desirable to completely automate the ETL process in an environment where the data sources and systems are volatile. As a result, such automation would likely result in a continuous dependency of additional technical expertise. Conversely, the development of semi-automated processes allows existing hospital staff to adjust their processes to changes and to better understand the origins of their data.

A full diagram depicting the stroke data warehousing architecture developed here is given later in the discussion and evaluation.

IV. **Implementation & Data Feed**

This step is concerned with the implementation of the infrastructure and data model developed in the previous steps and with populating the data warehouse with records. The implementation and data feed should also be considered a validation step in its own right, where the milestones achieved in the previous steps are finally implemented and tested.

In the stroke case study, the implementation of the data warehouse was

carried out seamlessly. Both the database and software (implemented as part of a portable framework) were stored in a research laptop and later made available in a secure shared folder in the hospital's intranet under controlled access. Fine tuning was required when populating the warehouse with records. The previous steps were often revisited in order to adjust the ETL processes (for example, introducing further constraints or deduplication) and data types (for example, inconsistencies with dates or unwanted conversion of real numbers into natural numbers by third party software). There were no issues with semantic heterogeneity at this stage as they have been resolved previously.

Quality handling is a key part of this step. Processes should be in place to ensure any retrieved data meets quality standards and a number of queries were built to this end. For example, when populating a later version of the data warehouse, summary queries allowed the identification of a sudden drop in certain biochemistry records from a particular time-period. This was the result of coding changes in the pathology laboratory IT system of which we had not been aware. The issue was resolved upon consultation with pathology IT staff and by backtracking to the previous step. Logs of all transformations and dates when data were fed into the warehouse were kept.

V. **Release & Maintenance**

Upon a successful implementation and data feed, the data warehouse is released with a new version after further functionality checks are carried out. This step may involve the logistics of moving a fully working solution to an area where it is accessible by the end-users. In the case study the query engine required additional software (a portable version of Python 2.7) and the database requires hospital computers to have a version of MS Access.

The methodological steps to develop the data warehouse (as seen in figure

3.18) should create a continuous looped cycle while the project and/or the data warehouse remain active. When the Release & Maintenance step is reached, the cycle becomes idle (in its testing state) until either a problem is found, additional data is needed, or a new request for a feature emerges. When a data top-up or refresh is needed, the previous step is revisited, and when a new feature request or a problem is detected the methodology returns to its initial step to assess the requirements for a particular change or problem.

3.3.2.5 Results and Discussion

The above methodological steps were followed and, at the time of writing, two versions of the research stroke data warehouse have been completed and a third was being finalised for release. In the first version, the old and the new stroke registers were matched to produce the main strokes table, whilst keeping the additional information in separate tables. This was a time consuming exercise as it required significant consultation work with clinical and administrative staff as well as the integration of several data fields (semantically and in format). In this version, the most laborious steps were Requirements and Data Design.

The second version included data from the biochemistry system (LAB source), comorbidities and follow-up (PAS source), the TIAs table, and the Patient demographics table (derived from PAS and NSTS). In this version all the required tables were included but consistent database relations and anonymisation processes (for example, automatically assign an anonymised number to each new patient) were not fully operational. Nevertheless, because of time constraints and the need to demonstrate and discuss progress with the end-users, this version was released as such. Linkage and additional cleansing of some of the data were carried out after any required study datasets were exported from the ware-

house as needed. In this version, both the Data Design and the Architecture Design were the steps requiring most resources and time.

Version three was, at the time of writing, under a second iteration of the Implementation & Data Feed step where final adjustments were being made with regards to refreshing data. The first iteration of this step revealed issues with data quality as the LAB system recently changed the coding of some of the blood tests, resulting in a loss of information, that was identified by running data quality queries in the staging area. The LAB system experienced further issues with their system and halted any back-end operations, resulting in a delay in the delivery of a fully working version. Nevertheless, version three also included querying interfaces for the warehouse provided by a MS Access form (built in the data warehouse) and a separate Python software was developed to facilitate the export of linked data in a CSV format via ODBC. Further work is also underway at the time of writing to provide additional patient demographics data to the TIA table. The most laborious steps in this version were Architecture Design and Implementation & Data Feed. A full description of the data warehouse tables and number of records currently in version three is given in Appendix A.

Across the three versions, a shift in the time spent on each step was observed. This was expected as the work needed in the second and third versions focused primarily on architectural designs and feeding new data. Nevertheless, at the time of writing, further work is planned to improve the functionality, embellish the querying interface and feed additional data. An additional challenge has also been identified that will require further work across all steps. A new database system was recently introduced in the stroke services department and, as such, new methods for retrieval and matching of this data source with the data warehouse are needed. A preliminary inspection of this new system reveals that similar ETL processes are required and the methodology presented here provides the necessary methodological steps to continue to bring new stroke data into the

data warehouse.

Overall, the methodology presented in Figure 3.18 accounted for all operations and interactions needed to develop and maintain the stroke data warehouse and its three versions. A summary of how the methodology steps presented here are aligned with Kimball's methodology and Szirbik's work is given in table 3.15. Previous work has been carried out on the alignment of Sbzirks's steps and the Rational Unified Process (RUP) [3]. The latter, in turn, fits easily with Bohem's Spiral model. For these reasons RUP and Spiral models are not shown in Table 3.15 and only the two methodologies specifically tailored to data warehousing (Kimball's and Szirbik's) are included. The methodology presented here, upon iterating through a first full cycle, becomes a data-driven methodology in that it builds on previous efforts utilising both code and processes that have already been developed. This is similar to the Spiral model and has been previously discussed by Inmon [2].

Step	**Szirbik's Step**	**Kimball's Step**
I. Requirements	Prove hardest task can be solved; Quality and scope of the sources; Validate Data Collection.	Requirements & Realities.
II. Data Design	Data Requirements; Ontological Alignment.	Architecture.
III. Architecture Design	Establish update policy; Develop software to adjust granularity.	
IV. Implementation & Data Feed	Validate and Fine-Tuning.	System Implementation; Test & Release.
V. Release & Maintenance		Test & Release.

Table 3.15: Alignment of the methodology steps with Szirbik's and Kimball's approaches.

As seen in table 3.15 the methodology follows Kimball's overall planning and design steps, which had been previously said to be simpler and engaging for the end-users [2]. The methodology presented here allows several versions to be

released with different degrees of completeness and this was found to be advantageous when engaging with domain experts and end-users.

Also similar to Kimball's approach, multiple data marts were built first into a pseudo-integrated ODS and only later were they fully integrated in a working data warehouse. Kimball's overall approach is ideal for domain specific projects such as the one presented in this section. However, the proposed approach to build the stroke data warehouse did not rely heavily on the dimensional model. Indeed, Inmon's model was found to be more robust for a clinical data warehouse where no predefined or recurrent database queries are expected. As such, the solution presented here is a hybrid one between Inmon's relational model and technical focus, and Kimball's methodological principles and bottom-up approach that work well in smaller environments within a large organisation.

In addition to this, Szirbik's methodological steps were extremely helpful in validating and refocusing the work on clinical data warehousing. Indeed it was found that identifying and proposing a solution for the hardest problem was the most important starting step. Without a demonstrable technical solution to retrieve and integrate biochemistry and comorbidity data (discussed here and in the previous chapter) this data warehousing project would have not been possible. Furthermore, it was important to take into account Szirbik's step on ontology building when developing a dictionary to map terms and concepts and to facilitate data integration.

An important feature of the methodology presented here is the staging area and its interactions with the methodology steps. Throughout the steps, the staging area was used to retrieve and store the required data in an ODS, to design and test the data model and conceptual representations, or to design and test ETL processes. The "adjuvant" database allowed the architecture, data model and overall database software to be designed, put together and tested before a working version of the data warehouse was released. It was particularly important to

keep the retrieved data in the ODS in its original form, not affected by data transformations, so that it can be accessed for reference, testing, validation, and data lineage analysis.

The full life cycle of the stroke register data warehouse, developed along with the above methodology, is depicted in Figure 3.20. The life cycle shows the interactions with end-users (researchers), a data warehousing developer and manager (practitioner) and the system. Three major cycles are observed: the update cycle, responsible for feeding new data into the ODS and the data warehouse; the integration cycle, where the information is integrated from the ODS into the data warehouse using a dictionary or an ontology; and the reporting cycle, where the data warehouse is used by the end-users to extract study specific cohorts. This higher level life cycle is compatible with the above methodology and it represents the overall interactions between users, practitioners and systems.

3.3.2.6 Conclusions

This section reviewed the state of the art in data warehousing both in industry and in academia since the concept was first introduced in the 1980s. An extensive literature survey has provided a factual and complete historical perspective that has not been, to our knowledge, previously covered in specialist books or research papers on the subject. The literature survey was expanded to include the developments in health informatics and this revealed how the discipline has traditionally been at the forefront of technological advances. This review allowed a thorough understanding of the current data warehousing methods and their origins and, in particular, the leading works of Inmon and Kimball.

An understanding of the latter two, seemingly opposing views, together with reports of data warehousing challenges in medicine led to the development of a hybrid approach to clinical data warehouse development that can be applied to

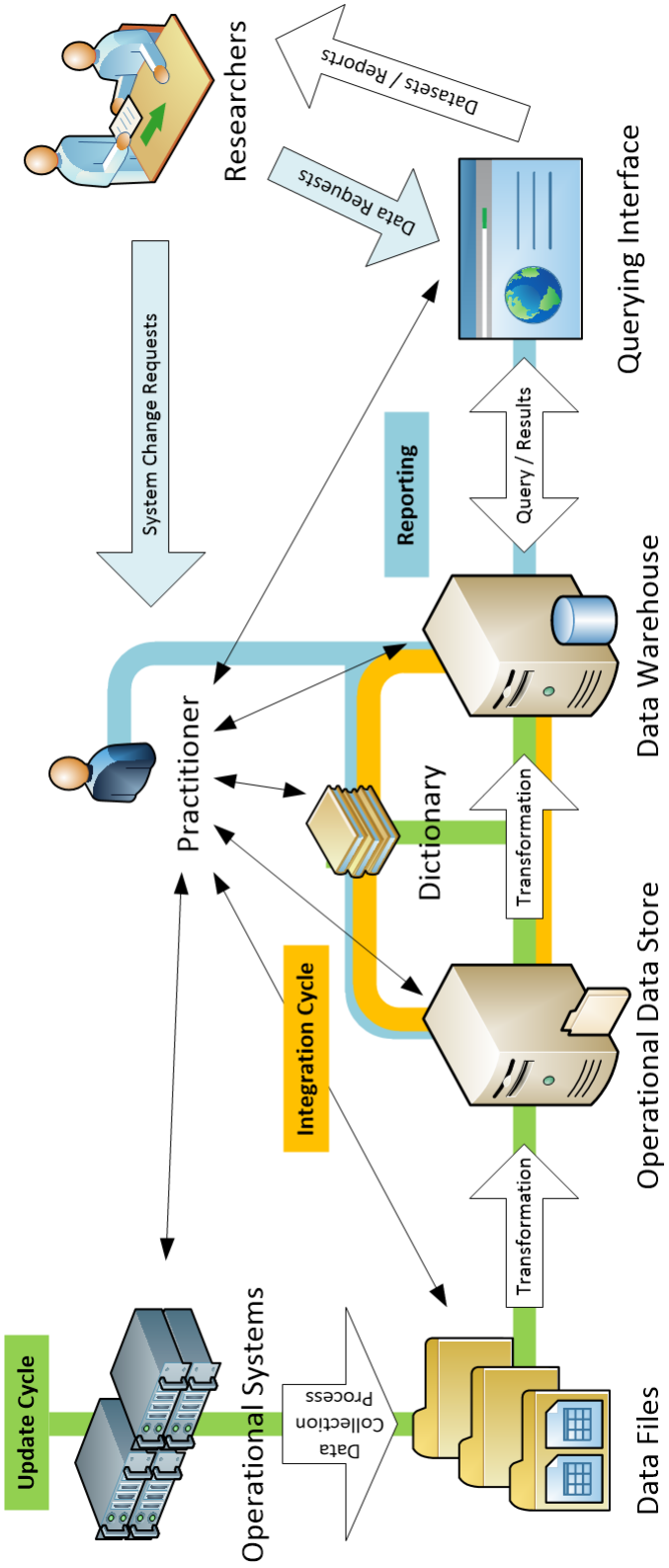


Figure 3.20: Architecture and Life Cycle of the Stroke Register Data Warehouse.

other domains. This complements the work carried out in the previous chapter by introducing a methodology to develop a domain specific data warehouse with limited resources. The work carried out here allowed the Norwich Research Stroke & TIA Register to be developed with over 25 thousand strokes and TIAs (1997 to 2013) and corresponding biochemistry data (from 2003) and comorbidities.

An empirical assessment of the factors affecting data warehousing success [193] revealed that higher levels of system and data quality are associated with higher levels of perceived net benefits. The methodology presented here focuses on data quality, not only by the recursive and iterative nature of its cycle, but also by the exhaustive use of the staging area where operational data is stored together with metadata and conceptual mappings (data dictionaries).

Further work is underway to continue to test and improve the methodological steps. In particular, future work should focus on the improvement of the querying interface by developing functionalities similar to those of the i2b2 system. Indeed, a framework such as the i2b2 could be used as a data warehouse with the proposed methodology and that would likely reduce the amount of work needed in the architectural design step.

Additional work is also envisaged on a proactive linkage and retrieval system, akin to a web crawler used in information retrieval, whereby patient-centric records are found in existing data sources and, when relevant, they are automatically retrieved and mapped to the ODS and data warehouse. An agent-based system of this kind would be placed between the operational systems and the ODS in the data warehousing life cycle. Automatic semantic integration of heterogeneous information sources has already been suggested in the literature [194] yet further work is needed to make these and similar approaches available to data warehousing implementations.

3.4 Discussion

Data preprocessing pertains to a set of actions taken before any data analysis processes start. In this chapter, four different preprocessing methods and techniques were studied. Regarding the case study on prostate cancer, a technique was developed to extract medical concepts from histopathology text reports and another to estimate and input the correct patient age. In the stroke case study, a record linkage technique was applied and a data warehousing methodology was developed to build a unique infrastructure for research. Concluding remarks of each of the techniques are given in the following sections. A discussion of the value of the preprocessing techniques in the improvement of data quality in hospitals is given in section 3.4.3.

3.4.1 Prostate Cancer: Text Extraction and Imputation

In section 3.2.1, a technique was developed to accurately extract Gleason grading from histopathology reports. This exercise was important as current hospital databases do not store this information in a canonical form. The developed algorithms successfully allowed the information to be extracted from the reports and an inspection of the reporting trends in prostate cancer histopathology reports was carried out. The extracted information is used later in this thesis when prostate cancer data is analysed.

Data editing and imputation was reviewed in section 3.2.2 and an algorithm for value estimation was introduced. The algorithm solves the age problem that happens to patients without a matching hospital or NHS record but where demographic information is available. In such cases, patients who are not registered with the local hospital yet have their blood specimens sent there for analysis may have an inconsistent age, recorded at the time of their appointment. The

algorithm uses multiple readings of the patient over time to determine an accurate age. Because the local laboratory is responsible for the analysis of bloods for most patients in the region, this algorithm enables the use of their biochemistry records (no clinical or demographic information) with accurate ages. In this thesis, however, only prostate specific antigen readings were used.

3.4.2 Stroke: Record Linkage and Data Warehousing

Section 3.3.1 dealt with record linkage by undertaking two exercises, the first linking the most relevant biochemistry records to stroke episodes and the second validating recorded locally dates of death. The first exercise allows biochemistry data to be made available for further analyses and clinical research on stroke. The second exercise allows an inspection of the reliability of the date of death field in hospital, in particular for those who died away from hospital. This was important as researchers were made aware of the potentials and limitations of using hospital records for follow-up and, in cases where discrepancies occurred, the stroke records were made accurate. As a result, the two exercises demonstrated the value of using record linkage by providing additional data for research and by validating and correcting follow-up information for the stroke patients.

Data warehousing was reviewed in section 3.3.2, and a methodology was developed for the design and maintenance of domain-specific clinical data warehouses. The case study on stroke resulted in the development and successful implementation of the Norwich Research Stroke & TIA Register data warehouse, containing over 25 thousand strokes and TIAs for a period of over 15 years. The methodology, based on previously published research in the area, accounted for all challenges in the development of three versions of the stroke data warehouse and is applicable to other domains. The operational data store (ODS) introduced in the previous chapter proved crucial to the development of the data warehouse. Nevertheless,

the next chapter will discuss another use of the ODS in the context of the prostate cancer study.

3.4.3 Data Quality

One of the key objectives of preprocessing methods is improving the quality of the data. Hospital information systems and, in particular, routinely collected data, are infamous for the dubious quality and heterogeneity of their data. In particular, the Norfolk & Norwich University hospital (NNUH) data quality strategy identifies eight consequences of potential errors in data [195]:

- breakdown in communications;
- wrong patient being contacted;
- incorrect decisions being made;
- good reputation of the hospital being tarnished;
- inaccurate 18 week RTT (referral to treatment period) status of patients;
- inaccurate recording of cancer waiting time targets;
- hospital rating being compromised;
- incorrect level of income being received by the hospital.

The same document reveals that the central hospital information system (the patient administration system) at the NNUH should ensure that [195]:

- Patients receive the best possible service from the Trust;
- No patient is overlooked for treatment;

- Assistance is given to clinicians and managers in maximising the service for the patient;
- The tracking of patients through their 18 week pathway is facilitated by capturing their RTT history;
- Cancer patients are monitored and tracked through their cancer pathways.

It is important to note the hospital's efforts to ensure the quality of their data [195]. However, the sheer number of departmental systems and domain-specific data being collected results in differences in the way the data is managed by individual departments with different priorities. It was noted that data with significant importance for management, auditing or key operational procedures was substantially checked, and hence, of higher quality. However, as demonstrated in this chapter, "less important" data for hospital operations such as dates of death may not have the same level of reliability. Throughout this chapter, several methods and techniques have been presented that can help to identify issues and improve the quality of this data. It is important to continue to develop methods and techniques that enable the secondary uses of routinely collected hospital data. Further discussions on data quality are given in the next chapters.

Chapter 4

Pathways Modelling, Mining and Visualisation

The previous chapter introduced key concepts, tools and techniques on data preprocessing, linkage and data warehousing. The latter highlighted existing data modelling techniques such as Kimball's dimensional model [159] and the Entity-Attribute-Value (EAV) [168]. This chapter introduces a new data model, tailored to data-driven patient-centric pathways. Together with the data model, a novel framework is proposed for the summarisation, visualisation and querying of complex clinical information as well as the computation of quality indicators. A new graphical representation of the data-driven pathways is also introduced and allows the synthesis of such information. The case study on prostate cancer was used in this chapter yet the methods and techniques have a wider applicability to other domains.

Section 4.1 introduces the background and concepts on clinical pathways, data and process mining techniques and prostate cancer. Section 4.2 introduces system-level paths, examines how patients and data flow between hospital systems and

the utility of data and process mining techniques. Section 4.3 describes the framework for the development of prostate cancer pathways and the software developed for their analysis and visualisation.

4.1 Introduction

"Data modeling is not optional; no database was ever built without a plan" [196]. As seen in the previous chapter, data modelling emerged in the late 1960s as database management systems were introduced but the concepts changed over time. Earlier data models typically represented data horizontally while more recent approaches attempt to represent it vertically or across dimensions. With the advent of the world wide web, additional data models were introduced to arrange information on the web. The Resource Description Framework (RDF) is a semantic web model for data interchange and it is based on statements (RDF triples with a subject, predicate and object) that relate objects together [192]. This is similar to the EAV model as data is represented vertically, and each row contains three columns describing the entity, the type of information, and its value for each row [168]. Such models are flexible for environments where schemas change often and for performing entity-centered queries as no joins are required to retrieve all facts about entities [197].

Aligning ontologies with data models is critical to introducing meaning and enforcing standards. Data models and ontologies must consider the structure and the rules of the domain [190] yet be flexible enough to adapt to new knowledge, particularly in biomedicine where information and knowledge grows rapidly [198]. Recent approaches have relied on semantic web technologies such as the Web Ontology Language (OWL) [192] and the use of metadata is seen as key [198]. However, drawbacks have been identified with EAV-like models. Poor data scanning performance, increased complexity in writing queries and consid-

erable programming are needed to perform tasks that conventional architectures would do automatically [199]. As noted in the previous chapter, there are trade-offs in choosing data models and hybrid approaches may provide overall better results [197].

This chapter introduces a new data model, based on the EAV model and RDF triples, that describes the routes that patients take through care. In a first instance, section 4.2 introduces system-level paths, a chronological sequence of activities or events used to study how patients interact with hospital departments. The utility of current data and process mining techniques was explored in this context. Later, complex clinical information is modelled into pathways (section 4.3) enabling further analysis of prostate cancer data and the mapping of patient journeys' through care. A framework and decision support system were developed to accommodate and analyse this information as well as to produce visual representations of the pathways. The next sections introduce further background on clinical pathways and data and process mining techniques.

4.1.1 Clinical Pathways

Clinical pathways, also known as care or critical pathways, have been introduced in healthcare systems to improve the efficiency of care whilst maintaining or improving its quality [200]. There are several definitions of clinical pathways in the literature but in this chapter they are defined as an ordered set of patient-centric events and information relevant to a particular clinical condition. Using the prostate cancer case study, particular attention is given to the use of clinical biomarkers and other indicators (such as blood readings) in pathways, as they enable a thorough inspection of data quality as well as further clinical studies observing trends over time. This is discussed in more detail later in this chapter.

In 1995, Pearson et al. [200] described critical pathways as a management plan

that displays goals for patients and provides the sequence and timing of actions necessary to achieve these goals with optimal efficiency. More recently they have been described as a concept for making patient centred care operational and for supporting the modelling of patient groups with different levels of predictability [200]. Clinical pathways are developed by multidisciplinary teams and rely on evidence from the literature, operational research and patient involvement methodologies [200].

Over the years, pathways evolved from paper-based to computerised pathways [201; 202] and there have been efforts to integrate them with electronic health records[202; 203]. One of the most promising fields for knowledge-based systems in health care is the support for guidelines and pathways [204]. The standard functions of pathways have been proposed in [202] and a strong emphasis is given to the statistics function to implement automated methods for checking the occurrence of variance (i.e. discrepancies between planned and observed events).

The analysis of clinical pathways is a topic receiving increasing attention from the field of medical informatics, but techniques often require extensive clinical expert knowledge and can be laborious. Huang and Duan [205] used process mining techniques to measure clinical behaviour derived from clinical workflow logs and to help identify novel process patterns. According to them, clinical pathway analysis has been defined as the process of discovering knowledge about clinical activities in patients care journeys. Ultimately the goal is to utilise the discovered knowledge for pathway (re)design, optimisation, decision support, audit or management. One of the major challenges reported was the derivation of compact yet high quality patterns that cover the most useful medical behaviours in clinical practice. Furthermore, the detection of complex patterns within patient data require higher levels of temporal abstraction [206] and the pathways introduced in this chapter take this into account.

Process mining techniques are promising analysis techniques in the context of

clinical pathways. However, it has been reported that traditional process mining algorithms have problems dealing with unstructured processes like those commonly found in a hospital environment [207; 208; 209] and that they may not produce clinically meaningful visualizations. The heterogeneity and incompleteness of the data are major obstacles in achieving meaningful models, yet an application to stroke has proved fruitful [209]. One of the aims of this chapter is to produce clinical pathways that may be suitable for process mining. For this, data quality is key, but consensus and definitions are lacking [37; 210] and intelligent agents that can explore quality issues are needed [210].

The use of routine data or workflow logs in the construction of clinical pathways is a key topic receiving increasing attention in the literature [205; 207; 209]. In hospitals, such efforts rely heavily on the hospital information systems (HIS) and electronic health records (EHR), and the availability and quality of the information conveyed in them. Indeed hospitals often opt for implementing several commercial departmental systems, creating "islands" of information across various departments [21]. This can significantly hinder the process of extraction and collation of detailed patient-centric information to create clinical pathways. The methods presented in this chapter attempt to overcome these difficulties.

A review on data quality in electronic health records [37] identified five data quality dimensions described in the literature: completeness, correctness, concordance, plausibility and currency. However, the authors identified that not all dimensions are commonly or consistently assessed and further work is needed towards the adoption of systematic, statistically based methods of data quality assessment. The work presented in this section encompasses the inspection of all of the above dimensions with a particular emphasis on assessing the completeness of pathway information using biomarker expert rules.

Overall, this chapter describes a framework for building and visualising prostate cancer pathways using routinely collected. This approach does not involve work-

flow logs produced by HIS or EHR, but rather, the patient-centric data conveyed in them. The previous work on methods for the collection of patient-centric data from multiple HIS has underpinned this research.

4.1.2 Mining Techniques

Process mining is the extraction of valuable process-related information from event logs; it is an area related to business process management (BPM), that is, the combination of computing and management sciences and their application to operational business processes [211; 212]. Workflow management (WFM) is also a related discipline that focuses on automation of business processes but it is often regarded as a more automated approach with less human interaction than BPM. The latter is more encompassing in that it ranges from process automation and analysis to management and the organisation of work [211]. Business processes are defined in [213] as sets of partially ordered and coordinated activities by which organisations accomplish their missions. Such processes can only be modelled and automated if they have the same structure and are repeatedly performed. This is the aim of WFM systems [212] and process mining aims at analysing event logs. WFM systems are often implemented in business contexts, and are common in the manufacturing of products or provision of services. However, as we have seen in the previous chapters, hospital information systems are not process-aware, do not incorporate WFM technology, and do not provide event logs. Some systems, however, will provide event logs as audit trails but their purpose is to aid with information governance rather than tracking any particular set of clinical processes. A way to overcome this limitation is to develop event logs based on live information from the systems, and this is possible by using the data model presented and used in the previous sections.

A Scopus bibliographic search for process mining in keywords revealed that, up

to the time of writing, only 1% of process mining publications (n=16) were in medicine and 2.3% were in biochemistry, genetics and molecular biology. Furthermore, the results are also positively skewed to the university that pioneered most of this research in the Netherlands [211]. Process mining has not been extensively applied to health care yet there is a growing interest in this field both from a service management and optimisation perspective and as well as in its application to clinical pathways. Challenges and limitations have been reported in all publications surveyed where process mining is applied to health care [207; 208; 209; 213; 214; 215], in particular, processes are often dynamic, *ad-hoc*, unstructured and multidisciplinary in health care domains. Furthermore hospital information systems are also not process aware which further hinders the discovery of useful process models.

In lieu of process mining techniques, some authors have reported using association rule mining to uncover relationships among data [215]. Association rules, also referred to as link or affinity analysis [36], were introduced in the early 1990s and aim at uncovering relations between items and item sets, typically in transactional databases. They are often used in the retail industry to identify items that are frequently purchased together [36]. Their application to biomedical domains has focused on biological data such as gene expressions [216] and other domains where data is more conveniently organised. The application of association rules in clinical data, however, is not trivial due to the nature of medical data, and hence greater efforts in data modelling are needed. Nevertheless association rules have been suggested for clinical domains in the late 1990s and early 2000s [217] and, more recently, have been applied to the diagnosis of diseases based on a list of findings [218] and in a paediatric primary care database [219]. However, authors concluded that the derived rules were often previously known or explained by confounding variables.

With the advent of *big data* and new modelling techniques, association rules are

becoming increasingly meaningful since datasets and the number of data elements have been growing. As a result, an increase in the number of publications since 2008 is noted when searching the Scopus bibliographic database for publications in medicine (a total of 142 documents retrieved between 2000 and 2013 where association rules appear in the articles' keywords). Association rules have been more recently applied to detect factors which contribute to heart disease in males and females [220], to identify clinical parameters akin to occurrence of brain tumor [221], and in hospital order-entry systems for management decision making [222]. Binary semantic association rules have also been introduced to cope with the increasingly complex environments [223].

Association rules and process mining algorithms are discussed further in the next section and in the context of the system-level pathways. State-of-the-art process mining and association rule mining techniques are applied and their utility in understanding frequent patient behaviours across hospital systems is examined.

4.1.3 Prostate Cancer

The latest estimates of global incidence indicate that prostate cancer has become the second most common cancer in men [224]. In the UK, it is the most common male cancer, accounting for 25% of all malignancies [42].

The National Institute for Clinical Excellence (NICE) in the UK publishes clinical guidelines and has recently developed the NICE pathways, a tool that visually represents the recommendations and guidelines on a specific clinical or health topic [225]. Following the NICE pathway, patients with suspected prostate cancer are directed through from referral, to assessment, diagnosis and communication; their needs are then often discussed at a multidisciplinary team meeting; admission and treatment options are selected as appropriate and ultimately patients are followed-up and outcomes assessed. During each step of the pathway, rele-

vant patient-centric data is produced and often stored in a variety of different hospital information systems. Clinicians wishing to investigate prostate cancer, say to establish the merits of alternative treatment and management options, would have a powerful tool if access to the integrated data was facilitated in an electronic and canonical form. However, as is often the case with hospital information systems, their data are heterogeneous, and data quality, accessibility and interface vary considerably. The work presented in this Chapter makes it possible to apply rules to the pathways, inspect their quality and guideline adherence. Furthermore, prostate cancer is an interesting yet non-trivial domain to build and analyse pathways as it is a chronic disease that often spans over large periods of time and consumes resources from several different hospital departments.

The next section presents the lowest level of granularity in the prostate cancer data, by examining how patients flow to and from hospital departments and their systems. Later, these are extended to pathways, where higher levels of granularity model more complex clinical information.

4.2 System-Level Paths

Chapter 2 presented a framework to collect patient-centric data from multiple hospital systems that emphasised on mapping the host environment. An interesting first exercise is to observe how patients interact with hospital departments. This can be achieved by listing frequencies and key statistics for each system, but more interestingly, by comparing and grouping patients' paths through the hospital systems they visited. This is a first high-level (low granularity) approach to modelling patient's journeys.

In this section, a system-level path is defined as a chronological sequence of *digital footprints* left in hospital information systems from patients diagnosed

with prostate cancer. This creates a single path per patient. Each footprint identifies a boolean (positive) occurrence of an event in a given system with an associated time stamp and patient identifier. The way in which the events were collected from the systems is restricted to prostate cancer patients and it implies some knowledge of the problem domain (previously discussed in Chapter 2). However, it is still possible that certain events are not specific to a prostate cancer journey. This needs to be taken into account in this chapter, and it is first addressed in the next section (4.2.1) when selecting an appropriate cohort of diagnosed prostate cancers.

A formal definition of system paths is given in section 4.2.2 and the understanding and analysis of the paths is described in section 4.2.3. The latter explores how association rule mining techniques can be used to mine frequent associations and how process mining techniques may yield more interesting results.

4.2.1 Data Selection

In this exercise, only patients with a diagnosis of prostate cancer should be considered from the available data in the operational data store (ODS). Given an inspection of the prostate cancer ODS, the hospital cancer register system (CRE) would be the most appropriate source to retrieve a list of diagnosed cancers, including those whose treatment is shared with other health care providers. However, this system only began to store accurate data from 2010. The data stored in the ODS spans a longer time period and the most relevant systems where prostate cancer patients are likely to be present were introduced in 2003 (histopathology, biochemistry, radiotherapy) or earlier (administration, oncology, radiology). In order to capture a complete count of prostate cancer patients that visited the hospital between 2003 and 2010, it was necessary to investigate three data sources: the patient administration system (PAS), the histopathology system (HIS) and

the oncology department system (ONC).

The PAS system contains clinical coding from discharge letters and hospital admissions other than outpatient events (at present these are not coded in NHS hospitals). The histopathology system contains results of biopsies and the confirmation of cancer, yet any patients who were clinically diagnosed may not have left a digital footprint in this system. Lastly, the oncology system will have diagnoses of patients who were referred to oncologists in this department instead of or, in addition to, urology. Any patient diagnosed with prostate cancer will have an instance with coding (or equivalent) from one or more of this systems.

The ODS contains a large number of patients, some with invalid local hospital numbers (either NHS numbers or other hospital's numbers). In this exercise, only records with valid hospital numbers were included from 2003 to 2010. A Venn diagram was produced from lists of all unique patients present in each of the three systems (Figure 4.1). The union of all three systems, $PAS \cup HIS \cup ONC$, yielded a list of 4,720 different hospital numbers and this was used to query the remaining systems.

Figure 4.1 shows additional interesting information that had previously not been looked at with this level of detail in this hospital. In particular, the majority of patients (38%, n=1,780) appear in the histopathology system alone. This would be largely due to biopsies that were carried out at this hospital yet no further treatment or consultations took place, or the patient decided to have treatment or monitoring elsewhere (including private hospitals). The second largest group in Figure 4.1 (29%, n=1,382) results from the intersection $PAS \cap HIS$. This is expected as patients who had a positive biopsy also have hospital episodes with a prostate cancer diagnosis. Only a small number of patients was present in all

three systems (6.3%, n=296). These are patients followed by the urology and oncology departments as well as having a biopsy. This interesting analysis, on its own, does not inform on data quality. It does show, however, that information on prostate cancers diagnoses is not confined to a single system. This posed difficulties in accurate data acquisition by the hospital's administration and for planning and performance. However, it was addressed when cancer waiting times were introduced and, in particular, from 2007, with the implementation of the first version of the hospital cancer registry system (CRE).

A first recommendation to collect prostate cancer diagnoses by hospital staff was to use the PAS system alone, however, Figure 4.1 shows a gain of 2,710 additional diagnoses by including HIS and ONC sources. Based on the union of the data in all three systems, a set of 4,720 different hospital numbers was used to query the other hospital systems. Table 4.1 summarises the list of selected systems, a short description of information conveyed in them, the number of unique valid hospital numbers in the ODS, and the total number of records in each source. The set of 4,720 diagnosed patients was linked to each of the sources, resulting in a table

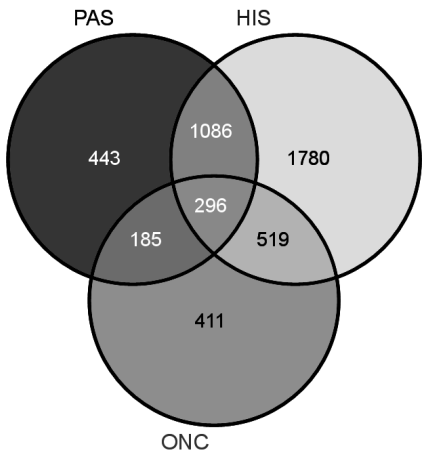


Figure 4.1: Venn diagram showing unique hospital numbers across Histopathology, Administration and Oncology systems.

with 52,475 records (average of 11 records per patient). Further details are given in the next sections.

System	Information	Patients	Records
PAS	Hospital episodes where Prostate Cancer is primary diagnosis.	2,010	7,773
LAB	PSA tests.	48,603	138,968
RAD	Radiotherapy to prostate (both palliative or radical).	2,012	2,768
ONC	Oncology department admissions (both new and old systems).	1,411	1,429
OPT	Operating procedures to prostate.	4,120	4,636
IMG	Radiology imaging to prostate.	866	1,281
HIS	Histology report with a Gleason grade.	3,681	4,363
ORT	Orthopaedics long bone fractures (only from 2008).	1,651	1,783

Table 4.1: List of selected systems and the number of records and valid patients in them.

4.2.2 Defining a System Path

Let D represent the predefined set of database systems such that $D = \{d_1, \ldots, d_n\}$ is a set of n systems as, for example, described in Table 4.2. A patient's footprint in a given system, d, is given by a three-tuple $F = (r, t, d)$ where r is the patient identifier, and t is the time in days from the first interaction associated with the hospital. A chronological sequence of systems' footprints (a system or system-level path) for patient, r, is represented here as $P = \langle F_1, \ldots, F_m \rangle$ where

i. F_i is of the form (r, t_i, d_i) for $1 \leq i \leq m$,

ii. $t_1 = 0$ is the time of the first interaction recorded,

iii. $t_i \leq t_{i+1}$ for $1 \leq i \leq m - 1$,

iv. when $t_i = t_{i+1}$ then $d_i \neq d_{i+1}$ for $1 \leq i \leq m - 1$, so that all F_i are unique with respect to (t_i, d_i) values (both time and system visited).

When $t_i = t_{i+1}$ and $d_i \neq d_{i+1}$ for $1 \leq i \leq m - 1$, however, the corresponding footprints F_i and F_{i+1} are said to be concurrent.

Hence, a simple system path might be $P = \langle F_1 = (1, 0, d_2), F_2 = (1, 1, d_7), F_3 = (1, 111, d_5), F_4 = (1, 176, d_2) \rangle$. In this example, a patient ($r = 1$) visited three different hospital systems. The first encounter was with the LAB system for blood tests, followed by a biopsy a day later (HIS), surgery at 111 days (OPT) and his last encounter with a hospital system was for another blood test at 176 days. The total length of time for this path is the latter number of days.

A sequence of the visited systems, S, for this example is can be $\langle d_2, d_7, d_5, d_2 \rangle$, or using the systems' codes for simplicity, $S = \langle LAB, HIS, OPT, LAB \rangle$.

Based on the above definition of a system path, the cohort of 4,720 patients identified in the previous section was used to create a data-driven system path for

each patient (one path per patient). Table 4.2 gives the predefined list of systems with their respective number of unique patient identifiers and total number of records.

D	System Code	Patients	Records
d_1	PAS	2,010	4,010
d_2	LAB	4,037	33,862
d_3	RAD	1,533	1,976
d_4	ONC	1,411	1,426
d_5	OPT	1,479	1,729
d_6	IMG	670	814
d_7	HIS	3,681	4,359
d_8	ORT	31	32

Table 4.2: Predefined list of hospital systems for system-level paths.

4.2.3 Understanding Paths

A first table, containing the ordered set of 4,720 patients (48,208 records) and their system-level paths was produced. However, this table included paths where only one system was visited. Although these are valid footprints and paths, they were not considered in subsequent analyses as no interactions between systems occurred. As such, any paths with a cardinality $|P| = 1$ were removed (n=283, 19 from PAS, 65 from ONC, and 199 from HIS). In addition to this, any duplicate records (records with the same patient identifier, system and day), as defined above, were also removed. For comparison purposes, only paths with sequential footprints of the same system on different days are considered. However, sequential footprints of different systems on the same day are accepted, although these only occur in 2% (n=1,096) of records in the transactional table. This is because some systems may have more than one record of the same appointment in the same day (often due to administration processes), while others will only register at most one episode a day. This leaves a total of 4,437 patient paths (47,925

records) each with two or more footprints.

System	Path Start	Overall Frequency
LAB	74.51% (n=3306)	90.98% (n=4037)
HIS	10.03% (n=445)	78.48% (n=3402)
ONC	6.15% (n=273)	30.34% (n=1346)
PAS	5.25% (n=233)	44.87% (n=1991)
OPT	3.31% (n=147)	33.33% (n=1479)
RAD	0.43% (n=19)	34.55% (n=1533)
IMG	0.25% (n=11)	15.10% (n=670)
ORT	0.07% (n=3)	0.70% (n=31)

Table 4.3: Distribution of hospital systems as a starting footprint of a path and overall frequency across paths.

Systems Visited	1	2	3	4	5	6	7
N Paths	20	1471	1055	1240	445	185	21
% Total Paths	0.45%	33.15%	23.78%	27.95%	10.03%	4.17%	0.47%

Table 4.4: Distribution of Number of Systems visited in all paths.

A first exercise was to list the distribution of starting footprint of paths (Table 4.3). This is important because it allows an understanding of how patients start their prostate cancer paths and it may also indicate potential erroneous data. The most common start is via the LAB system (74%) followed by the systems where diagnosis was made (HIS, ONC, and PAS total 22%) and 4% start from other systems. A LAB start would indicate previous PSA testing was done before diagnosis and a high volume of patients was expected. However, some of the least frequent starts are also plausible. Patients' cancers could be detected through imaging and even after an orthopaedics operation (pathological fractures or bony metastases). The radiotherapy system, however, is less likely to be a true start of a cancer pathway. Indeed upon manual inspection of paths starting with radiotherapy (n=19) it was identified that they were either closely followed by a diagnosis or the events were at the start of the cohort time (2003). This is taken into account later when analysing the paths.

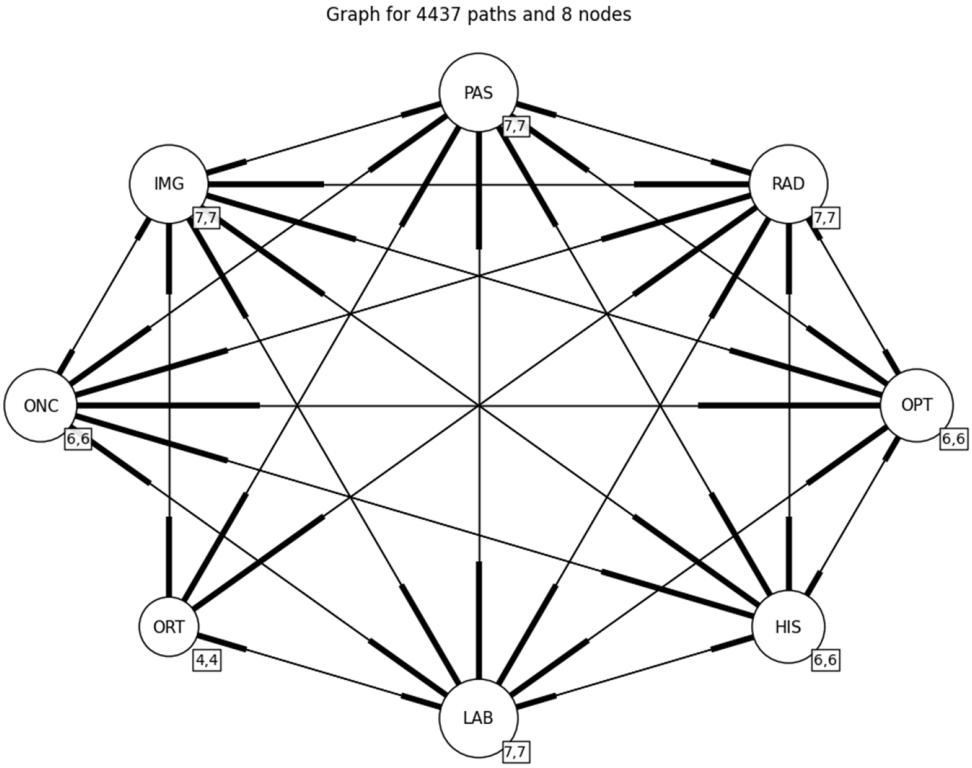
Another inspection, of the distribution of the number of different systems in each path (depicted in Figure 4.4), revealed that 95% (n=4211) of all paths visited between 2 to 5 different systems, and none visited all 8 systems. This warranted further investigations to understand how the systems interact with paths, detailed later in this section.

A first approach to understanding basic system path behaviour was to list all possible pairs of systems appearing across all paths (8^2 possible permutations given the list of systems). Software was developed to compute the frequencies of all possible pairs of footprints as well as a graphical representation using a directed graph (built with Python and networkx library). The latter is also drawn based on all paths and provides a visual representation of the existing pairs of systems. The frequency of the starting pairs across paths was also computed. Figure 4.2 shows the missing, most and least frequent sequence pairs between systems given the cohort of patient's paths.

This frequency analysis indicates the most common sequence pairs of systems visited and clues regarding the overall patients' journey through hospital departments and the findings are summarised:

- The most frequent sequence across paths is $\langle LAB, LAB \rangle$ (n=3508), indicating PSA retesting. It is expected that this could be the most significant activity for patients with prostate cancer and one that may add a substantial burden to hospital services also due to its frequency within paths (average 7 visits per patient).

- The second and third most frequent sequences are between the LAB system and the HIS system. The HIS system contains a histological confirmation of diagnosis and it is often preceded or succeeded by a PSA test (here shown as a visit to the LAB system). When preceded, a raised PSA value would trigger a biopsy result in HIS. When succeeded, the biopsy result in



Most Frequent	
Sequence	**Paths**
LAB, LAB	3508
LAB, HIS	2103
HIS, LAB	2031
RAD, LAB	1032
LAB, PAS	983
PAS, LAB	809
PAS, HIS	692
LAB, RAD	680
OPT, HIS	632
OPT, PAS	625

Least Frequent	
Sequence	**Paths**
IMG, ORT	4
ONC, ONC	4
RAD, OPT	3
OPT, RAD	3
RAD, ORT	3
PAS, ORT	2
ORT, RAD	2
OPT, IMG	1
ORT, PAS	1
ORT, IMG	1

Missing	
Sequence	**Frequency**
HIS, ORT	0
ORT, HIS	0
ONC, ORT	0
OPT, ORT	0
Frequent Start	
Sequence	**Paths**
LAB, LAB	1743
LAB, HIS	1111
LAB, PAS	201
HIS, LAB	153

Figure 4.2: Frequency of pairs of systems for the 4,437 patient paths drawn. The graphical representation shows the overall cohort and the existing links between each system base on the sequence of visits; it includes thicker stubs to represent direction (arrow heads) and each node shows its in-degree and out-degree respectively. Each table shows the most or least frequent sequence pairs.

HIS would trigger further monitoring of the PSA either because of active monitoring or a pre-treatment check (such as PSA check before surgery).

- The above sequences also appear in the most frequent starts of paths: $\langle LAB, LAB\rangle$ (n=1743), $\langle LAB, HIS\rangle$ (n=1111), $\langle HIS, LAB\rangle$ (n=153). This is important as it may indicate how patients were first diagnosed and whether they were "screening". However, biopsies may be undertaken as a result of a visit to PAS where they are undertaken as an outpatient procedure.

- $\langle LAB, PAS\rangle$ and $\langle PAS, LAB\rangle$ are similar to the interactions of LAB with HIS in that a hospital consultant appointment and/or outpatient biopsies succeeds or precedes a PSA test. Here, however, the appointment is not a biopsy. The sequence $\langle LAB, PAS\rangle$ (n=201) is also one of the most frequent starts of paths and more so than $\langle PAS, LAB\rangle$ (n=93).

- Sequence $\langle RAD, LAB\rangle$ (n=1032) revealed that a PSA test is often carried out after radiotherapy in order to monitor its effect.

- It was interesting to observe, as depicted in the graph, that almost all systems mutually followed each other. The only exceptions found were in the ORT system. This was expected due to the overall small number of ORT events in the dataset and served as a validation for the methods.

The preliminary analyses were carried out based on the system paths data model and the developed software. A full summary table with the results of this analysis is given in Appendix C 1.4. Two additional approaches to explore and understand how patients flow through systems are given in the next sections. A first approach relies on mining frequent itemsets from within the paths and a second approach focuses on mining similar sequential paths.

4.2.3.1 Itemset Mining and Associations

Given the transactional database of system paths and the frequency analysis previously carried out, an interesting exercise is to mine frequent sets of items appearing in each path. Mining frequent itemsets (groups of items) is a first step to mine association rules; it identifies frequent combinations of items that can later be used to create association rules with a given support and confidence [36; 226]. Traditional itemset mining or association rule mining does not take into account time or order, and as such, the work explored in this section will not have these elements and it focuses on answering questions regarding which items (systems) appear together in paths. Furthermore, association rule algorithms do not mine loops of length one but this was covered in the previous section.

The most prominent algorithm for mining itemsets and association rules was introduced in 1994 by Agrawal and Srikant [226]. The algorithm's input is a transactional database and a predefined minimum support value (threshold), often referred to as *minsup*. The output is a list of frequent itemsets. A frequent itemset is an itemset that appears in *minsup* transactions.

If A and B are two itemsets, an association rule $A \Rightarrow B$ is a relationship between the antecedent, A, and consequent, B, such that $A \cap B = \emptyset$. The support of a rule is the number of transactions that contain $A \cup B$ and the confidence of a rule is $(A \cup B)/A$ and can be interpreted as an estimate of probability $P(B|A)$. This traditional measure of confidence, however, ignores the support of the itemset in the rule consequent and hence some rules may be misleading[36]. An alternative measure that takes the support of B is defined as $lift(A \Rightarrow B) = \frac{P(A,B)}{P(A)P(B)}$. However, this introduces the problem of symmetry whereby there is no difference between rules $A \Rightarrow B$ and $B \Rightarrow A$. Other metrics of interest have been researched yet in this chapter a manual inspection of the generated rules was sufficient.

The Apriori algorithm (here with an extended step for creating association rules)

has as inputs a database of transactions, *minsup* and confidence, and the following generalised working steps:

1. From the database of transactions, mine frequent itemset patterns where each candidate itemset satisfies *minsup* and store the results in a working set. This process is repeated as many times as the cardinality of the largest itemset in the database.

2. The Apriori property is applied after each iteration of the previous step. The Apriori property (or anti-monotone Apriori heuristic [226]) states that any subset of a frequent itemset must also be frequent and, removing less frequent items is often called pruning.

3. Once all valid frequent itemsets have been generated, association rules are created in the manner described above based on the given confidence.

The Apriori algorithm has known performance limitations mostly due to repetitive database scans. This is particularly true when searching for large sets of candidates and mining long patterns. Over the years more efficient implementations of the Apriori algorithm have been introduced, for example, storing candidate itemsets in a hash-tree format [227]. In 2004, however, an algorithm that offered overall better efficiency, the Frequent Pattern Growth (FP-Growth) algorithm, was introduced [228] and it is now widely used with the same results and improved performance. For the work needed to be carried out in this thesis, however, performance was not an issue due to the size of the dataset and the processing speed of the computers used, and the Apriori algorithm was considered sufficient, in particular because of the small number of items (systems).

Given the transactional database of 4,437 system paths, an itemset of the visited systems was produced for each path. In this exercise, it is only possible to observe combinations of itemsets and time and order of the items are not taken into

account.

The Aprori algorithm to mine frequent itemsets was first applied to the system paths dataset with a $minsup = 0$ in order to retrieve all possible combinations (both frequent and infrequent), and to validate its implementation. Later, a threshold $minsup = 2.23\%$, was applied to retrieve itemsets appearing in just over 100 paths. Upon investigation of the first results, this threshold was found to retrieve a large enough number of paths for consideration. Setting a low *minsup* is known to cause *combinatorial explosion* (too many rules and many of them are meaningless) leading to the *rare item problem* [229] of deciding which rules are of value and which are not. Several mechanisms have been proposed to tackle this problem, such as using multiple minimum supports [230] or frequent closed itemsets. The latter were introduced by Pasquier *et al.* [231] with the aim of reducing, without information loss, the size of the mined association rule set, resulting in a condensed representation of the rules [232]. A frequent closed itemset is a frequent itemset that it not included in another itemset (superset) with the same support [231]. The process of generating frequent closed association rules was used to produce rules regarding the system paths with low *minsup* and confidence in order to retrieve a larger number or rules for inspection.

First, the Apriori algorithm ran in $\approx$ 15 miliseconds and returned 16 frequent closed itemsets (results in Table 4.5). Then, association rules were mined with the same *minsup* and a minimum confidence of 10% in order to obtain a larger number of rules for inspection and only frequent closed association rules were kept. The results of mining association rules are given in Table 4.6.

Table 4.6 shows the two rules with most confidence for six consequent systems. The strongest association rules (over 90% confidence) have as consequents LAB or HIS and this was expected as these systems have the most overall support. As a result a large number of rules are found with LAB as consequent (see Appendix C 1.4). The first rule in Table 4.6, $\{PAS, IMG, HIS\} \Rightarrow LAB$, indicates with

Itemset	Support
1. $\{LAB\}$	33862
2. $\{HIS\}$	4160
3. $\{PAS\}$	3991
4. $\{LAB, HIS\}$	2748
5. $\{RAD\}$	1976
6. $\{OPT\}$	1729
7. $\{ONC\}$	1361
8. $\{IMG\}$	814
9. $\{OPT, HIS\}$	762
10. $\{LAB, OPT\}$	613
11. $\{LAB, OPT, HIS\}$	554
12. $\{LAB, ONC\}$	360
13. $\{ONC, HIS\}$	147
14. $\{LAB, IMG\}$	143
15. $\{PAS, HIS\}$	115
16. $\{LAB, ONC, HIS\}$	103

Table 4.5: Results of the Apriori Algorithm with $minsup = 2.3\%$.

Rule	Support	Confidence
1. $\{PAS, IMG, HIS\} \Rightarrow LAB$	203	99.51%
2. $\{OPT, IMG\} \Rightarrow LAB$	166	99.40%
3. $\{OPT, IMG\} \Rightarrow HIS$	160	95.81%
4. $\{LAB, OPT, IMG\} \Rightarrow HIS$	159	95.78%
5. $\{RAD, ONC, OPT, HIS\} \Rightarrow PAS$	134	89.93%
6. $\{RAD, ONC, OPT\} \Rightarrow PAS$	149	89.76%
7. $\{PAS, HIS\} \Rightarrow OPT$	1162	84.08%
8. $\{PAS, LAB, HIS\} \Rightarrow OPT$	1073	83.11%
9. $\{LAB, ONC, IMG\} \Rightarrow RAD$	171	73.71%
10. $\{ONC, IMG\} \Rightarrow RAD$	175	73.22%
11. $\{RAD\} \Rightarrow ONC$	897	58.51%
12. $\{LAB, RAD\} \Rightarrow ONC$	742	58.20%

Table 4.6: Selected Closed Association Rules based on the Apriori Algorithm to mine frequent closed itemsets with $minsup = 2.3\%$ and 10% minimum confidence.

99.5% confidence (support 203) that when a patient visits the administration, imaging and histology systems, he will also visit the laboratory system for a PSA

test. The interest here is that both rules 1 and 2 show IMG as an antecedent and the latter is not a common system across all paths. Nevertheless, the LAB source is present in over 90% of paths and as a result it appears in a large number of rules including those with a relatively small support.

66% (n=135) of rules where RAD appears as a consequent (n=203) have ONC as antecedent. This, together with the rules where RAD appears both as consequent or antecedent, indicates a link between patients who see the radiotherapy department and those who see oncology. This is indeed expected as the course of radiotherapy treatment is often given by oncology consultants.

From the association rules it was also possible to identify a strong relationship between operating theatre (OPT) and histology system (HIS). Indeed a rule not shown in Table 4.6 but visible in Appendix C 1.4 shows that paths that visit HIS also visit OPT with 94.25% confidence and strong support (n=1394). The relevance of this rule is that for patients that had an operation (either extensive biopsy or prostatectomy) a histological report was produced.

One of the drawbacks of using a predefined *minsup* is that any itemsets that fall short of this value will not be included. For this reason, the orthopaedics system did not appear in any of the results. However, upon manual inspection, all paths where the ORT system was visited (n=31) had the following other system visits: 100% LAB, 87% HIS, 51% PAS, 39% IMG, 35% OPT, 32% RAD, and 23% ONC. This is interesting because out of all patients who visited orthopaedics only 39% had an imaging event where the word prostate was mentioned. This reassures that *minsup* values should be set to a low value. This does, in turn, require a greater effort to inspect the larger number of resulting rules and identify the meaningful ones.

The Apriori results and association rules provided some interesting, although to a degree, expected, results and allowed an understanding of which combinations

of systems are most frequently visited together.

Despite the fact that it is important to understand which systems are visited together in system paths, these rules do not inform on the timeliness or order of the visits, which is important if one is to understand the flow of information across systems. The next section investigates how sequential pattern mining may address this limitation.

4.2.3.2 Sequential Pattern Mining

The problem of mining sequential patterns was introduced in 1995 by Agrawal an Srikant [233] and later improved by the same authors [234]. As opposed to mining association rules, in sequential pattern mining transactions are grouped to form ordered sequences. A sequence string was exemplified earlier as $S = \langle LAB, HIS, OPT, LAB \rangle$ where all elements are ordered chronologically. The preliminary investigations (Figure 4.2) already covered ordered sequences but only for pairs of systems so as to gain a basic understanding of their interactions. The problem of mining sequential patterns is to discover "all sequential patterns with a user-specified minimum support, where the support of a pattern is the number of data-sequences that contain the pattern" [234]. This encompasses any sequences and not just pairs.

An early algorithm is the Generalised Sequential Patterns (GSP) algorithm proposed by Srikant and Agrawal [234] and it is based on Apriori. The algorithm works by making multiple passes over the data to determine candidates based on the item's support ($minsup$). Then, for each sequence of a determined length, the database is scanned to obtain support and generate candidate sequences using Apriori. The process is repeated until no frequent sequence or new candidate are found. As with Apriori, multiple database scanning make the process less efficient, particularly for long sequences, but pruning remains an advantage. Over

the years algorithms with increased performance were developed [235] and later began being used for sequence analysis of biological data such as DNA, RNA or proteins [236]. The first approaches were Apriori-based until pattern growth algorithms were introduced in 2000 [237]. Pattern growth algorithms differ in that they are based on recursively projecting segments of the sequences, resulting in a more efficient processing. Introduced by Pei *et al.*, PrefixSpan [237] is one of the most widely known algorithms and it is based on "projecting suffixes" for all given prefixes. This divide-and-conquer approach can be summarised in three steps: generating candidate sequences (sequence length - 1), divide search space according to prefixes, and scan the database to find subsets of sequential patterns. The algorithm is described in detail in [237] and it is not repeated here.

An open-source Sequential Pattern Mining Framework (SPF) that implements most sequential pattern mining algorithms [238] was used to run the PrefixSpan with the transactional database of 4,437 system paths and $minsup = 10\%$. As with the previous exercise, only frequent closed patterns were mined.

The most common sequence found was $\langle LAB, LAB, LAB \rangle$ with a support of 3390 paths. Indeed, when looking at most frequent sequences, the majority of those with support over 1000 had on average 1.6 (SD 0.5) unique systems and LAB was present in all. Again, the fact that LAB is so common both within and throughout paths has introduced additional difficulties in evaluating the results. As such, the rules are not included here but they are available in Appendix C 1.4. An additional challenge when applying sequential pattern mining algorithms is that the results included several similar rules with the same systems and identical support where only the order of the systems is different. The use of standard sequential pattern mining algorithms alone was not sufficient to produce meaningful patterns with discriminative support. However, other attempts were made using more recent algorithms that produce sequential rules. Introduced in 2012, the CMRules algorithm [239] focuses on mining sequential rules that are common

to many sequences rather than rules that appear frequently in sequences. The algorithm finds association rules to "prune the search space for items that occur jointly in many sequences" and then it removes those rules that do not meet the confidence and support thresholds according to the time ordering [239]. The output of this algorithm are rules of the form $A \Rightarrow B$ where A is an itemset with items that can appear in any order and B is an itemset of items that sequentially follow A. The rules reveal a sequential relationship between the two sets of items.

An implementation of CMRules algorithm was also available in the SPF framework and the latter was used to produce the rules with a minimum confidence of 10% and, because the previous exercises already yielded most association rules, a larger minimum support of 10% was selected. The algorithm returned 67 rules (included in detail in Appendix C 1.4) and selected rules are shown in Table 4.7.

The first three rules in Table 4.7 indicate, with a large support and confidence, the systems that precede LAB. They are RAD, HIS and PAS, respectively. These are similar results to those presented earlier but are here given with a degree of confidence. The two last rules (6 and 7) were selected to demonstrate the difficulty in interpreting the results. In these two rules OPT is kept in the antecedent set, HIS is kept in the consequent set, and LAB and PAS are swapped. However, there is more confidence that LAB and HIS are preceded by PAS and OPT.

Sequential Rule	**Support**	**Confidence**
1. $\{RAD\} \Rightarrow \{LAB\}$	959	94.13%
2. $\{HIS\} \Rightarrow \{LAB\}$	2598	91.83%
3. $\{PAS\} \Rightarrow \{LAB\}$	1464	91.83%
4. $\{PAS, LAB, OPT\} \Rightarrow \{HIS\}$	688	91.77%
5. $\{OPT, HIS\} \Rightarrow \{PAS\}$	515	83.91%
6. $\{PAS, OPT\} \Rightarrow \{LAB, HIS\}$	688	83.54%
7. $\{LAB, OPT\} \Rightarrow \{PAS, HIS\}$	475	76.66%

Table 4.7: Selected Sequential Rules from the CMRules Mining Algorithm for System Paths.

Sequential pattern mining provided rules that indicate which systems are followed by others but, given the work carried out in association rule mining and initial work on sequential pairs there was little new information given by applying ordering constraints. However, introducing interval time constraints to sequential pattern mining might provide more interesting results and some knowledge regarding time windows between systems. This is discussed in the next section together with other techniques to inspect the timeliness of systems' visits.

4.2.3.3 Time Constraints

A first approach to understanding how systems are visited in time can be to draw a scatter plot based on the information in patients' paths. A software program was developed to scan all system paths and produce a scatter plot (Figure 4.3) where the x-axis represents time, the y-axis shows the list of systems, and each footprint is associated to a system and represented by a tick in the plot. All paths were grouped together and the scatter plot shows the overall times when each system was visited in the cohort. In order to distinguish which systems were visited before and after diagnosis the paths' time were zeroed at diagnosis date. Figure 4.3 shows the scatter plot produced for all system paths. The maximum and minimum range was set to 3000 days in order to cover all paths.

Figure 4.3 is expressive in showing which systems are more frequently used before and after treatment. The most used system before treatment is LAB where blood tests are carried out in symptomatic patients but also in those who have no symptoms but have one or more risk factors. The second most common system used before diagnosis is the operating theatre OPT. This system would appear before the diagnosis because extensive biopsies will be carried out in the operating theatre. Other systems that appear before diagnosis are IMG where imaging was done to the pelvic region of suspected cancers and the radiotherapy RAD system.

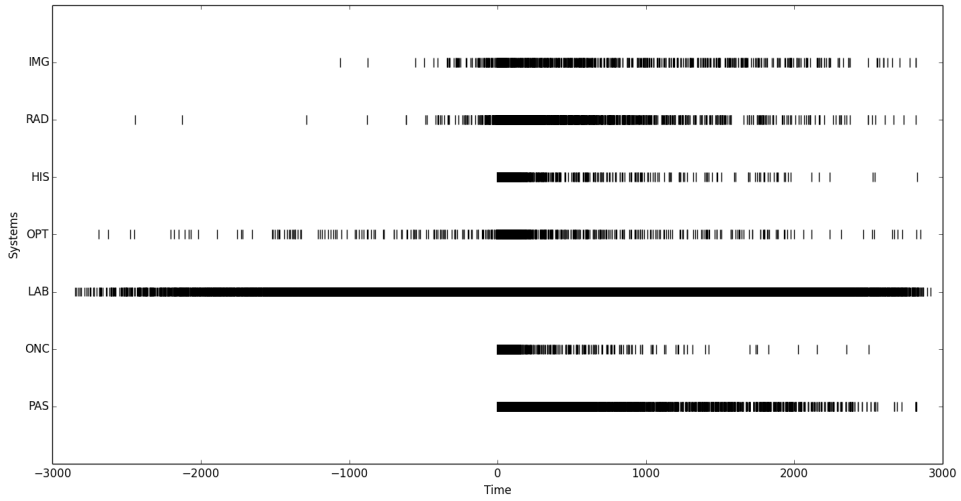


Figure 4.3: Scatter plot showing the systems visited and their times for all given system paths. The x-axis represents time zeroed at diagnosis, the y-axis shows the list of systems and the ticks show when a particular system was visited in time (range 3000 days).

Upon inspection it was found that the latter are mostly palliative radiotherapy treatments to the sacral area and hips or pelvic region.

It is also interesting to note the rate at which systems become less used after diagnosis. In particular, the oncology ONC system and the histopathology HIS system become less used more quickly than the others. This is because of the nature in which patients are registered in the ONC system and, in the case of HIS, once a conclusive biopsy has been carried out, and in particular if the treatment was surgery, further histopathology reports are less likely to be needed. Again, the LAB system is the most visited up to the end of all paths. Of interest are also the gaps in time where no path ever visited a particular system. These are visible in more detail in the second plot (Figure 4.4) where the time window has a smaller range of 100 days after diagnosis. The operating theatre system sees some interesting intervals. In particular, less operations (OPT) were carried out between 10 and 30 days. It was, however, not possible to accurately explain why

such gaps exist in this time frame.

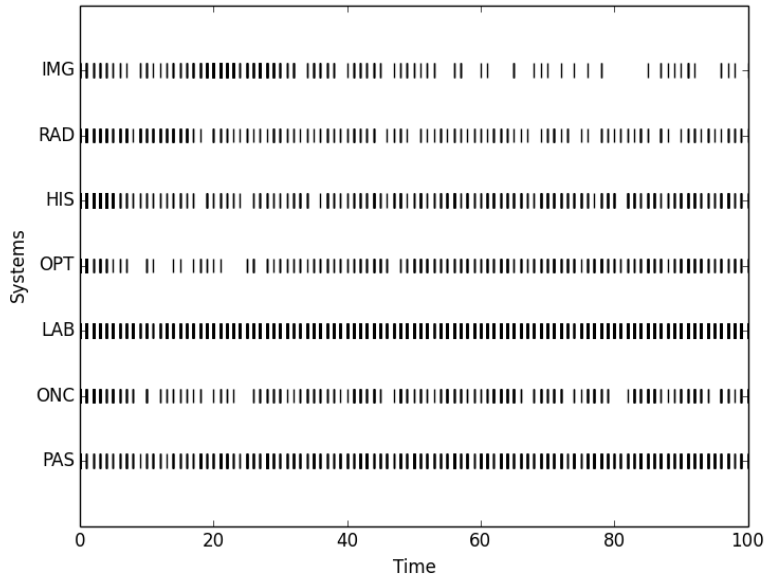


Figure 4.4: Scatter plot showing the systems visited and their times for all given system paths with a set time window of 100 days after diagnosis.

Another approach to study the timeliness between systems visits with more detail is to use Sequential pattern mining with constraints. In 2006, Hirate and Yamana introduced a novel algorithm to mine sequential patterns with time intervals [240]. Their algorithm discovers time-extended sequential patterns that are common to several sequences. The algorithm's input is a time extended sequence database where each itemset is annotated with a integer value denoting time. Also as inputs are the time constraints defined in [240]: minimum support, minimum time interval (between two subsequent itemsets of a sequential pattern), maximum time interval, and overall maximum and minimum time interval for the overall sequential pattern. The algorithm works in the same fashion as PrefixSpan in that it generates candidate sequences and then applies time constraints that satisfy minimum support and time intervals. A full description of the algorithm is given in [240] and it is not repeated here. The SPM framework

includes an implementation of Hirate and Yamana's algorithm to generate closed sequential patterns with time constraints. The algorithm was applied to the system paths dataset with the following constraints: minimum support of 2.3% to capture a wide number of rules; both time intervals minimum were set to 1 day and maximum set to 3000 days (maximum range) in order to capture as much information as possible. This resulted in 710 rules, 12 unique pairs of systems and 6 unique rules where more than two systems are present as itemsets.

Sequential Pattern	**Support**
1. $\{LAB, PAS\} < 1 > HIS$	421
2. $LAB < 1 > \{OPT, HIS\}$	253
3. $\{LAB, PAS\} < 1 > OPT$	196
4. $PAS < 1 > \{OPT, HIS\}$	191
5. $\{LAB, PAS\} < 1 > \{OPT, HIS\}$	137
6. $\{LAB, OPT\} < 1 > HIS$	131
7. $\{LAB, PAS\} < 3 > HIS$	118

Table 4.8: Selected sequential patterns with time constraints based on Hirate and Yamana's algorithm. Only rules where more than two systems are present in itemsets are shown. The angle brackets show the relative time stamp of the subsequent system, e.g. Rule 2 indicates that at time 0 LAB system was visited and 1 day later OPT and HIS were visited.

Table 4.8 shows rules that include more than two systems. The remaining rules, not shown in table 4.8, were less interesting in that they only revealed sequential pairs with different times and support. The rule with most support indicates that at time 0 systems LAB and PAS were visited and that one day later the HIS system was visited. The first and last rules (1 and 7) are the same but were considered separately by the algorithm due to the time interval. Overall, the rules produced were difficult to interpret as the order of the systems in the itemset is not accounted for and also because of the sheer volume of similar rules with different time cut-off points. Different input parameters were attempted but yielded a smaller number of similar or uninteresting rules. The algorithm was therefore not found to be suitable for this particular dataset. However, in

other situations, it might help to unearth particular combinations of systems that are consistently followed by others within a given time interval.

As a result, the average times between pairs of systems are explored in more detail. Software was developed to compute the averages and standard deviations for all possible permutations of systems. The full list of results is given in Appendix C 1.4 and two graphs were drawn to illustrate pairs with the shortest ($\leq$ 60 days) and longest times ($>$ 60 days) having a support of at least 100 paths (Figure 4.5).

The standard deviations for the computed averages in Figure 4.5 indicated much dispersion in the data (descriptive statistics in Appendix C 1.4). Attempts were made to reduce dispersion by imposing cut-offs for extreme values at 1 year and later, 100 days, yet only marginal improvements were observed. A better way to cope with extreme values was, in this case, to use percentiles. As such the 90th percentile was computed for all connections. Every arch in Figure 4.5 indicates the computed average and, within brackets, the percentile. The 90th percentile indicates the threshold (in days) where at least 90% of the data lies.

Overall, the shortest average time between systems occurs from PAS to HIS (average 2 days, SD 2) and its 90th percentile is 3 days. The latter indicates that over 90% of the connections from PAS to HIS are within 3 days. This connection reveals that a histopathology report is produced shortly after a consultant episode or a procedure in PAS. Indeed, the same percentile and a small average is observed in another connection, from the operating theatre (OPT) to PAS. This reassures us that the shortest observed times occur when a histological confirmation is needed given a biopsy, here represented by a visit to the operating theatre (OPT, for an extensive biopsy) or the outpatient clinic (PAS, for a less extensive biopsy). A small number of PAS visits in this connection, may, however, not be to perform a biopsy and that would account for the remaining 10% of connections.

In addition, Figure 4.5 A shows an interesting relation between the systems where

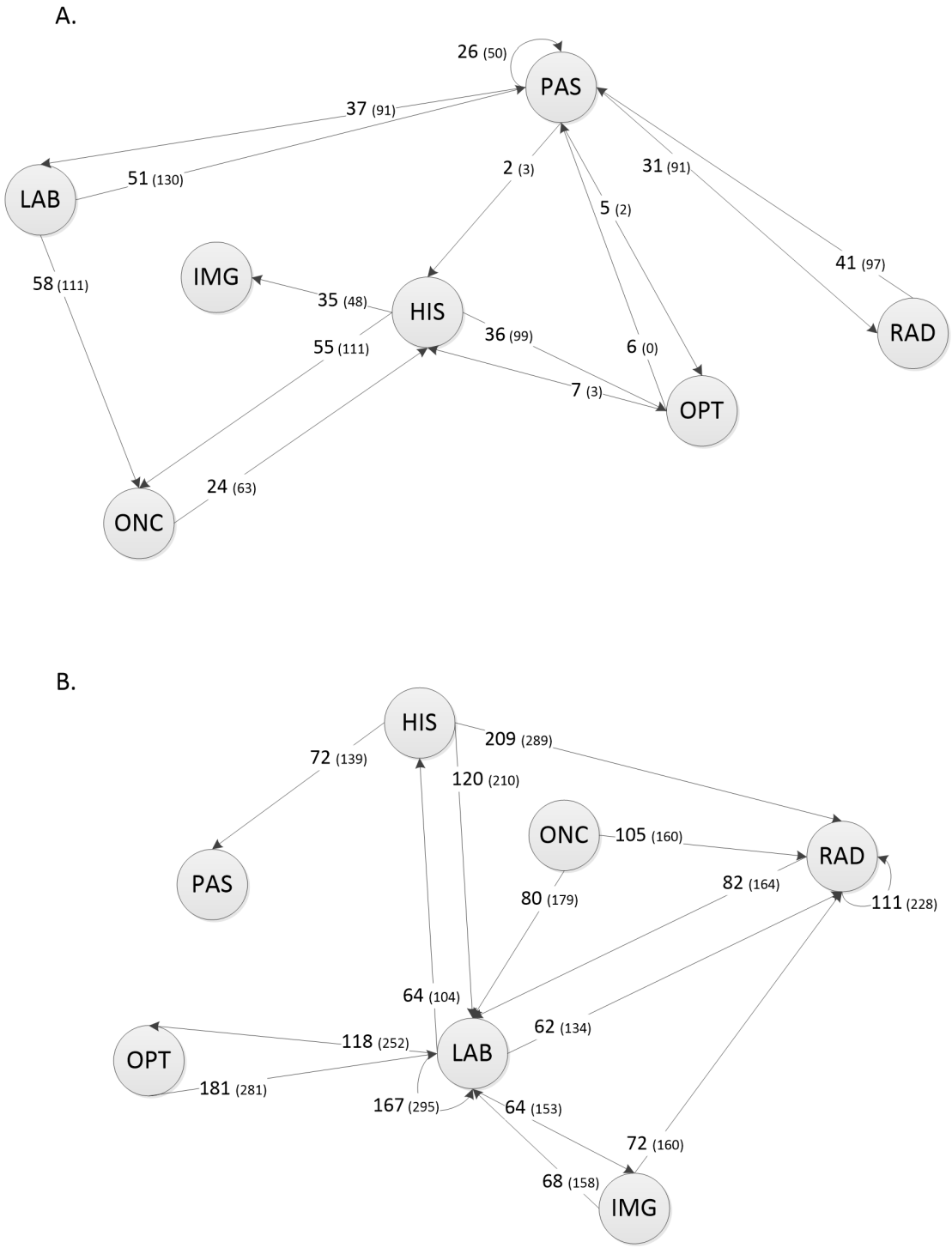


Figure 4.5: Two graphs (A and B) showing the average times and the 90th percentile between systems. Graph A shows the shortest times (less than 60 days) and the graph B shows the longest times (over 60 days). Only connections with a support over 100 were included. The graphs' arches show the average time in days between systems and the 90th percentile.

shortest connections are observed. These are from PAS to OPT or HIS and from OPT to HIS or PAS. The latter (OPT to PAS) occurs on the same day more than 90% of the time indicating that these patients are taken to a hospital bed after visiting the operation theatre. This analysis could also provide an insight into the systems' interoperability in that the administration system (PAS) should record most of the procedures, diagnoses and their respective clinical codes. However, the available data and limited interoperability identified earlier in this thesis together with the chronic nature of prostate cancer lead to mostly large time intervals between systems. Exceptions are the interactions between systems when patients are admitted as inpatients to undergo surgical interventions.

Figure 4.5 B shows the longest times between systems and the longest time observed (with a support of 100 paths or more) is between HIS and RAD with an average time of 209 (SD 173) and a 90th percentile of 289 days. This indicates the amount of time patients take to begin radiotherapy (here including both radical and palliative) after a histopathology confirmation of diagnosis. However, in Figure 4.5 it is only possible to assess the times between each two systems individually and hence it cannot be regarded as an accurate measurement of the average times taken to begin radiotherapy treatment after a diagnosis is made. Furthermore, radiotherapy may be delivered as adjuvant or neoajduvant therapy combined with other treatment types. These are limitations to the interpretation of these results yet they may be overcome by modelling and analysing the data differently. This is discussed in more detail in the next sections.

Overall, extending the system paths to include time is an important and worthwhile exercise that provided insights into the timeliness of the sequential flow of patients between systems. In particular, it highlighted the time intervals where patients were admitted to a hospital bed. Further work is needed to account for times across several systems such that, for example, the average time between PAS and OPT can also be computed in situations where other systems' visits

exist between these two. Hirate and Yamana's algorithm, however, did not discover any interesting rules that covered more than two systems and, as a result, it was considered sufficient to compute the averages and percentiles between any two systems. The results from Hirate and Yamana's algorithm provided similar results with this dataset.

Overall, algorithms that introduce time constraints such as Hirate and Yamana's can provide additional information that allow a better understanding of the flow between systems. In addition, scatter plots produced by the developed software also provided interesting insights into the time distribution of systems across all paths. The next section explores the utility of process mining to discover interesting processes within the system paths.

4.2.3.4 Process Mining

This section discusses the use and utility of process mining in finding additional patterns in the system paths. As previously mentioned, most process mining algorithms do not provide insightful results in health care domains due to noise, complexity and heterogeneity of the data. Classical approaches such as the alpha algorithm [241] require complete logs with no noise, yet this is not possible given the nature of health care data, as seen in this thesis. Nevertheless, other algorithms have been proposed to deal with noisy data [212].

The Heuristics Miner [242] is an advanced process discovery algorithm, and it has been previously used in a health care domains [208; 215]. Limitations were identified in [215] and the models produced were *spaghetti*-like, too complex due to the sheer size of connections and nodes. A later study [208] informed that the produced models were able to focus on the most frequent paths but they were too complex to analyse due to the fact that, in health care, they do not have a single kind of flow. The authors also reported that breaking down large event logs

into smaller logs with similar properties (by using clustering techniques such as self organising maps) provided better results, easier to analyse. Other additional clustering techniques that facilitate the understanding of *spaghetti* models were later introduced [243].

Nevertheless, because the system paths dataset has a small number of unique tasks (yet a large number of possible interactions), an attempt to produce a process model was carried out using the Heuristics Miner algorithm. Full details of the application of the algorithm and the generated process models are available in Appendix C. The system paths dataset was converted into the appropriate workflow log format and the Process Mining framework (ProM) software was used as it implements several of the process mining algorithms including the Heuristics Miner [244]. In process mining terminology, the dataset including all system paths is called the event log, a system path is a process instance (or case), and a patient footprint in a given system is a event (or activity) in a given system (task) although the latter two are sometimes used interchangeably. In the system paths dataset there are eight possible tasks (systems).

The Heuristics Miner algorithm was introduced in 2006 by Weijters [242], and uses a heuristic approach to address the problems faced by the classical alpha algorithm and, in particular, to deal with noise in event logs. The algorithm is explained in detail in [211; 215; 242] yet its three key steps are summarised here:

1. **Construct Dependency Graph**
   In this step a dependency measurement is computed for each pair of tasks in the event log. This is a measurement of strength of an ordered relation between two tasks given the support of relation.

2. **Determining Dependency Types**
   Establish the input-output expressions between activities so that parallel or exclusive activities can be captured.

3. **Discover Long Distance Dependency Relations**
   In this step, indirect relations between tasks are discovered (non-local behaviour).

An important concept is the computation of the dependency measure in the first step of the algorithm [211]. Let A and B be two tasks and $|A \rightarrow B|$ the number of times A is directly followed by B. The dependency relation measure between A and B is

$$|A \Rightarrow B| = \begin{cases} \frac{|A \rightarrow B| - |B \rightarrow A|}{|A \rightarrow B| + |B \rightarrow A| + 1} & \text{if } A \neq B \\ \frac{|A \rightarrow A|}{|A \rightarrow A| + 1} & \text{if } A = B \end{cases}$$

A dependency value is calculated for every pair of tasks in the dataset and gives the relative strength of A to B in relation to its inverse, B to A. It's values range between 1 and -1 and values close to 1 indicate a stronger positive dependency between A and B. The latter can only be achieved if A is often followed by B but B is hardly ever followed by A [211]. In the HeuristicsMiner algorithm a dependency matrix and subsequent dependency graph are constructed based on this measure. The determination of dependency type and long distance dependency relations are not explained in detail here but are given in [211; 215; 242].

The system paths dataset was divided into a training set (n=2212) and a test set (n=2225) and the HeuristicsMiner algorithm was applied to both independently. In both occasions the models produced had a poor fitness (i.e. how the observed process complies with the control flow specified by the process model). The adjustment of the algorithm's default parameters resulted in a marginal improvement yet when artificial start and ends were added to all paths, the best fitting and least *spaghetti*-like models were produced. Artificial starts and ends can be seen as a way to normalise the data so that all paths' starts now converge into an overall start task and end points converge into a unique end task. De-

tails of the models produced as well as input parameters are given in Appendix C. When comparing the training set with the test set, differences regarding the placement of the tasks and their connections were observed. As such, only tasks and connections that were common to both sets could be taken into account and the overall models were not used.

The following two interesting relations between tasks were identified in both the training set and the test set:

- $ArtificialStart \Rightarrow ONC \Rightarrow RAD \Rightarrow ArtificialEnd$

  Sequence pair ONC, RAD had been previously identified in the simple sequential pairs (support 485) and as the fifth most common start sequence pair (support 138). Rule $ONC \Rightarrow RAD$ was also identified using the CM-Rules sequential pattern algorithm with a support of 668 and confidence of 53%. Furthermore, an interpretation of the Apriori association rule results indicated that 66% (n=135) of rules where RAD appears as a consequent (n=203) have ONC as antecedent.

  In all three cases, the rule did not stand out as particularly relevant. Using the HeuristicsMiner, however, the dependency measurement value for relation $|ONC \Rightarrow RAD|$ was 0.96 both for the training set and the test set. This indicates a strong relation $|ONC \Rightarrow RAD|$, when compared to $|RAD \Rightarrow ONC|$. This would confirm a link between patients who visit the radiotherapy department followed by the oncology department, as their treatment plan is often agreed in oncology.

- $ArtificialStart \Rightarrow HIS \Rightarrow IMG \Rightarrow ArtificialEnd$

  A connection between HIS and IMG alone had not been previously reported in any rule generated by the associations and sequential rule mining algorithms. It was, however, identified in the sequential pairs summary,

where HIS, IMG with a support of 207 and IMG, HIS with a support of 19.

The dependency value given by the HeuristicsMiner algorithm was 0.93 for the training set and 0.90 for the test set. This rule indicates that imaging is often carried out after histology and the reason for this might be an elevated Gleason grade or extracapsular extension of the tumour. In this case imaging would help to investigate the extension of the invasion and existence of metastases elsewhere.

Given the above results, an alternative approach to capturing interesting connections using the dependency measure was undertaken. The approach relies on building a dependency map rather than a process model and to apply a lower threshold filter. This approach was carried out in the following steps:

1. Compute dependency values and create a dependency matrix for the training set and another for the test set.
2. Eliminate any dependency values below a predetermined threshold from the matrices.
3. Compute the difference between the values obtained with the training and test sets and consider removing any values from the matrices below a second, cross-validation threshold.
4. Draw a dependency map based on the dependency values that meet the above criteria from the test matrix.

This approach uses the first step of the HeuristicsMiner algorithm and adds additional validation to ensure that the resulting dependency map is consistent with both the training and test set (given the cross-validation threshold). The dependency threshold used was 0.5 and, in this case, a cross-validation threshold filter

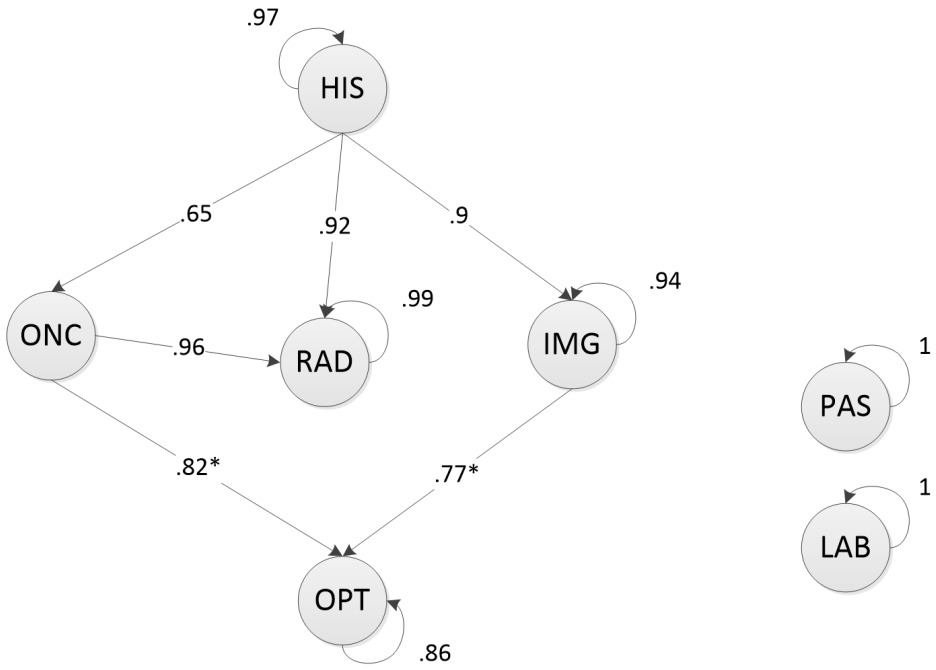
was not used as almost all values were within 0.05 and two larger values, 0.26 and 0.15 were highlighted. The resulting dependency matrix and map are given in Figure 4.6.

This approach bypasses the second and third steps of the HeuristicsMiner algorithm and as such it does not give long distance relations or dependency types, however, it is simpler to understand and the dependency matrix can be considered the backbone of process models [211].

Using this measure, two of the most dominant systems in the previous analyses, LAB and PAS, became disconnected (dangling) activities with a self-loop dependency of 1. They are disconnected because, with the exception of self-loops, the dependency values were consistently below the threshold, 0.5. In particular, dependency $|PAS \Rightarrow HIS|$ had the highest dependency value of 0.41, and all other connections had values lower than 0.3. According to the dependency relation, pairs where A is followed by B as many times as B is followed by A have a dependency value of 0. Values close to 0 will therefore indicate that it is difficult to assert the true direction of the relation.

This analysis highlighted interesting results and among them were the rules previously identified in the process models ($|ONC \Rightarrow RAD|$ and $|HIS \Rightarrow IMG|$). Apart from the latter and self-loops, the relations with greater dependency values are $|HIS \Rightarrow RAD|$ (0.92) and $|ONC \Rightarrow OPT|$ (0.82). Figure 4.6 shows that the system with most relations is HIS. The relations indicate that after a patient is diagnosed histologically, he is likely to visit either oncology (0.65), radiotherapy (0.92), or radiology for imaging (0.9). A further link between oncology and radiotherapy (0.96) is also observed as it is common for oncology to plan the radiotherapy treatment.

It is also interesting to note the routes through to the operating theatre (OPT) in the dependency graph. Patients may flow to from oncology (0.82) or radi-



	HIS	IMG	LAB	ONC	OPT	ORT	PAS	RAD
HIS	0.97 (0.02)	0.9 (0.03)		0.65 (-0.04)				0.92 (0.02)
IMG		0.94 (0.03)			0.77 (0.15)			
LAB			1 (0)					
ONC					0.82 (-0.26)			0.96 (0)
OPT					0.86 (0.1)			
ORT								
PAS							1 (0)	
RAD								0.99 (0.01)

Figure 4.6: Dependency map (top) showing all connections in the test set with a dependency value over 0.5. The two values highlighted with a star have higher cross-validation thresholds (0.15 or -0.26). Dependency matrix (bottom) showing all dependency values present in the map and the difference between the value in the training set and the test set.

ology (0.77). These contrast with the previous average times analysis in that the relations do not feature in the latter due to their small cardinality (under 100 paths). Furthermore, the previous analysis revealed a strong short time dependency between OPT, PAS and HIS, and that no similar association rules to those in the dependency graph were produced. This reinforces that, although the routes to OPT in the dependency graph may be interesting, they do not include a minimum support, and that the heuristics miner algorithm does not distinguish low frequent behaviour from noise. In this exercise, minimum supports were not included in the dependency graph so that more relations could be discovered.

Process mining techniques may provide some interesting results but they are not, on their own, considered appropriate in heterogeneous domains and respective datasets. Further work on combining these techniques with association rules, time constraints and robust data models is needed to ensure that meaningful processes can be derived from large datasets.

4.2.4 Results and Key Findings

This section introduced system level paths and explored currently available data and process mining tools and techniques for the discovery of trends and associations. The system paths data structure allowed the production of data-driven and patient-centric paths where the complete set of system visits for a patient is arranged sequentially. The data structure has also allowed for the system paths data to be easily analysed using both developed software and exiting mining software toolkits.

Overall, and partially due to low granularity and nature of the data, most of the meaningful results revealed association between two systems. In addition, an average of 7 visits to the LAB system per path introduced additional combinatorial explosion and, as a result, most rules with highest confidence have this

system as consequent. This was aggravated by multiple bi-directional relations where almost all systems flow to another and back. Nevertheless, the Apriori algorithm was able to answer questions such as which systems were often visited together. Rule $\{PAS, LAB, HIS\} \Rightarrow OPT$, for example, indicates with 83% confidence (1073 support) that the operational theatre is visited when all other three systems are visited (the administration, the laboratory for a PSA, and the histology systems). This, however, is not helpful in determining whether the latter systems were a consequence of a visit to the operating theatre or the contrary. The sequence in which systems are visited matters and sequential pattern mining algorithms were explored.

The sequential pattern mining algorithm used in this section narrowed down the total number of rules produced yet most of these would only include two systems. Despite this, the algorithm was able to provide more meaningful rules with greater support. Table 4.9 shows a list of the sequential and association rules with most confidence where only two systems appear. The sequential pattern mining algorithm provided, as expected, the same rules with consistently less confidence than the association rules.

Antecedent & Consequent	**Association Rule**	**Sequential Rule**	**Avg. Time (%ile)**
OPT , HIS	94.25%	88.62%	7 (3)
PAS , LAB	93.87%	91.83%	37 (91)
HIS , LAB	92.91%	91.83%	120 (210)
OPT , LAB	92.56%	90.75%	181 (281)
ONC , LAB	86.63%	84.98%	80 (179)
OPT , PAS	84.04%	83.28%	6 (0)
RAD , LAB	83.17%	94.13%	82 (164)

Table 4.9: Rules with most confidence from both association and sequential rule mining algorithms and their respective average time and 90th percentile in days.

A subsequent analysis using sequential mining with time constraints provided little new information and smaller support because of time cut-offs. As a result, an

inspection of the times between systems was undertaken and provided interesting results. In particular, a relation between systems with the shortest paths (PAS, OPT, HIS) was observed, indicating a time-dependent relation between the operation theatre, histology, and the administration system. From all relations where it was possible to compute averages, 75% (n=43) had an average time over 30 days and 26% had an average time over 3 months.

The chronic nature of prostate cancer contributes to larger time windows between events and consequently large average times and standard deviations. The use of percentiles was found to provide additional meaningful information on the time intervals between systems. However, a full map with novel information on the shortest and longest times between systems was produced as well as a scatter diagram. Both provided a new way of looking into this data that had not previously been looked at. Further work is needed to explore how this information can be made useful for hospital management, clinical staff and patients. The results presented in this section provide a general picture of the times taken for patients to flow between systems and hospital departments and a basis for further work, some of which is undertaken later in this chapter.

Process mining algorithms were also investigated and found unsuitable to model unstructured processes such as those found in health care, even with small number of items (systems) as observed in this section. *Spaghetti*-like models were produced, however, the underlying dependency measurement was considered of interest as it highlighted the direction of the relationships between systems.

Table 4.10 shows a list of all the relations included in the dependency model on the system paths dataset. When available, the respective association and sequential rules' confidence are given together with the average time (and 90th percentile) in days. This provides a basis for comparison between the results of the dependency measurement and the association rules; it shows that the relations with higher dependency value have similar rules with low confidence.

Antecedent & Consequent	Association Rule	Sequential Rule	Dependency Value	Avg. Time (%ile)
ONC , RAD	66.64%	53.05%	0.96	105 (160)
HIS , RAD	29.63%	26.31%	0.92	209 (289)
IMG , OPT	24.92%	-	0.77	68 (102)
HIS , ONC	23.40%	22.89%	0.65	55 (111)
ONC , OPT	20.50%	-	0.82	20 (55)
HIS , IMG	14.30%	-	0.9	35 (48)

Table 4.10: Relations included in the Process Mining Dependency Matrix and their respective association and sequential rules' confidence, and the average time and 90th percentiles in days.

The striking difference between confidence levels and the dependency values is illustrated in table 4.10. The selection of rules with high support and confidence based on frequent itemsets provides different results to those where a unidirectional tendency is used. The latter tackles the problem of symmetry and provides interesting information that is not covered in association rules but future work on a hybrid approach could provide better rules that cover both frequency and direction between two and more itemsets.

The data structure proposed in this section allowed the aggregation of information into system paths for analysis. The work carried out in this section reassured that association rule mining and process mining algorithms, on their own, do not provide significant results in heterogeneous and unstructured environments such as the one found in health care. Different ways of modelling the data can provide more meaningful results with the above methods. Process mining models, even when data modelling techniques impose some degree of organisation, do not cope well with increasing granularity in unstructured environments. Furthermore the association or sequential mining algorithms techniques are not aimed at discovering end-to-end processes and other techniques such as hidden Markov models [211] should be explored.

Further research is needed on methods tailored to unstructured environments and

hybrid approaches could provide meaningful results. Association rules with semantic constraints and based on a data structure such as the one proposed in this section are most promising. The work carried out in the next sections provides a framework for the modelling of complex clinical information that extends the system paths and introduces a novel approach to the visualisation and analysis of health care data from multiple heterogeneous sources.

4.3 Pathways

The previous section described the modelling of routine hospital data into system paths and assessed the utility of data and process mining techniques. Low granularity and scant semantics were reported limitations. Additional information, contained in each of the systems, can provide a more detailed understanding of the patients' journey through care and their clinical features. This section explores the development and modelling of pathways with complex clinical information with a focus on data quality and on the prostate cancer biomarker, the PSA.

The aims of the work presented in this section can be divided into three:

- to extend the system paths data model and generate data-driven patient-centric pathways for prostate cancer from routinely collected data,
- to develop a framework and decision support software for the integration, visualisation and analysis of pathways,
- to evaluate the completeness and utility of the generated pathways for investigating biomarker trends.

The pathways should allow for the selection of high quality data for clinical studies and decision making, which, in turn, enables the (re)design, management and

optimisation of clinical pathways. We focus on a definition of a pathway as a data structure that synthesises knowledge and facilitates the development of methods for the computation of variance and other statistics. The framework presented in this section, together with their formalisms, addresses the issues reported in the literature and should allow and encourage other tools and techniques, such as process mining or *ad hoc* algorithms to be used.

The next section formally defines a pathway and it is followed by a section that describes the methodology used to model the data and to development of a data dictionary (section 4.3.3).

4.3.1 Defining a Pathway

Let D represent the pathway dictionary where the i-th entry has a code c_i ($1 \leq i \leq n$) in a total of n possible codes described in detail in Table 2 and in section . C_E is the subset of codes containing timed events and C_I the subset containing informational elements, such as demographics. By associating a zero time with informational elements, all events in the pathway can be viewed as timed events.

A *pathway activity* A is then defined as four-tuple $A = (r, t, c, v)$ where

- r is the patient identifier
- $c \in C$ is an event code
- t is the time in days before or since the day of diagnosis recorded for patient r
- v is a value, numerical or categorical, associated with dictionary code c

A pathway for patient, r, is represented as a chronological sequence of activities, $P = \langle A_1, A_2, \ldots, A_m \rangle$, where

i. A_i is of the form (r, t_i, c_i, v_i) for $1 \leq i \leq m$,

ii. $t_i \leq t_{i+1}$ for $1 \leq i \leq m-1$,

iii. any A_i with $c \in C_i$ has $t_i = 0$,

iv. if $A_i = (r, t_i, c_i, v_i)$ and $A_{i+1} = (r, t_{i+1}, c_{i+1}, v_{i+1})$ then there is no activity $A = (r, t, c, v)$ where $t_i < t < t_{i+1}$ and

v. all relevant activities involving patient r appear in P.

A simple pathway for patient $r = 1$ might be $P = \langle A_1 = (1, -28, P, 45), A_2 = (1, 0, D, 2), A_3 = (1, 1, G, \text{"4 + 3"}), A_4 = (1, 1, H, \text{"}CyproteroneAcetate\text{"}), A_5 = (1, 151, R, \text{"37"}), A_6 = (1, 260, P, 0.2), A_7 = (1, 340, P, 0.05), A_8 = (1, 539, P, 0.05)\rangle$.

In this patient's pathway the first PSA test was elevated at 45 ng/ml and this led to the diagnosis of stage 2 prostate cancer with a Gleason grade of 4+3. Note that the biopsy was performed as an outpatient event and hence it is unavailable in this pathway, however, the histopathological findings of that biopsy are present. The patient then agreed to undergo hormone therapy (cyproterone acetate) and a subsequent 37 sessions of radiotherapy. The number of radiotherapy sessions is recorded as value of element code R. This was followed by PSA readings of 0.2 ng/ml and two readings < 0.1 ng/ml which indicate a good response to treatment.

This pathway is summarised in Table 4.11 and a description of the pathway codes and dictionary is given in section 4.3.3.

A pathway may also be succinctly expressed as a sequence of its activities codes such that, for patient r, $S = \langle c_1, c_2, \ldots, c_m \rangle$ is the set of c sequence codes belonging to a pathway P with m activities.

Using the above example, for patient $r = 1$, the corresponding pathway sequence would be $S = \langle P, D, G, H, R, P, P, P \rangle$. However, sequences may be ambiguous

Activity, A	**Identifier,** r	**Time,** t	**Event Code,** c	**Value,** v
A_1	1	-28	P	45
A_2	1	0	D	2
A_3	1	1	G	4+3
A_4	1	1	H	Cyproterone Acetate
A_5	1	151	R	37
A_6	1	260	P	0.2
A_7	1	340	P	0.05
A_8	1	539	P	0.05

Table 4.11: Tabular summary of a patient's pathway with 8 activities and a total elapsed time of 567 days.

when activities are concurrent in time. A way to overcome this limitation is by programmatically sorting every set of concurrent elements in S. A sequence may also be truncated so that repeated elements are removed. The latter enables a rough aggregation of pathways with similar sequential activities.

The above model of expressing pathway activities is similar to the entity-attribute-value (EAV) data model [245] where concepts are described in an attribute in a row. Later, the i2b2 data model [178] expanded on the EAV model to account for time (start and end dates for each observation). This, together with a star schema, has been described as an extremely efficient way of querying data as a large index can be built to encompass all patients' data in the master table [178].

The above definitions rely on a pathway dictionary that describes the semantics of patient activities and this is described in detail in the next section.

4.3.2 The Operational Data Store

The data extraction process described in chapter 2 and published in [52] was used to collect patient-centric data from HIS and to create an operational data store (ODS). The ODS has been previously discussed in the context of creating system

paths but it is used more extensively in the development of the pathways and framework described in this section.

As previously noted, the ODS replicates the hospital environment and it may contain more data than required for particular clinical studies as the retrieval process is based on minimum use of constraints. However, this provides a holistic representation of the patients, including their demographics, comorbidities, test results and other information, and is limited by the availability of electronic information in the HISs.

The prostate cancer ODS contains information from the following sources: administration, cancer waiting times, histopathology, radiology, biochemistry, operating theatre, orthopaedics, oncology, radiotherapy, and an external source, the cancer registry. However, not all sources were used in the development of pathways as later explained in 4.3.3. The data sources in the prostate cancer ODS were previously identified in chapter 2 (Table 2.2) and are summarised later in this section.

The methodological process used to select appropriate data elements from the ODS and to create a pathways data dictionary is given in section 4.3.3. First, a formal definition of a pathways is given in section 4.3.1.

4.3.3 Building the Pathways Dictionary and Repository

The ODS contains data from the retrieved hospital sources and metadata which allows for the inspection, linkage and integration of semantic and syntactically different data. Nevertheless, the ODS may contain information outside the domain of a specific pathway. Therefore, in order to build a pathway dictionary, it is crucial to identify, select and retrieve key data elements. The process of building a pathway dictionary from the data in the ODS is illustrated in Figure 4.7 and

is inspired by the similar data warehousing technique of Extract-Transform-Load [6]. The pathway dictionary can be regarded as a simple ontological knowledge base, built by a bottom-up process, from available data to concepts. Temporal ontologies have been developed [246] yet for the definition of pathways, the above time-oriented data structure together with a pathway dictionary was considered sufficient for this case study.

The dictionary building process, based on input from domain experts, a survey of the literature and current prostate cancer guidelines, involves gathering relevant data elements and applying transformations to either create new features or strip out irrelevant elements (e.g. hospital events that are neither exclusive nor relevant to the treatment of prostate cancer). At the end of this process, and for each data element, a flat file with the data corresponding to that element is created in the four-tuple transactional format described in 4.3.1. The steps involved in this process are described in detail below.

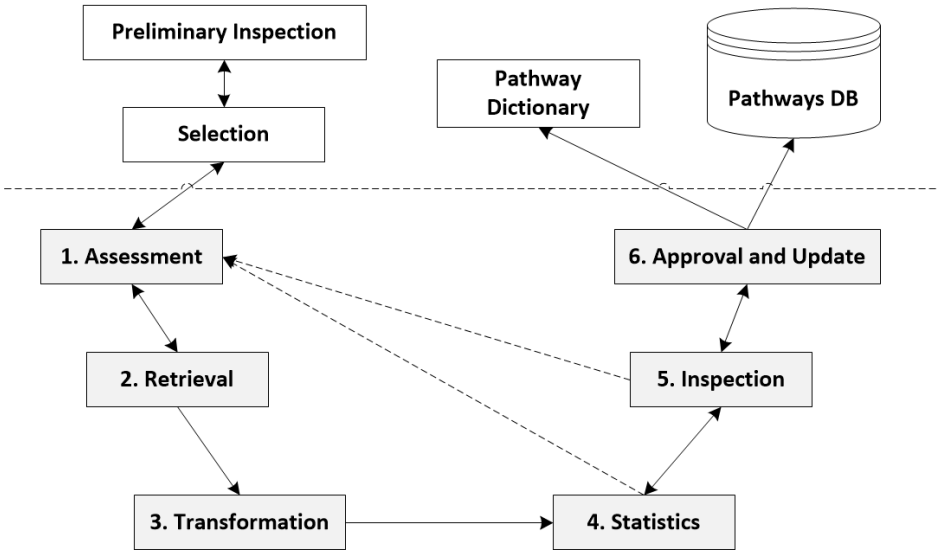


Figure 4.7: Methodology to build pathways dictionary and database.

1. **Preliminary Inspection and Selection**

   The domain experts collaborate on a first inspection of the available data in the ODS to help with the identification of key data elements to be included in the pathways. This involves examining summary statistics (such as frequencies of biochemistry tests) and metadata (such as attributes descriptions, semantics or expected outliers) from the ODS and is important as it sets the granularity of the pathways and the extent to which they can be meaningful for a particular disease.

   For the prostate cancer study, three classes of information were defined: demographics, diagnostics (including investigations) and treatment. Hence the selected elements in this step have an associated class. Further to this, each element type can either be a timed event, describing a particular activity at a given time, or auxiliary information such as demographic data or other non-event data such as a patients participation in a clinical trial. Both class and type are two properties common to all elements of the pathway and can be determined *a priori* or throughout the process of building the pathways as explained below. The use of routinely collected hospital data for timed events indicates with certainty that a particular activity occurred; however, its absence may not indicate the opposite. Existing data may be used in validity checks for the completeness of the data, for example, as we will see later the PSA biomarker can act as an alert for potential missing activities at particular time intervals.

   For each selected data element, the following six steps are carried out to create a complete pathway dataset and dictionary. Throughout the following steps we will use the example of the biomarker test for prostate cancer (PSA test) as a data element.

2. **Assessment**

   The first step is to inspect the elements values as well as its semantics,

syntax and data type, and any potential limitations that may interfere with the consistency of the data element. Additional mapping, linkage and transformations may be necessary to enforce a consistent format and these should be identified here. An example arising from the PSA test was the need for the removal of values that include symbols, such as " < 1", meaning the PSA test value is less than one. In this case such values were replaced by 0.5. A first classification of the element is also given by assigning a dictionary code and the element type (informational or timed event); in this case the code for the timed event PSA test is P.

3. **Retrieval**

   The set of attributes and values for the data element are retrieved from the ODS. In the case of the PSA test, the attributes in the ODS include dates of test authorisation, date of entry, value, comments, clinical history, fasting, blood reading thresholds and the patient identifiers. Rule-based deterministic record linkage can be used to enforce constraints. In the course of the case study, this step was used to select data within the study time period as well as validating data from the hospital sources against the cancer register datasets, where possible, in terms of completeness, correctness and concordance. The retrieved attributes must have the information required by pathway definition. The data for the particular element is then stored and in this case study a comma separated file is created to this effect. For PSA tests the attributes selected from the ODS to be included in the pathway were date of entry (date when the sample was taken from the patient within the selected time period), the value, and the identifier that allows linkage.

4. **Transformation**

   The retrieved data file is converted into the pathway data structure, with attributes Identifier, Code, Date (instead of time), and Value, where Code

is a constant. Date is used here but it will later be converted into time, t, zeroed at diagnosis date. The latter, by removing full dates, allows an additional layer of anonymity to the pathways as well as a basis for comparison among patients. Any necessary transformations identified in the previous steps are undertaken here.

5. **Summary Statistics**

   Summary statistics are produced in this step. These include distributions of the Value attribute, which can help to detect potential bias, together with overall support (i.e. total number of patients), value-specific support (i.e. number of patients on each value category), and extremes. Such statistics may help to detect and correct quality issues, by assessing completeness (missing data), correctness and plausibility. Additionally, other statistics may be produced, such as the number of values within a range; this is particularly useful for producing a summary of abnormal blood readings, such as raised PSA tests.

6. **Inspection**

   Together with the domain experts, the retrieved data and descriptive statistics are inspected. The values of the attributes are also checked for format consistency and the quality dimensions described above. At this stage, a decision regarding the data element is reached.

   The element may be:

   (a) kept as is, should it contain sufficient information and adequate support;

   (b) rejected, because there may not be enough information or support;

   (c) subject to decomposition, into two or more elements, should the values of the element vary qualitatively creating a source of ambiguous information, or should the requirements of a particular study involve

inspecting a particular quantitative range such as the abnormal range of a blood reading.

In the example of the PSA test, the data element was kept after the values were set to a canonical form. An example of an element that was rejected in the case study is biopsy because of insufficient support (this is further discussed below). A further example of an element that was split was surgery, into orchidectomy and surgery (prostatectomy). Another example of an element that was split was radiotherapy, where, for the analysis of the trend of PSA, only radical radiotherapy was interesting to investigate as it affects the PSA.

7. **Approval and Update**

   Upon inspection a decision is made regarding the data element and its values. When the decision is favourable an update is carried out. The update is concerned with the technical work of merging the table containing the data element and its values with the pathways database master table. Further transformations are also carried out to sort the master table by date and patient identifier, and to compute time t zeroed at diagnosis date. This can be achieved by either creating an informational element providing the date of diagnosis or by programmatically isolating the specific date from an existing element and subsequently setting t for all activities in a pathway. The pathways dictionary is then updated with summary information.

The first version the pathway dictionary and how the cohort was selected is shown in the next section.

4.3.4 Selected Data and Core Dictionary

A list of diagnosed prostate cancer patients was previously selected for the system-level paths. This was sufficient for the analysing the paths that patients take through systems, however, this data was not consistently checked for quality by the hospital until after the CRE system was introduced. Because it is now pertinent to look at detailed clinical information, a smaller list of patients, with verified demographics and cancer waiting times information was selected from the ODS. This list was then sent to the local cancer registry (Eastern Anglia Cancer Registry) so as to obtain accurate dates of death and additional information, such as treatments performed elsewhere, when available. Patients who were not in the registry, who reside in another registry catchment area, or whose tumours were awaiting a final report or were referenced only, were excluded from the list.

The selected cohort dates were patients diagnosed between 2004 and 2010. This time period allowed, overall, one year of data on PSA "screening" (2003) as well as one year follow-up of data (2011). The date when patients were last checked to be alive or dead was the 1st of April 2012 by the cancer registry. A total of 2,979 patients were identified in the CRE system, including those patients where treatment is shared with other hospitals. Deterministic record linkage was performed and for 149 patients it did not provide accurate matches (these were removed). A total of 2,830 patients were included in the original dataset. This data, however, is still part of a large cohort that warranted additional data quality inspection.

Upon transformation of the data into pathways, a data quality exercise aimed at matching radical treatment data between the hospital and the cancer registry. A total of 32% (n=926) of records were labelled as mismatching on treatments. Overall the treatment types with most mismatching cases were Hormone Therapy (19%) and Radiotherapy (13%). This is explained as the cancer registry collects

information that extends beyond the NNUH and it also captures data from GP practices, hence, providing a more comprehensive information on patient's treatments. The full implementation of the hospital cancer registry system (CRE) made possible for this information to be accurately matched between the hospital and the East Anglia Cancer Registry, in particular from 2010. As a result, a cohort of 1,904 patients was selected for further study and it is used henceforth in this chapter.

The data dictionary was created based on the selected data sources from the ODS, listed in Table 4.12. A total of 16 core elements was selected and it is shown in Table 4.13. The elements frequency indicates the percentage of pathways in which that particular element is present. Additional elements were added later and this is explained in the next sections.

Data source	**Description of selected data**
Administration (PAS)	Patient episodic information, co-morbidities and clinical coding.
Histopathology (HIS)	Histopathology reports and extracted Gleason grades.
Radiology (RAD)	Radiological imaging limited by reports where the word prostate occurs.
Biochemistry (LAB)	Prostatic Specific Antigen (PSA) tests. However, other blood tests can be added.
Operating Theatre (OPT)	Operating theatre procedures and coding.
Radiotherapy (RAD)	Radiotherapy treatments dates and number of sessions.
Cancer Registry datasets (CR)	The cancer registry dataset includes some of the above data, which can be used for quality checking purposes, and additional data such as cause of death.

Table 4.12: Data sources used for the development of the pathways.

Class	Code	Name	Type	Data source	Frequency
Demographics	Q	Deprivation Score	Information	CR	100%
Demographics	A	Age at Diagnosis	Information	CR	100%
Demographics	Z	Death	Event	CR+ODS (ADM)	21.11%
Demographics	L	Clinical Trial	Information	ODS (ADM)	1.16%
Demographics	X	Other Cancers	Event	CR	21.32%
Diagnostics	D	Diagnosis & Staging	Event	CR+ODS (HIST+ADM)	100%
Diagnostics	G	Histology Grade	Event	CR+ODS (HIST)	84.51%
Diagnostics	I	Imaging	Event	ODS (RAD)	15.28%
Diagnostics	P	PSA Test	Event	ODS (LAB)	95.27%
Treatment	S	Surgery	Event	CR+ODS (OT)	33.61%
Treatment	R	Radiotherapy	Event	CR+ODS (RT)	20.75%
Treatment	C	Chemotherapy	Event	CR+ODS (ADM)	0.42%
Treatment	O	Orchidectomy	Event	CR+ODS (OT)	0.11%
Treatment	H	Hormone	Event	CR+ODS (ADM)	50.42%
Treatment	W	Active Surveillance	Event	CR+ODS (ADM)	22.16%
Treatment	N	No treatment	Information	CR+ODS (ADM)	2.21%

Table 4.13: Pathway dictionary for prostate cancer.

4.3.5 CaP VIS: A Visualisation and Integration System

Recommendations for further research in process mining [209], clinical decision support systems and expert systems [247] suggest that software that integrates complex data and generates graphical representations is needed to support the analysis and understanding of such data. This is particularly relevant in systems that provide unique integration by collating routine data, such as the one presented in this section. Therefore, a system that enables and supports both analytical processes (such as data/process mining) and clinical decision support was explored.

This section introduces a novel software and framework developed for the analysis and visualisation of the pathways and their integration with further clinical information. The CaP VIS (Carcinoma of the Prostate Visualisation and Interpretation System) software and framework were developed based on some of the guiding principles of model-driven architecture (MDA). The three goals of MDA are to provide portability, interoperability and reusability [248]. As such, MDA-

based approaches implement models and are capable separating the operation of a system from the details of its environment or platform. Further details on MDA are given in [249] and [250] and are not repeated here.

Although MDA architectures often require significant amount of planning, the development of the CaP VIS software relied on a Rapid Application Development (RAD) methodology. RAD refers to "a development life cycle designed to give much faster development and higher quality results than the traditional software development life cycle"[251]. RAD approaches favour the rapid implementation and delivery of prototypes and this is important to communicate effectively with the domain experts during the development. The RAD methodology consists of the following phases [251]:

- Requirements planning phase: the scope of the project, requirements and constraints are discussed with the domain experts.
- User design phase: this is a continuous interactive process that allows users to understand, modify, and eventually approve a working model of the system that meets their needs
- Construction phase: tasks include programming and application development, coding, unit-integration and system testing
- Cut-over phase: further testing and user training is carried out and a full working version of the software is complete.

The above steps were used as guiding principles in the development of the software and framework presented in this chapter. The software was developed in a platform independent programming language, Python, and the framework architecture is described in detail in the next section.

Pathways Database

The pathway database consists of an organised filing system where each patient pathway is stored in a single file and linked to an inverted index. A pathway engine is responsible for performing typical database operations as well as advanced operations and script programming. Although this implementation affects performance, typically due to slow file IO scanning operations, it also ensures that the database is fully portable and that changes can be made to individual patient files by non-technical staff with access. Full password protection of the environment was implemented and additional encryption was possible but not deemed necessary due to the limited number of users requiring access in the course of this research. Adding a new patient to the database simply requires uploading a comma separated file in the pathways tabular format into the specific database folder. Reading operations may be performed concurrently, however, write and change operations are exclusive to a master (administration) user with access to the system and pathways engine. Advanced operations including insertion of new data elements or propagation of model changes through the cohort are also possible using the pathways engine. Developed scripts also allow the database to be converted into a single table or multidimensional format for use with other database management systems. This system was considered robust and adaptable to changes in the data model or dictionary and it may hence be considered a model driven system, where the model is the backbone of the system design process. Further details on the pathways database are given later in this chapter. The next section describes the development and version history of CaP VIS.

4.3.5.1 Development and Version History

The first most important system requirement was that the system should enable the inspection and study of the Prostate Specific Antigen (PSA) biomarker trends.

A first system was developed in Microsoft Visual Basic (VB) and VB for Applications and, based on the pathways data model, allowed the automatic computation of plots with the complete PSA trend for each patient in the dataset. The resulting plots were then divided by treatment type and this provided interesting results. Together with the domain experts, an analysis of the produced plots was critical to determining further system requirements and the future developments, including a novel graphical representation of pathways data. An excerpt of results of this process, for each treatment type, is available in Appendix C 1.5.

Single Plot Interface, CaP VIS version 1

Upon inspection of the PSA trend plots with the domain experts, it became clear that these should contain additional information in order to explain the changes in the PSA. For example, the most significant drops in PSA reading should be associated with a particular treatment type. This led to the development of a more sophisticated visualisation system, capable of interpreting the pathways and transforming them into meaningful yet concise graphical representations. A single plot interface software was developed (in Python) together with the pathways engine, and comprised an architecture similar to that of the Model-view-controller [252] (MVC). In this implementation, the architecture encompasses three elements with specific purposes:

- the model, responsible for the read, transforms and interpretation of the pathways data using an extended dictionary (containing information on how events are drawn);
- the view, that receives the instructions based on the model and generates a graphical representation of a pathway;
- the controller, that communicates user or system requests and is responsible for the interaction between the model, the view and the system.

This MVC architecture implementation allows for changes to happen to the model without affecting the view or controller. Each element can be updated or changed independently. In this version, the view produced a static picture file containing the graphical representation (plot) of a pathway.

Figure 4.8 shows the layout of a plot and the areas where information is displayed. The y-axis represents the biomarker values (in this case, PSA) and the x-axis represents time t as defined in the pathway data structure. The biomarker readings are plotted in the centre and events (such as treatments or death) are marked with a vertical line (*V-Line*). Different treatments can be colour-coded differently and, above the plot, the corresponding pathway code, c, is shown in the *V-Line headings* area. The *footer text* area displays additional information pertaining to events (such as Gleason grades or patient age at diagnosis) and the *right column text* area on the right of the plot displays additional information on the patient that is not time-dependent, such as deprivation score, additional diagnoses or comorbidities. The model and the graphical representation are discussed in more detail in the next section and examples of pathways are also shown later.

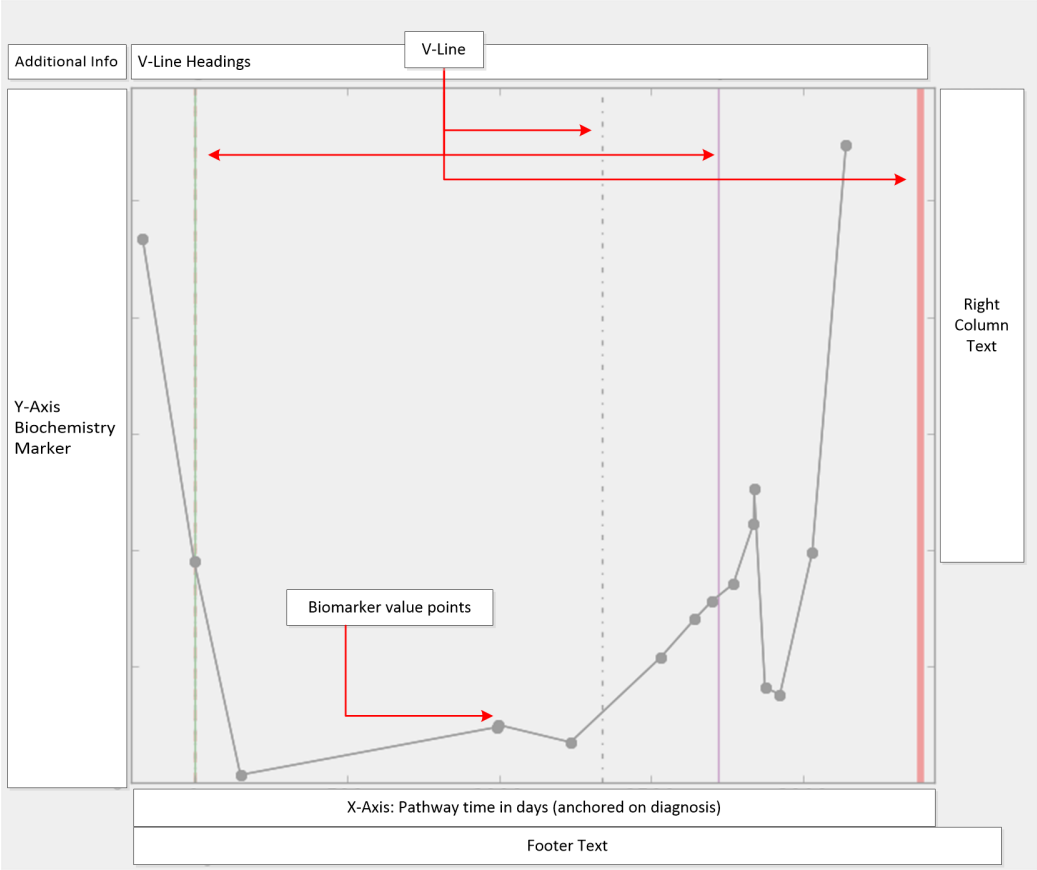


Figure 4.8: The schematic layout of a pathway plot.

Multiple Document Interface, CaP VIS version 2

Upon achieving a working solution to automatically produce graphical representations of pathways, a multiple document interface software was built in Python. This version of the software relies on an organization of the screen into mutually non-overlapping frames, one showing the static graphical representation of a pathway for a given patient, and four frames displaying the histopathology reports, the pathway details (in a similar fashion to Table 4.11), and pathway statistics. In this version, the software displays the static graphical representations already produced and saved as picture files. The software then reads information from the pathways database to compute basic statistics and display additional details.

This version of the software allowed a detailed inspection of patient pathways and this was particularly relevant and useful at meetings with the domain experts. A screen shot of the version 2 of the software is included in Appendix C 1.6.

Dynamic Multiple Document Interface, CaP VIS version 3

The final version of the software, at the time of writing, was comprised of all the elements in the previous versions with additional interaction capabilities and analyses tools and techniques. Instead of relying on static graphical representations of the pathways, the MVC architecture was embedded within the software, producing real-time plots of the pathways, as they are read from the database. This version also introduced dynamic visualisation where users have the options to zoom in, re-scale and navigate the pathway plot. This is particularly important as the scales of the plots may render some drawn objects too close to each other. A mechanism for graphical conflict resolution (i.e. avoiding overlapping elements) was also introduced. Additional windows and analyses were introduced and a full description of the architecture and the system functionality is given in the next section.

4.3.5.2 Architecture and Functionality

This section describes the CaP VIS version 3 architecture and system functionality. Figure 4.9 depicts the ways in which the data flows from the sources, and the steps involved in bringing detailed pathways into the visualisation and interpretation system or into the analysis or query engines. The file system and the ways in which the system is organised is also shown in Appendix C 1.7.1.

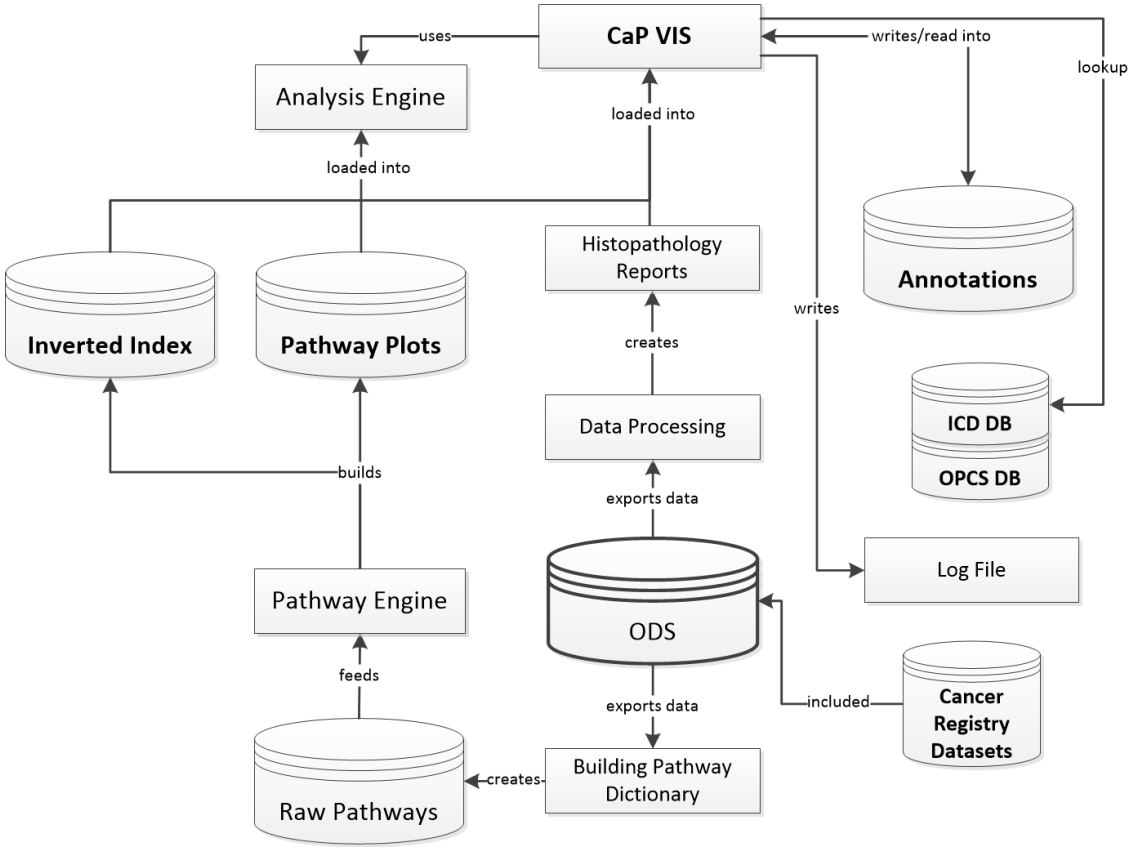


Figure 4.9: Data flow diagram illustrating the relationship between the operational data store (ODS, in bold), the pathway and analysis engine, the visualization and interpretation software (CaP VIS) and other interactions.

Figure 4.9 shows the process in which datasets were extracted from the ODS and in the pathway format defined in section 4.3.1 and used to build the pathway dictionary and raw pathways database. The pathways engine, which works with the information stored in the raw pathways database, is responsible for the segmentation, summarisation, cleansing and indexing of the raw pathways. Such operations together allow for the mapping, selection and retrieval of individual or groups of similar paths using regular expressions or *ad hoc* algorithms. The detailed pathways are organised by patient identifier and stored as plots that allow

an interpreter (the MVC) in the visualisation software (CaP VIS) to produce a detailed graphical representation. The interpreter will parse each activity from a pathway and, based on the dictionary and a set of rules determined for each element code, plot the corresponding graphical representation. An important feature of the visualisation system is to integrate the pathways with histopathological or further clinical information. A coding lookup table was added in order to translate and present diagnosis (ICD) and procedures (OPCS) codes. Because the time length of different pathways can vary considerably, it was important for the plot to be interactive, allowing zoom and re-scale as well as mechanisms for graphical conflict resolution. Sample output from the CaP VIS software is depicted in Figure 4.10 and shows a patient pathway and related information, including the pathway data format. The analysis engine can be used by the CaP VIS software to compute statistics for the pathways but it can also be used on its own to develop algorithms that work with the pathways data. Section 4.3.9 demonstrates the use of the analysis engine in computing completeness (quality) scores for the PSA values in pathways.

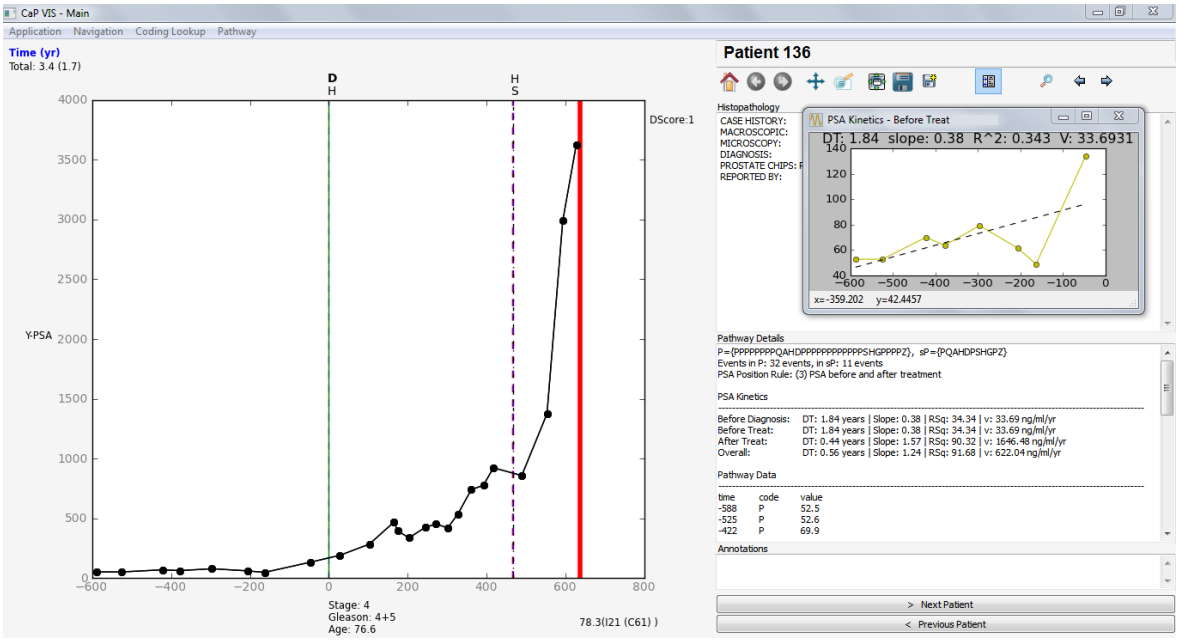


Figure 4.10: The CaP VIS system illustrating a castration resistant patient pathway and related information.

Figure 4.10 shows a patient that was first treated with hormone therapy (code H) and had a subsequent palliative prostatic resection (code S). The pathway is plotted on the left side of the screen, where the trend of the PSA biomarker is visible, together with diagnosis line (code D) and treatments (codes H and S). On the right, there are three sections showing the histopathology report, the pathway details in the expected format, the annotations section, and a fourth section (hidden) contains pathway statistics. An overlapping window shows in detail the PSA kinetics before treatment, computed for this pathway (including PSA doubling time (DT) and Velocity (V)). The toolbar above the histopathology section allows the user to zoom and pan the plotted pathway as well as to save the plotted figure to file, search for particular patient pathway, navigate to the previous or next patient pathway, and toggle between showing the histopathology report or additional pathway statistics. The latter shows additional information such as data quality, computed NICE risk, basis of diagnosis, and availability of

data elements from the pathways dictionary. An anchor window (hidden) is also available and presents a summary of the cohort that allows for fast browsing of the cohort.

CaP VIS's graphical user interface is underpinned by a console system that logs the user interactions and provides additional details on operations and data. A screenshot of a typical console usage is given in Appendix C 1.7.2.

A detailed list of the program's menus and options is given in Appendix C 1.7.3. An additional and important functionality, explored in more detail later, is the system's capability to plot an additional curve alongside the biomarker curve. This allows the inspection of two curves (for example, PSA and Alkaline Phospathase) and their interaction over time. The first version of the pathways data dictionary included the 16 core data elements already shown, however, in subsequent versions, additional biochemistry and hospital events were added. This is explained in more detail later in section 4.3.8 and demonstrates the flexibility of the model, software and framework in adapting to new data and specifications.

The next two sections show two additional software modules developed to interact with the CaP VIS framework but that are independent from the main software described above.

4.3.6 Cohort Visualisation

One of the advantages of using the pathway data structure is the ability to produce succinct sequences of activity codes. Truncating the sequence strings (i.e. collapsing sequentially repeating elements into one) enables the aggregation of pathways with similar sequential activities. A web-based software, called ExploraTree, was developed tp produce and display an interactive tree of the full cohort of 1,904 prostate cancer patients, and it is based on the data elements available in the pathways data dictionary. The technologies used include HTML, CSS, JSON, JavaScript and the InfoVis toolkit. The pathways engine was used to produce the correct data format for a tree representation using JSON and the InfoVis JavaScript toolkit.

In order to accurately aggregate patients with similar sequences of activities, new data elements were introduced in the pathway dictionary for this software. In the core data dictionary, a patient's death was encoded by only one data element (code Z). Patients who died of prostate cancer were kept with code Z while those who died of other causes were identified with code Y and those who survived, with code X. This ensures that all patients have a terminal element indicating whether they are alive. Because in this cohort not all patients are followed-up the same amount of time, all terminal elements (X,Y,Z) were given additional nodes that represent the amount of time the patients were followed-up in years (1 to 5 and '+' for over 5 years). This is illustrated in Figure 4.11.

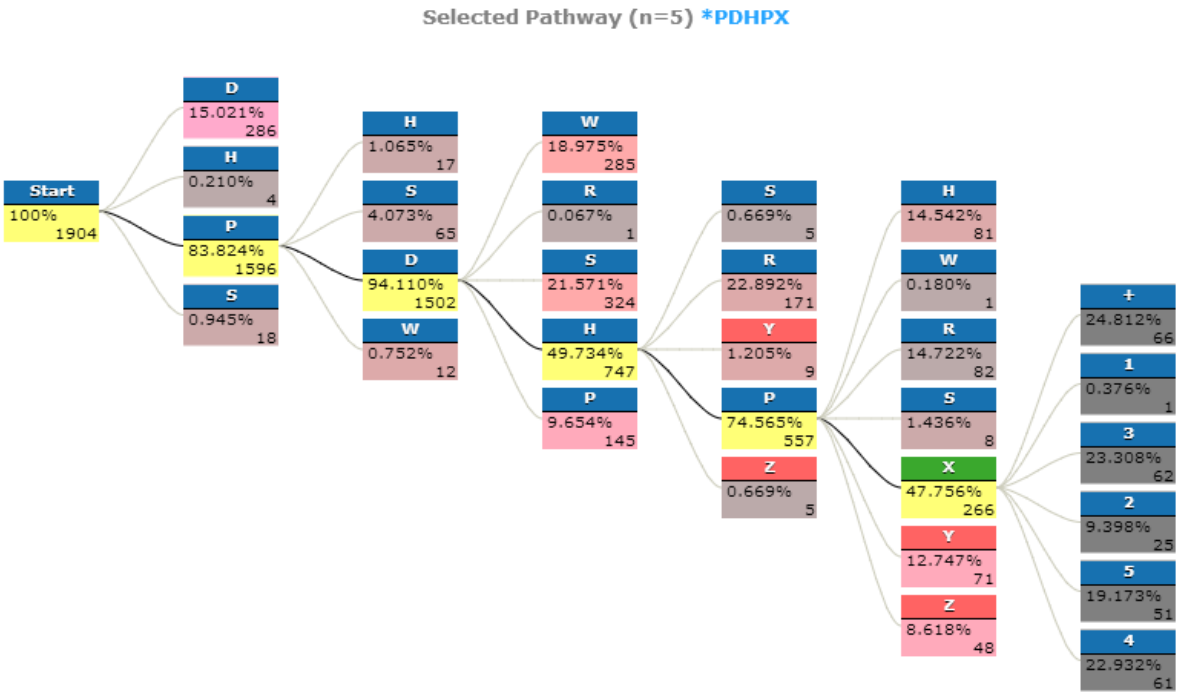


Figure 4.11: CaP VIS ExploraTree software displaying a selected pathway (patients with the same sequential activities). The selected pathway nodes are highlighted and terminal nodes are marked as red for patients that died and green for patients that were last seen alive in this cohort.

Figure 4.11 shows the cohort tree and highlighted sequence $\langle P, D, H, P, X \rangle$, that is, patients who started their pathway with one or more PSA tests (code P, n=1596), followed by a diagnosis of cancer (code D, n=1502), hormone therapy as first treatment (code H, n=747), other PSA test(s) (n=557) and finally were last seen alive in this cohort (code X). 90% of patients with this pathway (n=266) were followed-up 3 or more years and one patient was only followed-up less than one year.

This aggregation also allows comparing patients that followed similar pathways but who died of prostate cancer ($\langle P, D, H, P, Z \rangle$). In the case of patients with a

sequence prefix $\langle P, D, H, P \rangle$, 9% (n=48) died of prostate cancer (code Z), 13% died of other causes (code Y), 48% survived, and the remaining patients continued with other activities (H,W,R,S).

Visualising the cohort in this manner is important as it enables the selection of subsets of data for specific clinical studies as well as an inspection of the sequential routes that patients take through care. The sequence highlighted in figure 4.11 corresponds to the most common route (with most support on each node sequentially).

It is possible to add more meaning to the visualisation and to the pathways by introducing additional data elements and remodeling the data dictionary. For example, instead of using a single code for diagnosis it is possible to have a breakdown of the tumour staging or Gleason grade at diagnosis so as to group similar sequences with this information instead. However, due to the small size of this cohort, increasing granularity in the pathways dictionary would result in fewer patients in each node. For this reason no additional changes were made to the pathways dictionary used for the ExploraTree, but future work with a larger cohort is envisaged and the remodeling of the pathways dictionary is discussed again later in this chapter.

4.3.7 Exploring Diagnostic Profiles and Mortality

A second system developed within the CaP VIS framework, called RECON Diagnosis, explores the survival of patients with a given diagnostic profile based on the pathways data in the cohort. The system was also developed in Python and it works with the pathways database and engine. The user interacts via a predefined set of commands in a console window. The system requires the user to enter a diagnosis profile containing one or more of the following variables: PSA at diagnosis, Gleason grade, tumour staging and age group. The system then computes the number of patients in the cohort with the same profile and, given a 3 year survival period, the likelihood of those patients dying from prostate cancer within that time period. Because not all patients in the cohort were followed up the same amount of time, a smaller subset was selected. All patients diagnosed between 2005 and 2009 were followed-up at least 3 years and this was the subset of the cohort used by this program (n=1,416).

Figure 4.12 shows an example output from the system. In this example, the selected profile was patients with a low PSA at diagnosis (as defined in NICE guidelines, a reading under 10 ng/mL) and a Gleason grade of 9. The program identified 52 patients (19% deaths) with this profile and computed a 2x2 table that compares patients with the profile against those that did not have the profile. Odds ratios (OR), p-value and confidence interval were also automatically computed. The program then provides an interpretation of the results and in this example, patients with the selected profile were at a higher risk of death (odds ratio 2.558) and the result is statistically significant (p-value < 0.05). This confirms that a high Gleason grade is associated with a greater risk of death from prostate cancer. In addition to survival, the program then shows the different treatment options that patients with the selected profile underwent and a breakdown of the number of deaths by treatment modality. In the example, hormone therapy had the least deaths (14%).

```
 CaP VIS - RECON Diagnosis
 Exploring options for given diagnosis profiles.

 When making a profile please give a numeric option
 or leave blank to remove filter (include all).

----------------------------------------
   NEW PROFILE
----------------------------------------
1. PSA at Diagnosis [0-none,low-1,med-2,hi-3]: 1
2. Gleason Grade [0-none,6,...,10]: 9
3. TNM Stage [1-in situ,2,3,4,6-not possible to grade]:
4. Age Group [1-(<=55), 2-(56-65), 3-(66-75), 4-(76-85), 5-(>=86)]:


----------------------------------------
   RESULTS
----------------------------------------
Profile: low PSA, Gleason 9, *any* TNM, *any* Age Group
Cohort: 1460 patients, 131 deaths (8.97%)

N in Profile:   44      (3.014%)
N Deaths:       8       (18.182%)

Status Alive/Dead checked at 3 years.


----------------------------------------
   MEASURE OF ASSOCIATION
----------------------------------------

2x2 Table:
                -Dead-  -Alive-
Profile         8       36
No Profile      123     1416

Odds Ratio:     2.558 Increased Risk
p-value:        0.025 Significant
95% CI:         2.558 to 2.558

Interpretation: The risk of death within 3 years among individuals with the sele
cted profile at diagnosis ( lowPSA,G9 ) is 2.558 times that of those who do not
have the same profile.


----------------------------------------
   TREATMENT OPTIONS
----------------------------------------
Type            N       %N      Deaths  %Deaths
Hormones        29      65.91%  4       13.79%  *
Horm+Radio      1       2.27%   0       0.0%    *
Surgery         11      25.0%   3       27.27%
WatchWait       3       6.82%   1       33.33%



 Quit program? [y/N]
```

Figure 4.12: CaP VIS RECON Diagnosis System output showing patients with a profile made up of a low PSA at diagnosis and a Gleason grade sum of 9.

Variable at Diagnosis	Odds Ratio	p-value
PSA high >20 ng/mL	6.268	<0.005
PSA med 10-20 ng/mL	0.508	0.009
PSA low <10 ng/mL	0.255	<0.005
Gleason sum 10	11.661	<0.005
Gleason sum 9	3.882	<0.005
Gleason sum 8	1.679	0.154 *
Gleason sum 7	0.102	<0.005
Gleason sum 6	No Deaths	
TNM 4: Fixed or invades adjacent structures	8.369	<0.005
TNM 3: Through prostate capsule	0.678	0.178 *
TNM 2: Confined within prostate	0.327	<0.005
TNM 1: Not palpable or visible	No Deaths	
Age >= 86	4.263	<0.005
Age 76-85	2.562	<0.005
Age 66-75	0.534	0.002
Age 56-65	0.344	<0.005
Age <= 55	0.753	1 *

Table 4.14: List of variables at diagnosis and their effect on 3-year deaths from prostate cancer. P-values marked with * indicate the variable is not statistically significant at the 0.05 level.

The previous example indicated that patients with a Gleason 9 and a low PSA were 2.558 times more likely to die from prostate cancer within 3 years than those that did not have that profile. Table 4.14 shows the results of the program for individual variables and allows to compare the improvement in risk against those who had only a low PSA (OR 0.255) or a Gleason 9 (OR 3.882). The low PSA combined with a Gleason 9 lowered the risk for those who have Gleason 9 irrespectively of other variables. Table 4.14 also shows that Gleason 10 is the variable that contributes the most to a higher risk of death from prostate cancer within 3 years (OR 11.661). This is followed by tumour staging 4 (OR 8.369), high PSA (OR 6.268) and age over 86 (OR 4.253). The variables with most protective effect are Gleason sum 7 alone (OR 0.102), a low PSA (OR 0.255), TNM 2 (OR 0.327), and age 56-65 (OR 0.344).

The developed program is flexible and can be adapted to different data elements and changes in the pathways dictionary. Future work includes extending the profiles to include the computation of risk for any given pathway sequence (or segment) and different outcomes (for example, hormone escaped, development of metastases, biochemical recurrence after treatment).

4.3.8 Enhancing the Core Dictionary

The ability of CaP VIS and the pathways dictionary to incorporate new data elements is key; it enables the remodeling of the data and pathways, and as a result, a broader scope for studies and inspections that were otherwise not possible with routine hospital data.

With the CaP VIS framework in place and the core dictionary, new data elements were added on biochemistry and hospital events from the prostate cancer ODS. 22 biochemistry tests were selected together with the domain experts and are listed in Table 4.15. The percentage of patients who have at least one reading of the test in their pathways is also given in Table 4.15. This is important as it was previously unknown how many of these patients had particular blood tests throughout their pathway. The reasons for having these test may in most cases not be specific to the prostate cancer pathway yet they might enable studies with new data.

Biochemistry Test	% Before Diagnosis	% After Diagnosis	N (After Diagnosis)
Creatinine	98.79	88.13	1678
Urea	98.79	88.13	1678
Sodium	98.79	88.08	1677
Haemoglobin (Hb)	98.06	83.25	1585
White Blood Cells	98.06	83.25	1585
Mean Corpuscular Volume	98.06	83.25	1585
Mean Corpuscular Hb	98.06	83.25	1585
Albumin	94.75	78.31	1491
Alkaline Phosphatase	94.64	78.20	1489
Bilirubin	94.28	77.63	1478
Cholesterol	81.30	63.60	1211
Random Plasma Glucose	81.67	63.13	1202
HDL Cholesterol	73.37	54.15	1031
Total Cholesterol / LDL	73.16	53.57	1020
Triglycerides	64.23	49.37	940
LDL Cholesterol	64.18	49.37	940
Calcium	79.67	44.70	851
Fasting Glucose	45.69	22.74	433
Testosterone	12.50	4.41	84
Aspartate Transaminase	1.10	0.42	8
Vitamin D	0.68	0.26	5
Gamma-Glutamyl Transpeptidase	0.00	0.00	0

Table 4.15: List of additional biochemistry tests added to the pathways database and percentage of patients having each test before and after diagnosis of prostate cancer.

The new data elements demonstrate the flexibility of the data model and allow a novel inspection of routine data. Figure 4.13 shows four pathway plots for the same patient, a 69 year old diagnosed with stage 3 prostate cancer and a Gleason sum of 9. Plot A shows the original plot where the PSA is seen to have dropped after the patient underwent hormone therapy. The patient died of prostate cancer. When producing this pathway's plots, the dictionary was extended so that the treatments retrieved from the local cancer registry appear with a suffix "1" in the vertical lines' headings. In this case, regarding the date when the patient first

commenced hormone therapy, a time discrepancy of 51 days was seen between the two data sources, where the hospital recorded the later date. Indeed additional data quality inspections are possible and this is explored in the next section.

The discrepancy in dates, in this case, did not introduce uncertainty in this pathway as the effect of the treatment is seen in the subsequent PSA readings. The pathway plot then shows a PSA relapse (of two readings from the nadir) and shortly after the last PSA the patient died of a pulmonary embolism (ICD I26) and prostate cancer (ICD C61) as a secondary condition leading to death. Also shortly before death the patient was diagnosed with a secondary and unspecified malignant neoplasm of inguinal and lower limb nodes (ICD C77.4). This was revealed by the additional data collected on hospital episodes (and clinical coding from discharge letters) and discussed again later.

Figure 4.13 Plot B shows an additional blood test, Alkaline Phosphatase (ALP) and its normal range in the shaded area. When a patient's advanced cancer metastasises to the bones, ALP can be increased due to active bone formation. Indeed studies have shown that prostate cancer patients with a serum ALP reading of more than twice the normal upper limit had a significantly lower survival rate than their respective counterparts [253]. This is observed in this pathway, however, other pathways may show an increased ALP due to other reasons such as an obstructed bile duct or liver disease.

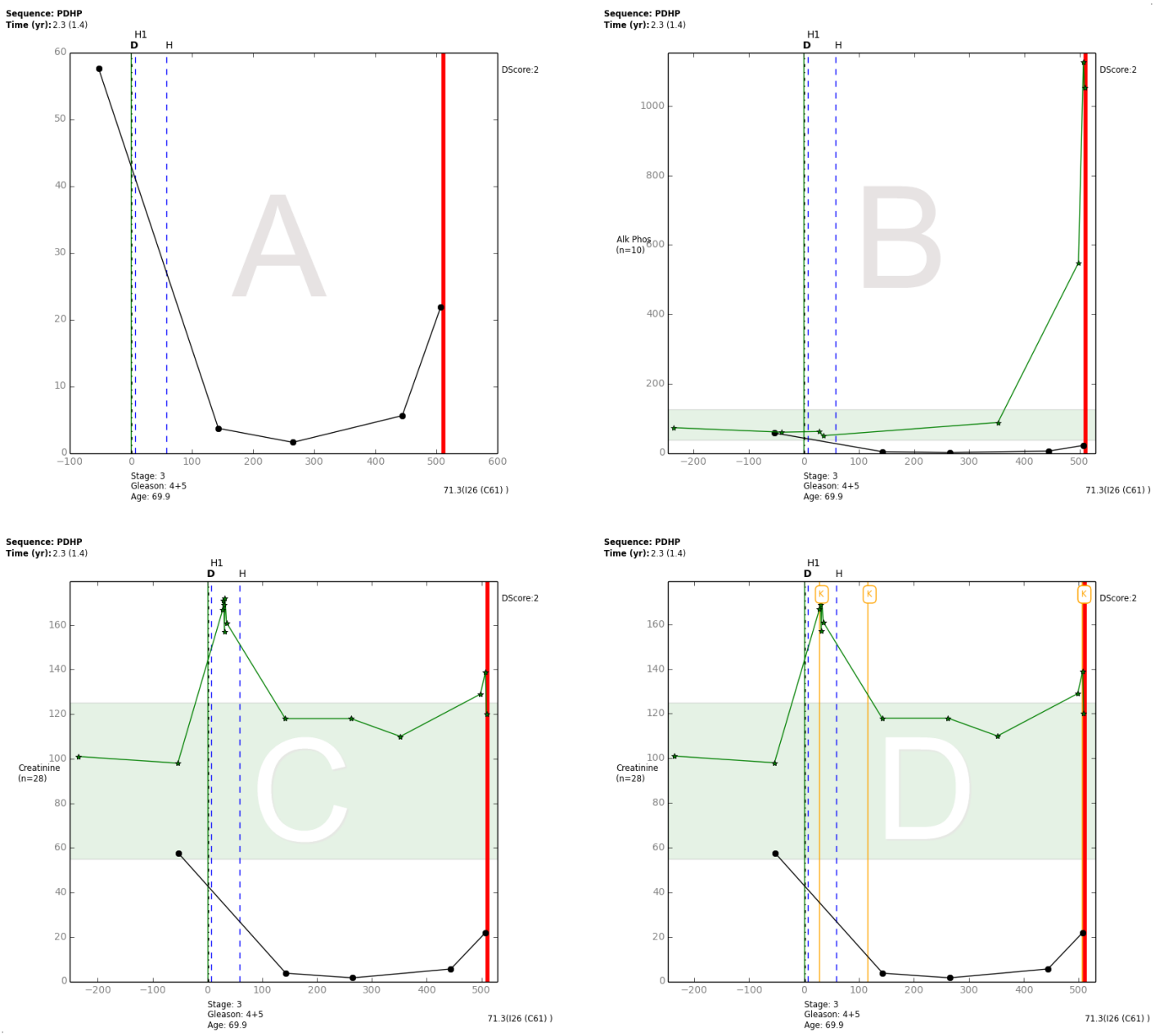


Figure 4.13: Four pathway plots of the same patient (175) with sequence $\langle P, D, H, P \rangle$. Plot A shows the original plot with the PSA trend alone. Plot B shows the same information as plot A with additional Alkaline Phospathase readings and their normal range (shaded area). Plot C shows Creatinine readings and Plot D shows the same information and hospital events (code K).

Lastly, figure 4.13 plot C and D show another blood test, Creatinine. Creatinine has been reportedly associated with more advanced disease and decreased survival [254], however, any condition that impairs the function of the kidneys is likely to raise the creatinine levels in the blood and act as a confounding factor. In plot C, a flare in the values of Creatinine readings was observed within the first 3 months. When introducing additional data elements from the hospital statistics, in plot D, a hospital episode (marked with pathway code K) was found with an associated primary diagnosis of acute kidney failure. Although a kidney stone was not coded in this (or any) episode for this patient, a catheterisation of the bladder was performed during the same hospital visit, and an inspection of the patient notes confirmed a kidney stone was the cause of the acute kidney failure. The second hospital episode in this pathway, also marked with code K, was for the removal of the catheter, and the last hospital episode included a diagnosis of a secondary and unspecified malignant neoplasm of inguinal and lower limb nodes and a pulmonary embolism, caused by the first. This level of information would also allow, for example, in other cases, to evaluate renal impairment and prostate cancer. Indeed, in this respect, it has been reported that renal impairment in men undergoing prostatectomy represents substantial and unrecognised morbidity [255].

The introduction of additional hospital data helped to explain the Creatinine flare for this patient and provided interesting insights that would otherwise not be explored, particularly in any retrospective clinical studies. The pathway plots provided sufficient information for the interpretation of this pathway and the patient notes were only used to confirm it. However, this is not possible across all pathways as information varies in respect to completeness and data quality. Furthermore, discrepancies in treatment dates across data sources may introduce additional challenges. As such, it is important to be able to differentiate between

pathways that have a sufficient information and those that do not. The evaluation of the completeness and utility of the generated pathways for investigating biomarker trends is explored in detail in the next section.

4.3.9 Assessing Data Completeness using Biomarker Information

Routinely collected data can vary in quality and it is important to assert the quality of all elements in the pathway so that particular paths can be selected or discarded for clinical analysis. We already discussed that upon extraction from the ODS, data elements were cross-validated against a trusted source, viz. the cancer registry. However, the biomarker information included in this study enabled additional quality assurance. To this effect, methods of computing the completeness of pathways from the biomarker information are investigated.

4.3.10 Rule Based Scores

Given the defined dictionary and its underlying format, it is possible to create a knowledge base of rules to aid the process of computing completeness scores for particular elements of the pathway. It is often difficult to convey and analyse biomarkers' information in pathways but here it was possible to compute their trends and to allow those computations to inform on the quality of the pathway. In the particular case of prostate cancer, the trend of PSA readings across the pathway is of interest. Two major sets of rules in which the biomarker can be used to assess the completeness of a pathway were identified, guided by domain experts. The first set of rules relies on the position of biomarker readings in the pathway, whereas the second relies on identifying clinical interventions that justify the changes in biomarker values. Rules can be applied programmatically

using the analysis engine.

Positioning of biomarker readings

As some of the intended clinical investigations pertain to PSA trends and associated treatments, it is important to have complete PSA trends within a pathway. In this context, a pathway should include biomarker readings before and after treatment so that the effect of treatment on the biomarker can be elucidated in posterior analyses. We can therefore compute a partial score of a pathway as a result of a set of rules on the occurrence of PSA readings. The rules are presented in Table 4.16. with their respective score and the percentage of pathways where the rule applied. The computation of the positioning score involves iterating through pathways codes and flagging occurrences of the PSA and their position with respect to treatments (excluding active surveillance). The most informative pathways should have one or more readings before and after treatment and the least informative have no PSA readings.

Positioning Score	Rule Description	Coverage
0	No PSA readings found.	4.70%
1	One or more readings found before treatment (or no treatment) and none after treatment.	4%
2	One or more readings found after treatment and none before treatment.	8.20%
3	One or more readings found before and after treatment.	82.90%

Table 4.16: PSA availability and positioning rules with respective scores and coverage.

Substantiation of biomarker variation

Further rules can be devised to ascertain quality. For example, biological variations, in this case expressed by the PSA, should often be accompanied by evidence of some clinical intervention. An analysis of the PSA curve can be undertaken

to identify major changes in PSA readings. The most significant drop in PSA should be associated with treatment to the prostate. A complete pathway should be able to provide explanations for such drops in the form of some clinical intervention. In this case, the computation of a score involves looking at every pair of PSA readings and then identifying the maximum absolute drop. This is followed by searching between the pair of values to identify an element of substantiation, which in this case study was set to be any radical treatment. The result of this rule is a Boolean value, stating whether substantiation of a large change in the biomarker trend was detected.

Overall score

An overall score for completeness can then be computed based on both positioning of biomarker readings and substantiation of major variation. It is worth noting that pathways that receive a positioning score of 0 or 1 could not have substantiation by definition as no PSA values appear after treatment. The overall score is an ordered set of values in which the highest score is awarded to the pathways with the highest positioning scores that are substantiated. The overall scores are given in the next section in Table 4.18.

The next section discusses the main results of the developed framework, the pathways data model, and the computation of pathway data completeness for quality assurance.

4.3.11 Results

The development of a framework to build, analyse and visualise pathways from routinely collected hospital data made it possible to create individual patient pathways and respective graphical representations for 1,904 patients whilst integrating clinical information from several HIS.

The core data dictionary contains 16 elements, described in Table 4.13. The data sources specify whether the elements were collected from the ODS (hospital systems, together with an abbreviation of the respective system) or the cancer registry (CR). Elements present on both sources have been cross-validated so their quality is assured. The data integration and the pathways data model, however, allowed discrepancies in the timeliness of events to be highlighted. As a results, not all data elements were kept due to data quality issues.

Regarding biopsies, they are only coded if performed as an inpatient event and hence only extensive biopsies were retrieved. As a result biopsy events were removed from the dictionary and are not used in this thesis, but can be kept for future studies. The frequency of imaging events was low (only captured imaging events on 15% of all pathways) and it reflects the nature of the retrieval methods from radiology, which are based on a text search of the word prostate.

The pathway dictionary in Table 4.13 also gives the percentage of patients who died in this cohort during the time of observation (i.e. pathways including a death event, 21%). These deaths are not exclusive to prostate cancer and the percentage should not be used to determine a measure of survival from prostate cancer. It was possible, however, for the developed CaP VIS RECON and ExploraTree software to interpret this further by extending the pathways data model. Future survival analyses are also planned.

Further data elements were also added to the pathways dictionary including biochemistry tests, as co-morbidities and hospital stays which may or may not be related to prostate cancer. This was explored earlier in this section and it is briefly discussed again later.

The analysis engine computed descriptive statistics such as the various frequencies of the elements of the dictionary. A summary of the pathway statistics for all pathways is given in Table 4.17. Descriptive statistics are important as they con-

vey information about the pathways. They can also give rise to quality indicators but these methods alone were not sufficient to determine quality.

Statistic	**Value**
Average number of unique activities	4.66 (SD 1.0)
Average pathway length	1,795 days (SD 1,724)
Average pathway length from diagnosis	1,017 days (SD 653)
Most common activity code	P (90.5%)
Five most common start codes	P (72.9%), X (11.7%), D (7.4%), G (4.1%), L (1.2%)
Five most common terminal codes	P (73.2%), Z (21%), G (3.1%), W (0.6%), R (0.5%)
Total number of unique pathway sequences	694
Most common pathways' sequence (repetitions truncated)	$\langle P, D, G, H, P \rangle$ 7.1%, $\langle P, D, G, W, P \rangle$ 6.8%
Most common treatment regimes (where first and second treatment modality are within 92 days of each other)	H (47.6%), S (27.2%), W (16.7%), SW (3.1%), SH (1.2%)

Table 4.17: Summary of pathway statistics from the core pathways dictionary.

The pathways data model and analysis engine allowed the computation of completeness scores for the purpose of selecting pathways with enough data to analyse the biomarker trend. The following sections show the results of the application of the rules and their impact on quality assessment.

Inspection of the positioning of readings

The application of rules on the positioning of the PSA biomarker allowed the identification of 1,579 (82.9%) pathways where it was possible to plot the trend of the biomarker through treatment (scores S3+S5 in Table 4.18). The framework presented above made possible the inspection of data elements in relation to other events plotted chronologically. It is also possible to compute the proximity between elements. For example, treatment elements within 90 days were grouped together to form treatment packages. The type of rules proposed here allow for the assessment of the timeliness and completeness dimensions of data quality.

Inspection of the substantiation rule

Overall it was possible to ascertain the biomarker variation substantiation rule for 61.1% (n=1,321) of the pathways. It was also identified that 4.1% (n=79) of all pathways with two or more PSA readings, had a constant or always rising PSA trend. These were merged with the overall substantiation number, making 65.2% (n=1,242) the total number of pathways with a positive substantiation rule (scores S4+S5 in Table 4.18). Substantiation does not occur when a treatment element is not present in the biomarker interval of interest, or if the treatment date is inaccurate. This may indicate missing information. The substantiation rule allows for the elimination of 20.6% pathways with insufficiently accurate information to study the biomarker trend. This rule enables the assessment of the plausibility, completeness and timeliness dimensions of data quality.

A hybrid scoring system

A hybrid scoring system for the completeness of the pathways combines both biomarker rules described above (positioning and substantiation) and it is given in Table 4.18. The overall score ranges from least complete (score S0) to most complete (score S5) and were automatically computed based on the criteria set in the rules above. This particular set of rules aims to identify the completeness of the pathways based on the prostate cancer biomarker. It is also possible to extend the framework presented in this chapter to create other quality scenarios involving more robust and detailed rules based on biomarkers or other aspects of the pathways. Examples of pathway plots automatically drawn by the CaP VIS system are available in Figure 4.14 and illustrate each of the five completeness scores.

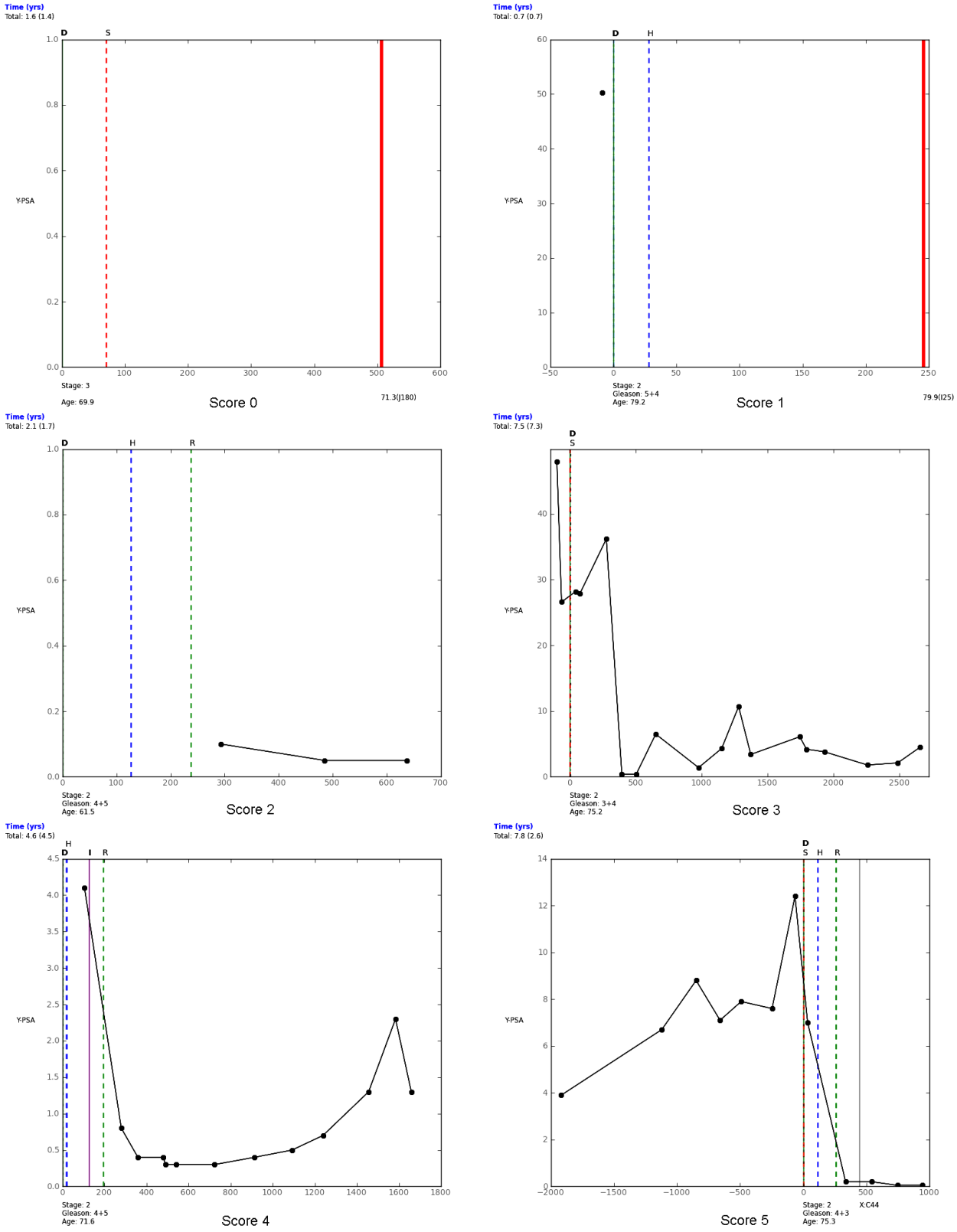


Figure 4.14: Examples of pathway plots drawn by the developed CaP VIS system for each of the six possible completeness scores.

Overall Score	Biomarker			Frequency	Average Number of Unique Elements
	Positioning	Substantiation	Description		
S0	0	N/A	No readings found.	4.70%	3.26 (SD .64)
S1	1	N/A	One or more readings found before treatment (or no treatment), and no readings after.	4%	4.72 (SD 1.02)
S2	2	N/A	One or more readings found after treatment, and no readings before.	5.40%	4.71 (SD 1.03)
S3	3	No	One or more readings found before and after treatment.	20.60%	4.56 (SD .99)
S4	2	Yes	One or more readings found after treatment and major biomarker variation explained.	2.90%	4.70 (SD .88)
S5	3	Yes	One or more readings found before and after treatment and major biomarker variation explained.	62.30%	4.80 (SD .92)

Table 4.18: Completeness scoring system for PSA trends in prostate cancer pathways.

Further analysis of data quality

In order to understand further data quality issues in relation to completeness scores, variations between key demographic information and completeness were explored using the framework presented above. For example, with regards to the UK indices of deprivation (ID 2004) obtained from the cancer registry, it was possible to observe that 72.2% (n=65) of score S0 pathways were associated with the least deprived patients (two quintiles). Similarly, the least deprived have the lowest proportion of score S5 pathways (47%, n=116). This could be largely due to wealthier patients seeking private health services, which would result in some of their data missing from the HISs used to build the pathways. However, it is difficult to confidently corroborate this observation given the collected data. Further data sources would be required, particularly from primary care, to accurately corroborate these findings.

A further analysis on surveillance regimes made possible the observation that 7% (n=25) of those on surveillance (as first treatment) had a subsequent treatment within at least a year, and therefore left surveillance. For those that did not have a subsequent treatment (93%, n=317), it was possible to investigate any substantial drops in PSA, which may be indicative of unrecorded treatments. By establishing a drop ratio calculated as the maximum PSA drop divided by the PSA at diagnosis, it was noted that, 31% (n=97) of pathways on surveillance regimes show a drop over a 0.5 ratio whereas 16% (n=50) had a drop > 1. This analysis is only preliminary, but it may indicate that patients received treatment yet these have not been recorded or carried out at this hospital. Such pathways could be excluded from analyses or be further explored to seek plausible reasons for the unexplained variation in biomarker trend. Again, this is an example of the type of analysis enabled by the framework and the pathways data structure.

The analyses on quality also led to improvements in the data collection process. It was possible, for example, to identify patients that only had PSA readings after

treatment as well as those without PSA readings before diagnosis. This process yielded a small number of pathways (0.2%, n=38) where there had been earlier PSA readings but these were not linked to the patients main hospital number in the hospital data warehouses and hence were missed on retrieval (not present in the ODS). Such cases are not expected to occur frequently and do not affect any of the hospital administration or clinical operations. However, they can diminish the amount of information available for the use of routinely collected hospital data for analysis. In this instance, as only a small number of cases were affected, they were manually fixed. The exercise, however, uncovered the need for further checks by the hospital on the data warehouse to ensure consistency of recordings.

Framework and developed software

The developed framework and visualisation software enabled the visualisation of all 1,904 patient pathways with their corresponding biomarker trends. This gives clinicians access to trends that may have been previously much harder to observe. Furthermore, the system is flexible and extensible to include other data elements such as blood readings. For example, Figure 4.15 plots the PSA values and the Haemoglobin (Hb) readings. The shaded area is the normal range for Hb. In this case, the drop in Haemoglobin on the day of surgery reveals perioperative bleeding. This information, when computed for all patients, would enable a study of the length of time that patients take to recover after surgery. This illustrates the flexibility of the combined framework and visualisation tool and provides access to a number of studies with data that was otherwise not readily available or contextualised.

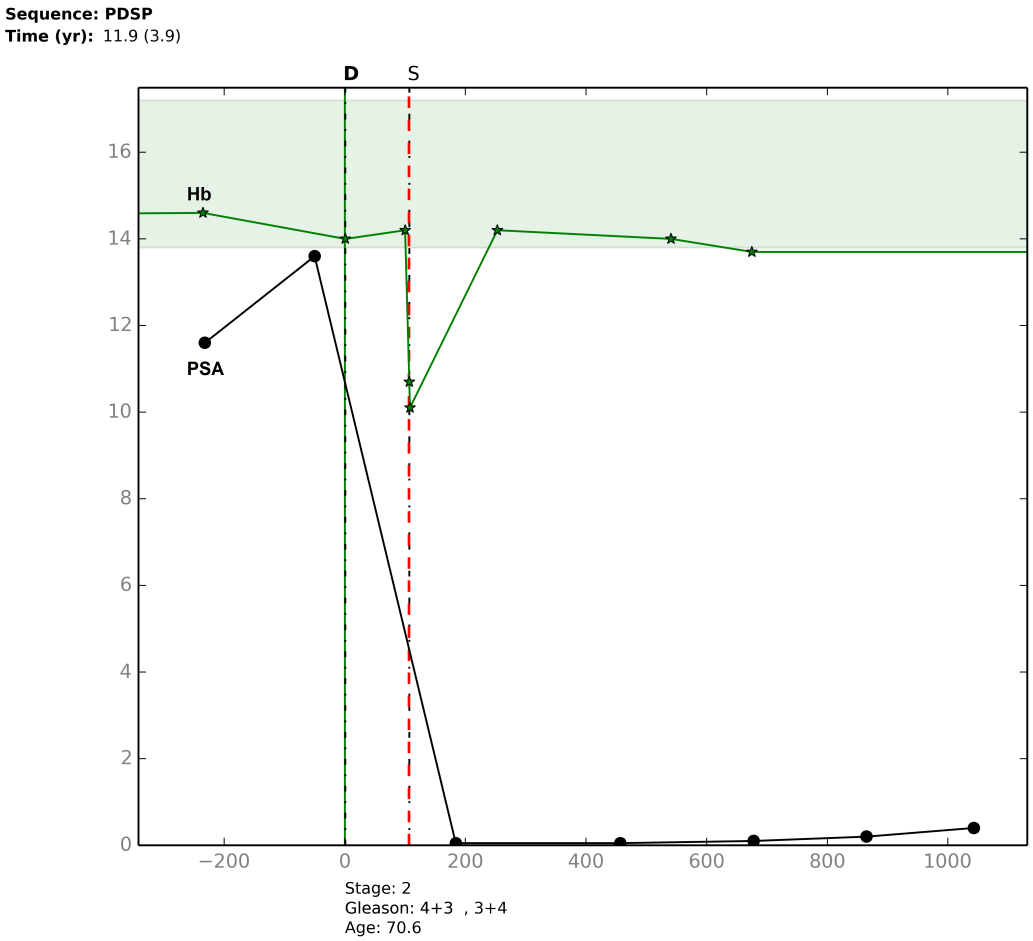


Figure 4.15: Pathway plot showing the PSA (round markers) and haemoglobin readings (star markers) together. As a result of the prostatectomy event (S) the PSA dropped and haemoglobin also dropped due to normal perioperative bleeding. The shaded area denotes the normal range for haemoglobin.

Additionally, the framework presented made possible the inspection of all five data quality dimensions described in [37], including those that are least often assessed. Currency (or timeliness) has been considered a fundamental dimension yet it is often not assessed and only measured using a single approach [37]. One of the key variables of the pathway data structure presented here is the inclu-

sion of time since pathways are arranged chronologically and allow for concurrent elements. For example, in the case study, treatments within 90 days were considered a treatment package. It is also possible to identify and discard data elements not relevant at particular intervals, as exemplified by the positioning rules. Furthermore, the plausibility and concordance dimensions were assessed using the substantiation rule, the completeness dimension using the positioning rule, and correctness dimension assessed by cross-referencing against the cancer registry.

The proposed framework and developed software should also allow for the selection and extraction of particular datasets with complete data for process mining and other analysis. It has been reported that the evaluation of the quality of process mining event logs relies on trustworthiness (recorded events actually happened), completeness and well defined semantics [256]. These can be achieved by selecting pathways with required data points using the proposed framework. Furthermore, the visualisation system allows for the close inspection and contextualisation of pathways, illustrating particular paths with similar features such as the ones exemplified in the Figure 4.14. In summary, the proposed framework, when used in hospitals, would facilitate the retrieval, selection and inspection of patient pathways and also the further steps of data mining analysis using appropriate methodologies.

4.4 Conclusions

This chapter introduced a novel data model tailored to data-driven patient-centric pathways and a novel framework for the construction of pathways and the summarisation, visualisation and querying of complex clinical data. This can be seen as an evidence based clinical decision support system that enables both clinicians and researchers to query individual pathways and select cohorts with similar features.

Despite current data and process mining techniques not being extensively applied in health care, they continue to be suggested to help uncover similar or frequent trends in health care processes. Some applications in specific domains have been fruitful but they rely on structured data carefully prepared for process mining and systems that produce event logs. Section 4.2 explored the feasibility of state-of-the-art process mining and association rule mining algorithms in detail and found these to be unsuitable to mine simple system-level paths with low granularity (where tasks are essentially binary). Therefore in more complex and unstructured processes such as those depicted in section 4.3, such techniques would also be found unsuitable.

A framework which enables the secondary uses of routinely collected hospital data was developed and presented in this chapter. The main components of this framework (Figure 4.9) are the ODS containing patient-centric data, used to build the pathways based on the methodology presented in Figure 4.3.3, the pathways engine, analysis engine, and the visualisation software. It has been noted that for immediate decision-making by clinicians at the point of care, information needs to be brief and easily interpreted [34]. The graphical representation of the pathways has facilitated interpretation, communication and debate between experts, and further work is needed to assess the pathway plots in other settings and domains.

The underlying pathway data structure proposed in this thesis, is in some aspects similar to the EAV data model; additionally, it retains patient privacy and together with the dictionary provides a simplified, yet flexible and powerful, platform for the complex querying and analysis of patient information and disease pathways. It enables the summarisation and extension of pathways as well as the aggregation of similar sequences. It is also possible to capture and plot pathways with concurrent elements and to develop algorithms to further explore the data and investigate quality issues. Furthermore the methodology used to build pathway dictionary as well as the formalisms presented here can be transported to

other domains and settings.

The process of integrating routinely collected electronic data may produce pathways that may not be informative or complete. A topic which has received little attention in the literature is the computation of quality indicators for data-driven pathways. Such indicators are important to enable the selection of study-relevant high quality data for clinical investigation. The methods developed in this chapter enable us to discard pathways, that because of the nature of electronically routinely collected hospital data as well as retrieval methods, fail to provide enough or sufficiently accurate information to be used in clinical analysis.

It has been shown that methods for pathway quality measurement can rely on biological marker trends, as they are often the response to some parallel process. In the case of the PSA, a sharp decline in the average readings would indicate treatment to the prostate, which suppressed the production of the antigen, so can be use to ascertain if treatment records are missing. Algorithms were written to compute completeness based on prostate cancer biomarker rules creating an overall scoring system (Table 4.18). Once researchers are satisfied that the PSA trends have sufficient data points and are substantiated (i.e. they receive a high completeness score), they can investigate those PSA trends as predictors of prognosis in the disease. Such research is seldom undertaken due to the unavailability of data but may lead to improved outcomes for patients and health services.

We investigated the cohort of 1,904 patients, automatically built their respective pathways, and computed completeness scores. Overall 65.2% of pathways attained the two highest scores, while 82.9% attained the highest PSA positioning rule. Hence, these pathways contain sufficient biomarker information to aid clinical investigations on the biomarker trends. It was shown that routinely collected data can be transformed and prepared for clinical research, decision making and decision support but limitations still exist. Routine hospital data is also heterogeneous when it comes to quality and not all required information that would

accurately explain a pathway is available using electronic health records alone. Data paucity refers to the instances when there is not enough information in the archives that satisfactorily describe a phenomenon or an occurrence [257]. As seen in the case study in prostate cancer, events have start dates but rarely have end dates, which makes it difficult to understand their timeliness and requires additional consultation with the domain experts. It is hoped and recommended that, in the future, hospital information systems collect more details and metadata as this would greatly improve the secondary uses of routine clinical data.

Nevertheless, the flexibility of the data structure presented in this chapter allows the insertion and removal of new dictionary elements including additional blood tests and comorbidities to the pathways, as depicted in Figures 4.13 and 4.16. It is also possible to extend the data structure to include end-times for every activity. The work presented here has also enabled future research into published pathway adherence and variance metrics, particularly with respect to the UK NICE guidelines. This chapter described methods for data collection, presentation and quality assessment that can be applicable to build other disease pathways in other settings. Further work is also envisaged on mining pathways, in particular, the computation of similarity of biomarker trends, and the application of clustering algorithms and survival analysis in the context of pathways.

In summary, this chapter described the following contributions:

- A data model that describes patients' system-level paths as the sequences of *digital footprints* they leave in hospital systems;
- The potentials and limitations of data and process mining techniques in analysing system-level paths and the trends found in this data;
- The flexible Pathways data model and structure, representing the journeys that patients take through care, allowing for temporal abstraction, aggregation, and summarisation of similar sequential pathways;
- A methodology to construct the pathways data dictionary from routine hospital data collected from an operational data store;
- A novel graphical representation of a patient-centric pathway using routinely collected hospital data;
- A framework and software for the visualisation and analysis of complex clinical information and pathways;
- A method of assessing the quality dimensions of routine hospital data.

Chapter 5

Conclusions

Routine clinical data is recorded on several different hospital systems but is rarely analysed or linked. Upon linkage, pre-processing and integration of this data, important and useful information can be obtained that benefits patients, their doctors and health service planners. This thesis explored in detail the journey from routinely collected hospital data to knowledge in a large NHS hospital with a catchment area of up to 822,500 people.

Methodologies and computational techniques were applied and developed, and the importance of clean integrated health records was demonstrated using two clinical case studies. The prostate cancer case study allowed us to examine over 10 years of hospital data and their systems. On average, prostate cancer patients' pathways spanned over a period of 7.6 years (SD 4.6) from their first recorded biomarker test. This contrasts with the stroke case study, where the acute episodic nature of the data in hospital showed an average length of stay of 14.4 days (SD 20.1). Both case studies complemented each other in this respect, yet collecting and linking prostate cancer data was more difficult due to the sheer number of systems as well as its lengthy time span. In contrast, the stroke register provided a centralised way of selecting cohorts; however, the fact that there were

three different versions of the database posed additional integration challenges due to their semantic differences.

Overall, 15 different data sources were explored, including the regional cancer registry and the NHS national tracing service (now Personal Demographics Service). This allowed for semantic differences as well as different types of heterogeneity between sources and their data to be investigated. It is crucial to understand the data, its semantics, changes over time, and its origins; the latter requires a thorough understanding of the system, that is also bound to change over time and which may or may not interchange data with other systems. Interoperability is a much sought after goal in hospitals yet this thesis has demonstrated that it still has not been achieved in a way that facilitates the linkage and data interchange across hospital systems and with national sources. This is confirmed in recent publications in other settings both nationally and internationally. The time and effort spent on understanding, linking and preparing routine data will continue to hinder research projects and prevent new research from being carried out. To facilitate the process of retrieving and creating study databases, a methodology was introduced in this thesis and applied successfully to the two case studies.

However, the journey from data to knowledge is far more encompassing than data retrieval and integration; it relies on quality assessment and cleaning, on additional semantic interpretation and contextualisation, and on preprocessing techniques to extract relevant features from unstructured data. Data modeling thus becomes a subsequent critical step to organise such information and to carry out analyses. This thesis introduced techniques for data preprocessing as well as modeling complex routine clinical information that enabled further research to be undertaken. A detailed summary of the journey from data to knowledge is provided in the next section.

5.1 The Journey from Data to Knowledge

i. **Access**

It has been and continues to be difficult and bureaucratic to obtain appropriate credentials to carry out research using hospital data in the NHS. Hospital database projects wanting to use anonymised patient data undergo the same ethics process as large drug trial projects from pharmaceutical companies. In addition, as a primary investigator for the prostate cancer study, little guidance was available at the initial stages and support from the clinical supervisor proved essential. The way in which clinicians and other staff request hospital data is idiosyncratic and unclear; in some cases junior doctors manually collect the data while, in other cases, the system's administrators or the information services department can provide some help in achieving this. However, some of the data available in hospital systems, and its quality, is unknown even to the system administrators. This is particularly true for data that is outside the scope of planning, performance or audits. As a result, investigating other available data with research potential, and negotiating system access, becomes a laborious process, that is repeated as many times as there are systems and managers. Because this data is not used, there is limited accountability and a lack of understanding, which naturally hinders the process of investigating its origins and its potential research uses. Methods for evaluating the quality, completeness and plausibility of routine data were explored in this thesis.

In this project, ethics approval was obtained for the prostate cancer study and patient identifiable information such as names and addresses was not used. The stroke case study was also approved by the local ethics and governance committees which granted full access to the data for the purpose of building the stroke register and linking information, yet only anomymised datasets could be extracted and used for research. Access is an essential part of

working with hospital data, and requires substantial amounts of time and effort, in particular when the extent and availability of the data is not known. Accessing data was accounted for in the proposed methodology, and was discussed in Chapter 2.

ii. **Data Retrieval**

The collection of patient-centric data from multiple hospital sources is a non-trivial task, and one that has received relatively little attention in the literature. Chapter 2 introduced a novel methodology for the retrieval of patient-centric data from multiple sources. The methodology can be used on its own or built into existing data mining or data analysis methodologies; its outputs are a study dataset for each source investigated, and a metadatabase that contains system and data details. This leads to the creation of an operational data store (ODS), a semi-integrated database that replicates the host environment and allows linkage, quality assurance and selection of cohorts. Furthermore, the methodology allows the study of the systems and their host environment; it highlights the changes in systems over time, and leads to a preliminary selection of cohorts.

Selecting an appropriate prostate cancer cohort was a difficult task as there was no single method or system to select patients diagnosed with the condition. This led to recursively backtracking between the ODS, the systems and the production of summary statistics for discussion with the clinical and research teams. The methodology proved useful in streamlining this process. Nevertheless, improvements in the way the hospital records information were noted during the 10-year study period. In particular, the national cancer waiting times audit required that hospitals account for the number of patients diagnosed with cancer, their waiting times and treatment modalities. This particular national initiative has had an extremely positive effect in the way information is recorded electronically in this particular hospital. Nev-

ertheless, information that is outside the scope of national audits or beyond the mandatory datasets continues to be "neglected". This is, arguably, an accountability issue that, despite not being a priority, should receive further attention and resources.

iii. **Preprocessing**

Preparing and cleaning the retrieved data as well as dealing with its "conceptual, organizational, logistical, managerial, and statistical-epidemiological aspects" [7] is essential so that any analyses can be carried out. However, this has not been studied comprehensively, and descriptions of preprocessing methods are lacking in the literature. Chapter 3 introduced selected techniques demonstrating the most relevant challenges in this respect. In particular, a technique to extract information from histopathology text reports provided interesting results, and enabled further study on the way histopathologists report their findings. Beyond histopathology, clinical reports continue to be written and stored with relatively little structure, and improvements are needed in order entry and natural language processing software so as to help authors assign conceptual mappings, keywords or relevant clinical coding. This structure would make the process of information retrieval more accurate and efficient rather than cumbersome and *ad hoc* as reported in this thesis.

Data editing and imputation techniques were also investigated in Chapter 3, and an algorithm for continuous value estimation was introduced. This was an important and relevant exercise as only some databases verify their information at the time of imputation. For instance, information resulting from a blood test for a patient that is not registered (or properly linked) in a hospital administration system will contain values that were entered at the time of consultation, but a percentage of these can be erroneous due to bias. For such cases, a look-up service using NHS number (to the Personal Demo-

graphics Service) could automatically return the patient's accurate identity and demographic information, which would ensure the quality of this data, even when patients are not registered with the hospital. Preprocessing tools and techniques will continue to be required not only to assess erroneous values or transform data and formats (for example, anonymisation) but also to accurately link information across databases.

iv. **Linkage and Integration**

Data linkage can be considered a preprocessing technique in that it is allowing distinct datasets to be merged before analyses can be undertaken. Two data linkage exercises were included in Chapter 3 and demonstrated the value of a technique that is often overlooked in intra-organisational exercises. One of the exercises assessed the reliability and accuracy of hospital databases in providing a correct date of death. The exercise compared the dates of death recorded for stroke patients in the administration system with data from the NHS national tracing service over a period of 15 years. A yearly average of 10% (SD 3.6) of mismatches (5.6% were missing and 4.5% were recorded incorrectly) with a highly varying distribution over the years was observed. Changes to hospital services, data quality processes and national systems (such as the introduction of the Summary Care Record) are likely to have contributed to the volatile distribution of the mismatches over time. Nevertheless, this illustrates that secondary care databases need to be used with caution when retrieving this type of information and, where possible, linkage to national systems should be sought. This exercise has also shown that even deterministic record linkage within a hospital is prone to error (2.2%) despite the use of unique identifiers.

The second linkage exercise aimed at providing additional biochemistry information to the stroke register, and provided new insights into the availability of electronic biochemistry records for stroke admissions and discharges. This

is important as the amount of patients having a particular blood test was previously not known, and cohorts including this information can now be selected and used. This exercise provided groundwork for the development of a new clinical data warehouse, the Norfolk & Norwich Research Stroke & TIA Register (stroke register) with over 25 thousand patients.

The linkage exercises provided a level of data integration across hospital databases and with national sources. However, the context in which information was created in each source required additional semantic integration, and the data collection methodology facilitated this. Nevertheless, further work on integration continued to be carried out throughout this thesis. Data interchange across hospital and other sources could be improved by the use of international standards such as the health level-7 (HL7); however, it was not possible to use any such standards with the systems studied in this thesis. Arguably, HL7 and other standards may be themselves complex to implement, do not provide a full integration between different coded values, and are difficult to reach without privileged access to the underlying information system. Perhaps simpler implementations and data models would provide better results and wider applicability.

v. **Data modeling**

Chapter 3 introduced the concepts of data warehousing and data modeling in the context of developing the stroke register clinical data warehouse. The latter was developed using a hybrid approach that took Inmon's relational model and technical focus, Kimball's methodological principles and bottom-up alignment, and Szirbik's overarching methodological steps. We found this approach to provide sound results in smaller environments within a large organisation.

Based on the above, a methodology for the development of domain-specific clinical data warehouses with routinely collected data was introduced and

addresses all the challenges identified when producing the three versions of the stroke register. The methodology is expected to have wider applicability and a research paper is currently being prepared on this topic.

In Chapter 4, a new data model, inspired by the EAV model and RDF triples, was introduced to describe the routes that patients take through care using routine data. The accompanying methodology made this model adaptable to new data elements and to changes in the host environment. The data model made possible the integration and organisation of patient-centric data and information for multiple purposes. The pathways model was extended to include complex clinical information and to include its own data dictionary, providing a level of ontological alignment and a description of the semantics of patient activities. This work was carried out to model prostate cancer pathways but it is expected to have wider applicability and reproducibility; work is currently underway to develop stroke pathways, and a paper has been submitted to a leading health informatics journal.

Overall, data modeling was shown to be an extremely important part of the journey, particularly in contextualising information on patient activities as well as providing new knowledge about the way in which activities relate to one another. Furthermore, the simplified data structure that is platform independent provided a straightforward way to manage the ever-changing data and its model. Nevertheless, models similar to the EAV are less efficient and require additional programming to perform tasks that other architectures would do automatically. Integrated platforms that facilitate the interactions between data model, system and users, could tackle these drawbacks.

vi. **Visualisation, Quality and Analyses**

Chapter 4 introduced a framework for the integration, visualisation and analysis of the prostate cancer pathways that could be applied to other domains. The framework included the CaP VIS software and an analysis engine that

enabled complex interactions with the pathways data model and its data. Other work on frameworks with similar functions was published while this research was ongoing [177; 178; 258]. However, at the time of writing, these do not focus on pathways nor, to our knowledge, do they compute visualisations or the analyses that were carried out in this thesis.

In this thesis, patient-centric pathways were created for each prostate cancer patient in the cohort and a novel graphical representation was introduced. The compilation of several data points in a single patient-centric pathway plot revealed aspects of the patient's journey through care that would otherwise remain hidden in databases. Furthermore, the framework enables the summarisation and extension of pathways as well as the aggregation of similar sequences. In turn, this allows the inspection of frequent paths that patients take, and the data can be remodeled to focus on particular activities or events. The framework and its software can be seen as a clinical and research decision support system that enables the inspection of complex clinical information, its analysis, and the selection of high quality cohorts for research. Software was also developed to compare the outcomes of given patients' profiles using the prostate cancer data.

Data and process mining techniques are often suggested for the analysis of workflows and pathways. These techniques were evaluated in this thesis and found unsuitable when applied to unstructured routine clinical data. The framework proposed in this thesis allows for quality data to be selected, and has the potential to facilitate data and process mining not only by providing more structured data, but also by enabling further interpretation of complex clinical data. Further work is required both on algorithms that can cope with the complex nature of routine data, as well as software that can integrate and prepare data for analyses.

The process of integrating routinely collected electronic data also produces pathways that may not be informative or complete. Chapter 4 also discussed

the additional data quality dimensions that are possible to explore thanks to the proposed framework. In particular, a scoring system was developed to rank pathways according to their quality in order to study the prostate cancer biomarker. Rules can be implemented to ascertain, for example, whether a treatment has indeed occurred at a particular time point based on the effect it had on the biomarker readings.

The overall quality of routinely collected electronic data, based on the observations made in this thesis, vastly depends on the uses and importance of that data (or system) to the clinicians and hospital. Quality data is often found in systems used for audits, commissioning and performance, or automated order entry systems such as the one used in biochemistry. Conversely, erroneous data, although limited, is mostly found in systems where backdated manual entry is performed, in systems where data points may not be consistently relevant to a particular disease, or in systems where this data is no longer used operationally. In such cases patient notes are still the most accurate source of detailed information yet this may not be recorded electronically. Nevertheless the work carried out in this thesis should reassure that de-identified patient data can be used for research and public good, and that, by creating a greater demand for this data, their quality and accuracy are improved.

vii. **Further Research and Ongoing Clinical Studies**

The work carried out in this thesis has already enabled clinical studies and further research on stroke and prostate cancer. Due to the cumbersome journey from data to knowledge, time constraints and other limitations, these are not included or discussed in detail in this thesis. Nevertheless the studies that have been carried out and submitted for publication as well as those still under preparation have been included as abstracts in Appendix E. Summary statistics of the findings in prostate cancer are given in Appendix D. Further-

more, data mining techniques such as the decision trees have been applied to the prostate cancer data (with a comprehensive list of biochemistry tests as potential predictors) and to the stroke dataset, and further work is currently underway to interpret the results. In addition, the work presented here has also enabled future research into published pathway adherence and variance metrics, particularly with respect to the UK NICE guidelines. For example, rules derived from the guidelines can be checked with the framework presented in this thesis, in a similar fashion to the work carried out in assessing the data quality dimensions.

The above steps map the comprehensive journey from data to knowledge in hospitals: access, retrieval, preprocessing, linkage and integration, data modelling, visualisation, quality and analyses. The steps were carried out mostly in this order, however, significant backtracking from the last steps to the first four was noted. These first four steps (access, retrieval, cleansing and linkage) were also the most time consuming, followed by data modelling and the development of the framework, which required a substantial amount of consultation with the clinical teams. One of the limitations of this work is that it studied a single large hospital. However, given the reports in the literature, the consultation with the clinical teams, and the two case studies, the steps and challenges are expected to be similar elsewhere and in most other *heterogeneous* domains.

5.2 Concluding Remarks and Further Work

This thesis has reported in detail the cumbersome journey from data to knowledge in secondary health care centres and, as a result, provided unique insights at every step of the way. The question of whether routinely collected data in large hospital databases can be used for research, was answered in that, with limitations, some of the information can indeed be used for this purpose. One of the reasons for these limitations has to do with the fact that, in hospitals, there is limited accountability for data that is not actively used for audits, management, or planning and performance evaluation. Nevertheless, we have seen that it is still possible to extract knowledge from large heterogeneous databases in hospitals and that the extracted knowledge be used for research. We have also seen that using additional data sources is beneficial in this respect as it validates and complements hospital data.

Because of some of the work presented in this thesis, the hospital R&D department has decided to invest in this area, and further work will continue in the development of a sustainable stroke register infrastructure. This work has also resulted in plans to expand and apply the pathways concept to other hospitals and domains and a bid has also been sent for the continuation of this study. Further work is needed to develop analytical tools that can cope with the unique nature and heterogeneity of routine health data.

This project has encouraged further work in other problem domains and in the development of tools to retrieve, model and analyse pathways from multiple sources. It is becoming increasingly evident that, with the explosion of *big data* in hospitals and in society, such tools will play an increasingly important role in the computerised analysis of trends and knowledge discovery. The extraction of knowledge from data is the goal of data science, and this emerging discipline shares features with the work carried out in this thesis, where a broad range of skills, from

analytical to “soft skills”, is required of its scientists and practitioners.

“For all its perils, medical data mining can also be the most rewarding. For an appropriately formulated scientific question, thousands of data-elements can be brought to bear on finding a solution. For an appropriately formulated medical question, finding an answer could mean extending a life, or giving comfort to an ill person. These potential rewards more than compensate for the many extraordinary difficulties along the pathway to success”, Cios and Moore [20].

References

[1] L.H. Sobin, M.K. Gospodarowicz, and C. Wittekind. *TNM Classification of Malignant Tumours*. Wiley, 2009. viii, 76, 78, 82, 90

[2] M. Breslin. Data warehousing battle of the giants: Comparing the basics of the kimball and inmon models. *Business Intelligence Journal*, 9(1):6–20, 2004. ix, 145, 150, 151, 156, 157, 158, 159, 176

[3] N.B. Szirbik, C. Pelletier, and T. Chaussalet. Six methodological steps to build medical data warehouses for research. *International Journal of Medical Informatics*, 75(9):683 – 691, 2006. International Council on Medical and Care Compunetics (ICMCC). ix, 159, 160, 163, 165, 167, 176

[4] W. Hersh. A stimulus to define informatics and health information technology. *BMC Medical Informatics and Decision Making*, 9(1):24, 2009. xi, 4, 5

[5] J.C. Krzysztof. *Medical Data Mining and Knowledge Discovery*. Physica-Verlag Heidelberg, 2001. xi, 21, 22, 23, 24, 26, 29, 42, 50, 51, 69

[6] O. Brazhnik and J.F. Jones. Anatomy of data integration. *Journal of Biomedical Informatics*, 40(3):252–269, 2007. xi, 21, 22, 30, 31, 34, 43, 51, 151, 235

[7] J. Van den Broeck, S.A. Cunningham, R. Eeckels, and K. Herbst. Data

cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, 2(10):e267, 09 2005. xi, 69, 70, 284

[8] D.F. Gleason. *The Veteran's Administration Cooperative Urologic Research Group: histologic grading and clinical staging of prostatic carcinoma. Urologic Pathology: The Prostate*, pages 171–198. Lea and Febiger, 1977. xi, 76

[9] C.D.J. Holman. Introductory analysis of linked health data: Principles and hands-on applications version 1.6. Technical report, Scottish Health Informatics Programme Training Workshop, St. Andrews, UK, 2009. xii, 117, 118, 121, 122, 144

[10] B. Feldman, E.M. Martin, and T. Skotnes. Big data in healthcare hype and hope. Technical report, DrBonnie360, 2012. 2

[11] D.S. Jones, S.H. Podolsky, and J.A. Greene. The burden of disease and the changing task of medicine. *New England Journal of Medicine*, 366(25):2333–2338, 2012. PMID: 22716973. 2, 3

[12] T. Ovitt, M.P. Capp, P. Christenson, H.D. Fisher, M.M. Frost, S. Nudelman, H. Roehrig, and G. Seeley. Development of a digital video subtraction system for intravenous angiography. In *23rd Annual Technical Symposium*, pages 73–76. International Society for Optics and Photonics, 1979. 3

[13] E.H. Shortliffe, R. Davis, Stanton S.G. Axline, B.G. Buchanan C.C. Green, and S.N. Cohen. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the mycin system. *Computers and biomedical research*, 8(4):303–320, 1975. 3

[14] R. Miller, H.E. Pople, and J.D. Myers. Internist-i, an experimental computer-based diagnostic consultant for general internal medicine. In J.A.

Reggia and T. Stanley, editors, *Computer-assisted medical decision making*, volume 2. Springer-Verlag, New York, 1986. 3

[15] Hospital episode statistics, admitted patient care, england - 2012-13. Technical report, Health and Social Care Information Centre, 2013. 3, 15

[16] G.M. Weber, K.D. Mandl, and I.S. Kohane. Finding the missing link for big biomedical data. *The Journal of the American Medical Association*, 311(24):2479–2480, 2014. 3, 4

[17] T. Cairns, H. Timimi, M. Thick, and G. Gold. A generic model of clinical practice - the cosmos project. In Klaus-Peter Adlassnig, Georg Grabner, Stellan Bengtsson, and Rolf Hansen, editors, *Medical Informatics Europe 1991*, volume 45 of *Lecture Notes in Medical Informatics*, pages 706–710. Springer Berlin Heidelberg, 1991. 3

[18] P. Taylor. *From Patient Data to Medical Knowledge: The Principles and Practice of Health Informatics*. Wiley, 2008. 3, 10, 70

[19] Academy of Medical Sciences. Personal data for public good: using health information in medical research. Technical report, London: AMS, 2006. 3, 20

[20] K.J. Cios and W. Moore. Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26(1-2):1–24, 2002. 4, 10, 20, 23, 24, 69, 72, 100, 101, 292

[21] W. Sujansky. Heterogeneous database integration in biomedicine. *Journal of Biomedical Informatics*, 34(4):285–298, 2001. 4, 9, 10, 21, 30, 153, 155, 189

[22] P.M. Nadkarni, L. Marenco, R. Chen, E. Skoufos, G. Shepherd, and P. Miller. Organization of heterogeneous scientific data using the eav/cr

representation. *Journal of the American Medical Informatics Association*, 6(6):478–493, 1999. 4, 153

[23] C.P Friedman. What informatics is and isn't. *Journal of the American Medical Informatics Association*, 20(2):224–6, 2012. 4

[24] R.A. Greenes and E.H. Shortliffe. Medical informatics: An emerging academic discipline and institutional priority. *The Journal of the American Medical Association*, 263(8):1114–1120, 1990. 5

[25] P.L. Miller. Medical informatics in clinical medicine and the biosciences. *Nature medicine*, 1(1):93, 1995. 5

[26] E. Coiera. *Guide to Health Informatics, 2Ed.* Taylor & Francis, 2003. 5

[27] Julie Louise Gerberding. Health-care quality promotion through infection prevention: beyond 2000. *Emerging infectious diseases*, 7(2):363, 2001. 6

[28] L. Garcia Alvarez, P. Aylin, J. Tian, C. King, M. Catchpole, S. Hassall, K. Whittaker-Axon, and A. Holmes. *Journal of Hospital Infection*, 79(3):231 – 235, 2011. 6

[29] D. Teodoro, E. Pasche, J. Gobeill, S. Emonet, P. Ruch, and C. Lovis. Building a transnational biosurveillance network using semantic web technologies: requirements, design, and preliminary evaluation. *Journal of medical Internet research*, 14(3):e73–e73, 2011. 6

[30] E.O. Wilson. *Consilience: the unity of knowledge*. Vintage Series. Abacus, 1999. 6

[31] C.P Friedman. Where's the science in medical informatics. *Journal of the American Medical Informatics Association*, 2(1):65, 1995. 6

[32] S. Gonzalez-Bailon. Social science in the era of big data. *Policy & Internet*, 5(2):147–160, 2013. 7

[33] C. Hsinchun, S.F. Sherrilynne, C. Friedman, and W. Hersh. *Medical Informatics: Knowledge Management and Data Mining in Biomedicine (Integrated Series in Information Systems)*. Springer, June 2005. 7

[34] J.D.S. Kay. Communicating with clinicians. *Annals of clinical biochemistry*, 38(2):103–110, 2001. 8, 276

[35] R. Haux. Health information systems: past, present, future. *International Journal of Medical Informatics*, 75(3-4):268–281, 2006. 8, 9, 20

[36] M.H. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2002. 10, 24, 191, 204

[37] N.G. Weiskopf and C. Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013. 11, 189, 274

[38] L.T. Kohn, J.M. Corrigan, M.S. Donaldson, S. Molla, et al. *To Err Is Human: Building a Safer Health System*, volume 627. National Academies Press, 2000. 11

[39] A. Bottle B. Jarman P. Aylin, S. Tanna. How often are adverse events reported in english hospital statistics? *BMJ*, 329(7462):369, 2004. 11

[40] B.M. Hales and P.J. Pronovost. The checklist-a tool for error management and performance improvement. *Journal of Critical Care*, 21(3):231 – 235, 2006. 11

[41] R. Kaushal, K.G. Shojania, and D.W. Bates. Effects of computerized physician order entry and clinical decision support systems on medication safety: A systematic review. *Archives of Internal Medicine*, 163(12):1409–1416, 2003. 11

[42] Cancer Research UK. Prostate cancer statistics, November 2012. 13, 192

[43] S.M. Collin, R. Martin, C. Metcalfe, D. Gunnell, P.C. Albertsen, D. Neal, F. Hamdy, P. Stephens, J. Lane, R. Moore, et al. Prostate-cancer mortality in the usa and uk in 1975–2004: an ecological study. *The lancet oncology*, 9(5):445–452, 2008. 13

[44] C.G. Roehrborn and L.K. Black. The economic burden of prostate cancer. *BJU International*, 108(6):806–813, 2011. 13

[45] B. Holmström, M. Johansson, A. Bergh, U. Stenman, G. Hallmans, and P. Stattin. Prostate specific antigen for early detection of prostate cancer: longitudinal study. *BMJ*, 339, 9 2009. 13

[46] Department of health: Progress in improving stroke care. Technical report, National Audit Office, 2010. 14

[47] P.M. Rothwell, A.J. Coull, L.E. Silver, J.F. Fairhead, M.F. Giles, C.E. Lovelock, J.N.E. Redgrave, L.M. Bull, S.J.V. Welch, F.C. Cuthbertson, L.E. Binney, S.A. Gutnikov, P. Anslow, A.P. Banning, D. Mant, and Z. Mehta. Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (oxford vascular study). *The Lancet*, 366(9499):1773 – 1783, 2005. 14

[48] L. Gray, N. Sprigg, P.M.W. Bath, P. Sørensen, E. Lindenstrøm, G. Boysen, P.P. De Deyn, P. Friis, D. Leys, R. Marttila, et al. Significant variation in mortality and functional outcome after acute ischaemic stroke between western countries: data from the tinzaparin in acute ischaemic stroke trial (taist). *Journal of Neurology, Neurosurgery & Psychiatry*, 77(3):327–333, 2006. 14

[49] National sentinel stroke clinical audit 2010. Technical report, Royal College of Physicians, 2011. 14

[50] D. O'Neill, F. Horgan, A. Hickey, and H. McGee. Stroke is a chronic disease with acute events. *BMJ*, 336(7642):461–461, 2008. 14

[51] P.K. Myint and A.A. Welch. Healthier ageing. *BMJ*, 344:e1214, 2012. 14

[52] J. Bettencourt-Silva, B. De La Iglesia, S. Donell, and V. Rayward-Smith. On creating a patient-centric database from multiple hospital information systems in a national health service secondary care setting. *Methods of Information in Medicine*, 51(3):210–20, 2012. 15, 233

[53] S. Pakhomov, S.A. Weston, S.J. Jacobsen, C.G. Chute, R. Meverden, and V.L. Roger. Electronic medical records for clinical research: application to the identification of heart failure. *American Journal of Managed Care*, 13(6):281–288, 2007. 20

[54] J. Powell and I. Buchan. Electronic health records should support clinical research. *Journal of Medical Internet Research*, 7(1):88–93, 2005. 20

[55] M. Berg and E. Goorman. The contextual nature of medical information. *International Journal of Medical Informatics*, 56(1-3):51–60, 1999. 20, 32

[56] H.T. Sorensen, S. Sabroe, and J. Olsen. A framework for evaluation of secondary data sources for epidemiological research. *International Journal of Epidemiology*, 25(2):435–442, 1996. 20, 32, 43, 47, 48

[57] P.L. Reichertz. Hospital information systems - past, present, future. *International Journal of Medical Informatics*, 75(3-4):282–299, 2006. 20

[58] C. Safran and L.E. Perreault. *Medical Informatics: Computer Applications in Health Care and Biomedicine*, chapter Management of Information in Integrated Delivery Networks, pages 359–396. Springer, 2003. 20

[59] M.L. Muller, T. Ganslandt, H.P. Eich, K. Lang, C. Ohmann, and H. Prokosch. Towards integration of clinical decision support in commercial hospital information systems using distributed, reusable software and

knowledge components. *International Journal of Medical Informatics*, 64(2-3):369–377, 2001. 21

[60] D.M. Mackay, C. Papi, N. Roberts, and N. Bexon. Ten ways to improve information technology in the nhs. primary care doctors need to become aware of training opportunities. *British Medical Journal*, 326:1034, 2003. 21

[61] J.D. Kay, D. Nurse, C. Bountis, and K. Paddon. The oxford clinical intranet: providing clinicians with access to patient records and multiple knowledge bases with internet technology. *Studies in Health Technology and Informatics*, 100:130–138, 2004. 21

[62] J.M. Teich, J.P. Glaser, R.F. Beckley, M. Aranow, D.W. Bates, G.J. Kuperman, M.E. Ward, and C.D. Spurr. The brigham integrated computing system (bics): advanced clinical systems in an academic hospital environment. *International Journal of Medical Informatics*, 54(3):197 – 208, 1999. 21

[63] N. de Keizer and E. Ammenwerth. The quality of evidence in health informatics: How did the quality of healthcare it evaluation publications develop from 1982 to 2005? *International Journal of Medical Informatics*, 77(1):41–49, 2008. 21, 69, 70

[64] J.C.W. Debuse, B. de la Iglesia, C.M. Howard, and V.J. Rayward-Smith. *Industrial Knowledge Management*, chapter Building the KDD Roadmap: A Methodology for Knowledge Discovery, pages 179–196. Springer-Verlag, 2000. 21, 22, 29, 42, 48, 49, 51

[65] C. Shearer. The crisp-dm model: the new blueprint for data mining. *Journal of Data Warehousing*, 5:13–22, 2000. 21, 22, 26, 28, 29

[66] J.C. Krzysztof and A.K. Lukasz. Trends in data mining and knowledge discovery. In *Pal N.R., Jain, L.C. and Teodoresku, N. (Eds.), Knowledge Discovery in Advanced Information Systems*, pages 200–2. Springer, 2002. 22, 24, 25, 42

[67] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in knowledge discovery and data mining*. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996. 22, 27, 29, 68

[68] P. Cabena. *Discovering Data Mining: from concept to implementation*. Prentice Hall International, 1997. 22, 26, 29

[69] K.J. Cios, A. Teresinska, S. Konieczna, J. Potocka, and S. Sharma. A knowledge discovery approach to diagnosing myocardial perfusion. *Engineering in Medicine and Biology Magazine, IEEE*, 19(4):17–25, 2000. 23, 26

[70] J.P Sacha, K.J. Cios, and L.S. Goodenday. Issues in automating cardiac spect diagnosis. *Engineering in Medicine and Biology Magazine, IEEE*, 19(4):78–88, 2000. 23, 26

[71] L.A. Kurgan, K.J. Cios, R. Tadeusiewicz, M. Ogiela, and L.S. Goodenday. Knowledge discovery approach to automated cardiac spect diagnosis. *Artificial Intelligence in Medicine*, 23(2):149–69, 2001. 23, 26

[72] G. Richards, V.J. Rayward-Smith, P.H. Sonksen, S. Carey, and C. Weng. Data mining for indicators of early mortality in a database of clinical records. *Artificial Intelligence in Medicine*, 22(3):215 – 231, 2001. 24, 25

[73] J.H. Bettencourt. Extracting patient-centric data from the nhs: A case study in prostate cancer at the norfolk & norwich university hospital. Mas-

ter's thesis, School of Computing Sciences, University of East Anglia, Norwich, 2009. 26, 46

[74] K.K. Hirji. Exploring data mining implementation. *Communications of the ACM*, 44(7):87–93, July 2001. 27

[75] Lukasz L.A. Kurgan and P. Musilek. A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(1):1–24, March 2006. 28

[76] W. Cheung and C. Hsu. The model-assisted global query system for multiple databases in distributed enterprises. *ACM Transactions on Information Systems*, 14(4):421–470, 1996. 30, 32

[77] A.P. Sheth and J.A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22(3):183–236, 1990. 30, 31, 32

[78] C. Hsu, M. Bouziane, L. Rattner, and L. Yee. Information resources management in heterogeneous, distributed environments: A metadatabase approach. *IEEE Transactions on Software Engineering*, 17(6):604–625, 1991. 32

[79] NHS Connecting for Health. Nhs data model and dictionary, 2009. 37

[80] A. Gupta and M.S. Lam. Estimating missing values using neural networks. *Journal of the Operational Research Society*, 47:229–238, 1996. 47, 49, 100

[81] K. Thangavel and A. Pethalakshmi. Dimensionality reduction based on rough set theory: A review. *Applied Soft Computing*, 9:1–12, 2008. 49

[82] E. Rahm and P.A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, December 2001. 52

[83] Norwich Research Park Cardiovascular Research Group. Protocol for norfolk and norwich stroke & tia register, version 3, 2013. 53, 54

[84] N. Gange and S. Marroqui. Trust guideline for the management of acute stroke in adults. Technical report, Norfolk & Norwich University Hospital NHS Foundation Trust, Norwich, UK, 2011. 54

[85] L. Garcia Alvarez, P. Aylin, J. Tian, C. King, M. Catchpole, S. Hassall, K. Whittaker-Axon, and A. Holmes. Data linkage between existing healthcare databases to support hospital epidemiology. *Journal of Hospital Infection*, 79(3):231 – 235, 2011. 65, 66

[86] A. Famili, W. Shen, R. Weber, and E. Simoudis. Data preprocessing and intelligent data analysis. *Intelligent Data Analysis*, 1(1-4):3 – 23, 1997. 68, 69

[87] D. Pyle. *Data Preparation for Data Mining*. The Morgan Kaufmann Series in Data Management Systems Series. Morgan Kaufmann Publishers, 1999. 69

[88] T. de Waal, J. Pannekoek, and S. Scholtus. *Handbook of Statistical Data Editing and Imputation*, pages 429–439. John Wiley & Sons, Inc., 2011. 69, 100, 101, 115

[89] S.B. Kotsiantis, D. Kanellopoulos, and P.E. Pintelas. Data preprocessing for supervised leaning. *International Journal of Computer Science*, 1(2), 2006. 69

[90] Society for Clinical Data Management. Good clinical data management practices, version 3.0. Technical report, Society for Clinical Data Management, 2003. 70

[91] K. Thiru, A. Hassey, and F. Sullivan. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ (Clinical research ed.)*, 326(7398):1070, May 2003. 70

[92] C. Safran and L.E. Perreault. *Medical Informatics: Computer Applications in Health Care and Biomedicine*. Springer, 2003. 72, 100

[93] M.A. Hearst. Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 3–10, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics. 72

[94] A. Tan. Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, pages 65–70, 1999. 72

[95] R. Baeza-Yates R. and B. Ribeiro-Neto. *Modern information retrieval*. Addison-Wesley Longman, 1999. 73

[96] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107 – 117, 1998. Proceedings of the Seventh International World Wide Web Conference. 73

[97] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *The Journal of the ACM*, 46(5):604–632, September 1999. 73

[98] J.P. Erinjeri, D. Picus, F.W. Fred, D.A. Rubin, and P. Koppel. Development of a google-based search engine for data mining radiology reports. *Journal of Digital Imaging*, 22(4):348–356, 2009. 73

[99] J. Mostafa and W. Lam. Automatic classification using supervised learning in a medical document filtering application. *Information Processing & Management*, 36(3):415 – 444, 2000. 73

[100] I. McCowan, D. Moore, and M.J. Fry. Classification of cancer stage from free-text histology reports. *Proceedings of The Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1:5153–6, 2006. 73

[101] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, and J.F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, pages 128–144, 2008. 74

[102] M. Bates. Models of natural language understanding. *Proceedings of the National Academy of Sciences*, 92(22):9977–9982, 1995. 74, 98

[103] S. Meystre and P.J. Haug. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. *Journal of Biomedical Informatics*, 39(6):589 – 599, 2006. 74

[104] J.E.F. Friedl. *Mastering Regular Expressions*. Second edition, 2002. 74

[105] A. Turchin, N.S. Kolatkar, R.W. Grant, E.C. Makhni, M.L. Pendergrass, and J.S. Einbinder. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *Journal of the American Medical Informatics Association*, 13(6):691–695, 2006. 74

[106] H. Liu, Y.A. Lussier, and C. Friedman. Disambiguating ambiguous biomedical terms in biomedical narrative text: An unsupervised method. *Journal of Biomedical Informatics*, 34(4):249–261, August 2001. 75, 99

[107] G. Dhom. Histopathology of prostate carcinoma. diagnosis and differential diagnosis. *Pathology, research and practice*, 179(3):277–303, 1985. 75, 76, 77, 78

[108] P.A. Humphrey. Gleason grading and prognostic factors in carcinoma of the prostate. *Modern Pathology*, 17(3):292–306, 2004. 76, 77

[109] W. C Allsbrook, K.A. Mangold, M.H. Johnson, R.B. Lane, C.G. Lane, M.B. Amin, D.G. Bostwick, P.A. Humphrey, E.C. Jones, V.E. Reuter, I.A.S. Wael Sakr, P. Troncoso, T.M. Wheeler, and J.I. Epstein. Interobserver reproducibility of gleason grading of prostatic carcinoma: Urologic pathologists. *Pathology, research and practice*, 32(1):74–80, 2001. 77

[110] A.W. Partin, M.W. Kattan, E.N. Subong, P.C. Walsh, K.J. Wojno, J.E. Oesterling, P.T. Scardino PT, and J.D. Pearson. Combination of prostate-specific antigen, clinical stage, and gleason score to predict pathological stage of localized prostate cancer. a multi-institutional update. *The Journal of the American Medical Association*, 277(18):1445–51, 1997. 78

[111] A.W. Partin, L.A. Mangold, D.M. Lamm, P.C. Walsh, J.I. Epstein, and J.D. Pearson. Contemporary update of prostate cancer staging nomograms (partin tables) for the new millennium. *Urology*, 58(6):843–8, 2001. 78

[112] R. Ayyathurai, K. Ananthakrishnan, R. Rajasundaram, R.J. Knight, H. Toussi, and V. Srinivasan. Predictive ability of partin tables 2001 in a welsh population. *Urologia internationalis*, 76(3):217–22, 2006. 78

[113] N. Bhojani, S. Ahyai, M. Graefen, U. Capitanio, N. Suardi, S.F. Shariat, C. Jeldres, A. Erbersdobler, T. Schlomm, A. Haese, T. Steuber, H. Heinzer, F. Montorsi, H. Huland, and P.I. Karakiewicz. Partin tables cannot accurately predict the pathological stage at radical prostatectomy. *European journal of surgical oncology*, 35(2):123–8, 2009. 78

[114] D.M. Fanning, Y. Fan, J.M. Fitzpatrick, and R.W Watson. External validation of the 2007 and 2001 partin tables in irish prostate cancer patients. *Urologia internationalis*, 84(2):174–9, 2010. 78

[115] D.M Fanning, F. Yue, J.M. Fitzpatrick, and R.W. Watson. Novel predictive tools for irish radical prostatectomy pathological outcomes: development and validation. *Irish Journal of Medical Science*, 179(2):187–95, 2010. 78

[116] National Institute for Health and Clinical Excellence. Prostate cancer: full guideline (cg58). Technical report, NICE, 2008. 78

[117] A.M. Cohen and W.R. Hersh. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71, 2005. 83, 99

[118] H.J. Murff, F. FitzHenry, M.E. Matheny, N. Gentry, K.L. Kotter, K. Crimin, R.S. Dittus, A.K. Rosen, P.L. Elkin, S.H. Brown, and T. Speroff. Automated identification of postoperative complications within an electronic medical record using natural language processing. *Journal of the American Medical Association*, 306(8):848–855, 2006. 97, 98

[119] D. Shalvi and N. DeClaris. An unsupervised neural network approach to medical data mining techniques. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence*, volume 1, pages 171–176 vol.1, 1998. 100

[120] S. Nordbotten. Measuring the error of editing questonnaires in a census. *American Statistical Association Journal*, 55:364–369, 1955. 101

[121] S.R. Cole, H. Chu, and S. Greenland. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology*, 35(4):1074–1081, August 2006. 101

[122] V. Tolkki. Automatic editing and imputation. Technical report, Statistics Finland, 2009. 101, 102, 103

[123] L. Granquist. An overview of methods of evaluating data editing procedures. *Statistical Data Editing, Methods and Techniques*, 2:112–123, 1997. 102

[124] L. Granquist. *Improving the Traditional Editing Process*, pages 385–401. John Wiley & Sons, Inc., 1995. 102

[125] I.P. Fellegi and D. Holt. A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71(353):17–35, 1976. 103

[126] F.E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969. 103

[127] UN/ECE Data Editing Group. Glossary of terms used in statistical data editing. Technical report, United Nations Statistical Commission and Economic Commission for Europe, 2000. 103

[128] P. Christen. Probabilistic data generation for deduplication and data linkage. In *IDEAL'05, Springer LNCS 3578*, pages 109–116. Springer LNCS, 2005. 117, 119

[129] G.R. Howe. Use of computerized record linkage in cohort studies. *Epidemiologic Reviews*, 20(1):112–121, 1998. 119

[130] H.B. Newcombe and J.M. Kennedy. Record linkage: making maximum use of the discriminating power of identifying information. *Communications of the ACM*, 5(11):563–566, November 1962. 119

[131] I. Fellegi, P. Ivan P., and A.B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. 119

[132] W.E. Winkler. Frequency-based matching in the fellegi-sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 778–783, 1989. 119

[133] C.W. Kelman, A.J. Bass, and C.D.J. Holman. Research use of linked health data - a best practice protocol. *Australian and New Zealand Journal of Public Health*, 26(3):251–255, 2002. 120

[134] D.E. Clark and D.R. Hahn. Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 397–401, 1995. 120

[135] M. Sariyar, A. Borg, and K. Pommerening. Evaluation of record linkage methods for iterative insertions. *Methods of Information in Medicine*, 48(5):429–437, 2009. 120

[136] S. Gomatam, R. Carter, M. Ariet, and G. Mitchell. An empirical comparison of record linkage procedures. *Statistics in Medicine*, 21(10):1485–1496, 2002. 120

[137] L. Gu, R. Baxter, D. Vickers, and C. Rainsford. Record linkage: Current practice and future directions. Technical report, CSIRO Mathematical and Information Sciences, 2003. 120, 121

[138] W.W. Cohen and J. Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the 8th ACM SIGKDD international conference*, KDD '02, pages 475–480, New York, NY, USA, 2002. ACM. 121

[139] T. Greenhalgh, G.W. Wood, T. Bratan, K. Stramer, and S. Hinder. Patients' attitudes to the summary care record and healthspace: qualitative study. *BMJ*, 336(7656):1290–1295, 6 2008. 133

[140] Coroners statistics 2012, england and wales. Technical report, Ministry of Justice, 2013. 133

[141] W. Alvey and B. Jamerson. *Record linkage techniques - 1997: proceedings of an international workshop and exposition, March 20-21, 1997, Arlington, Va.* Federal Committee on Statistical Methodology, Office of Management and Budget, 1997. 144

[142] T.B. Pedersen and C.S. Jensen. Research issues in clinical data warehousing. In *Proceedings of the Tenth International Conference on Scientific and Statistical Database Management*, pages 43–52. IEEE Computer Society, 1998. 145, 151, 154

[143] C.W. Bachman. The programmer as navigator. *Communications of the ACM*, 16(11):653–658, November 1973. 146, 147

[144] C.W. Bachman and S.B. Williams. A general purpose programming system for random access memories. In *Proceedings of the October 27-29, 1964, Fall Joint Computer Conference, Part I*, AFIPS '64 (Fall, part I), pages 411–422, New York, NY, USA, 1964. ACM. 146

[145] C.W. Bachman. The origin of the integrated data store (ids): The first direct-access dbms. *IEEE Annals of the History of Computing*, 31(4):42–54, October 2009. 146, 147

[146] K. North. Database systems: The first generation. *Ken North Computing, LLC*, 2012. 146, 147

[147] G.G. Dodd. Elements of data management systems. *ACM Computing Surveys*, 1(2):117–133, 1969. 146

[148] D.K. Burleson. *Inside the Database Object Model*. Taylor & Francis, 1998. 146

[149] E.F. Codd, S.B. Codd, and C.T. Salley. Providing olap (on-line analytical processing) to user-analysis: An it mandate. Technical report, 1993. 146, 150

[150] E.F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, June 1970. 146, 147

[151] D.L. Childs. Feasibility of a set-theoretic data structure : a general structure based on a reconstituted definition of relation. Technical report, The University of Michigan, 1968. 146

[152] S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. *SIGMOD Record*, 26(1):65–74, March 1997. 147

[153] B.A. Devlin and T. Murphy P. An architecture for a business and information system. *IBM Systems Journal*, 27(1):60–80, 1988. 148, 149, 152

[154] F. Hayes. The story so far. *Computerworld: Applications*, 101, 2002. 148

[155] P.P. Uhrowczik. Data dictionary/directories. *IBM Systems Journal*, 12(4):332–350, 1973. 149

[156] W.H. Inmon. *Building the Data Warehouse*. Wiley and Sons, 1992. 149, 150, 158

[157] OLAP Council. Olap council white paper. Technical report, OLAP Council, 1997. 150, 151

[158] N. Gorla. Features to consider in a data warehousing system. *Communications of the ACM*, 46(11):111–115, November 2003. 151

[159] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley & Sons, Inc., New York, NY, USA, 2nd edition, 2002. 151, 185

[160] D. Linstedt. Data vault series: Data vault series 1 - data vault overview. *The Data Administration Newsletter*, 2002. 152

[161] W.H. Inmon, D. Strauss, and G. Neushloss. *DW 2.0: The Architecture for the Next Generation of Data Warehousing*. The Morgan Kaufmann series in data management systems. Elsevier Science, 2010. 152

[162] R.A. Greenes, A.N. Pappalardo, C.W. Marble, and G.O. Barnett. A system for clinical data management. In *Proceedings of the November 18-20, 1969, Fall Joint Computer Conference*, AFIPS '69 (Fall), pages 297–305, New York, NY, USA, 1969. ACM. 153

[163] R.F. Walters. *ABCs of MUMPS: An Introduction for Novice and Intermediate Programmers*. Digital Press, Newton, MA, USA, 1989. 153

[164] C. Safran and L.E. Perreault. Management of information in integrated delivery networks. In Edward H. Shortliffe and Leslie E. Perreault, editors, *Medical Informatics*, Health Informatics, pages 359–396. Springer New York, 2001. 153, 154

[165] C. Safran, D. Porter, J. Lightfoot, C.D. Rury, L.H. Underhill, H.L. Bleich, and W.V. Slack. Clinquery: A system for online searching of data in a teaching hospital. *Annals of Internal Medicine*, 111(9):751–756, 1989. 153

[166] P.J. Haug, B.H. Rocha, and R.S. Evans. Decision support in medicine: lessons from the help system. *International Journal of Medical Informatics*, 69(2-3):273 – 284, 2003. Working Conference on Health Information Systems. 153

[167] R.M. Gardner, R.O. Crapo, A.H. Morris, and M.L. Beus. Computerized decision-making in the pulmonary function laboratory. *Respiratory Care*, 27(7):799–816, 1982. 153

[168] V. Dinu and P. Nadkarni. Guidelines for the effective use of entity–attribute–value modeling for biomedical databases. *International Journal of Medical Informatics*, 76(11):769–779, 2007. 153, 185, 186

[169] J. Annevelink, C.Y. Young, and P.C. Tang. Heterogenous database integration in a physician workstation. *Proceedings of the Symposium on Computer Applications in Medical Care*, pages 368–372, 1991. 153

[170] K. A. Marrs, S.A. Steib, C.A. Abrams, and M.G. Kahn. Unifying heterogeneous distributed clinical data in a relational database. *Proceeding of the Symposium on Computer Applications in Medical Care*, pages 644–648, 1993. 153

[171] A.R. Bakker. The development of an integrated and co-operative hospital information system. *Informatics for Health and Social Care*, 9(2):135–142, 1984. 153

[172] J.C. Prather, D.F. Lobach, L.K. Goodwin, J.W. Hales, M.L. Hage, and W.E. Hammond. Medical data mining: knowledge discovery in a clinical data warehouse. *Proceedings of the AMIA Annual Fall Symposyum*, 101(5), 1997. 154

[173] L. Ohno-Machado, A.A. Boxwala, J. Ehresman, D.N. Smith, and R.A. Greenes. A virtual repository approach to clinical and utilization studies: application in mammography as alternative to a national database. *Proceedings of the AMIA Annual Fall Symposium*, pages 369–373, 1997. 154

[174] T.B. Pedersen and C.S. Jensen. Clinical data warehousing-a survey. In *Proceedings of the VIII Mediterannean Conference on Medical and Biological Engineering and Computing*, page 20. University of Cyprus, 1998. 154

[175] J. Niinimaki, G. Selen, M. Kailajarvi, P. Gronroos, K. Irjala, and J. Forsstrom. Medical data warehouse, an investment for better medical care. In *MIE'96, Medical Informatics in Europe*, 1996. 154

[176] J. Kaufman. Healthcare and life sciences standards overview-technology for life: Nc symposium on biotechnology and bioinformatics. In *Biotechnology and Bioinformatics, 2004. Proceedings. Technology for Life: North Carolina Symposium on*, pages 31–41. IEEE, 2004. 155

[177] H. Hu, M. Correll, L. Kvecher, M. Osmond, J. Clark, A. Bekhash, G. Schwab, D. Gao, J. Gao, V. Kubatin, C.D. Shriver, J.A. Hooke, L.G. Maxwell, A.J. Kovatich, J.G. Sheldon, M.N. Liebman, and R.J. Mural. Dw4tr: A data warehouse for translational research. *Journal of Biomedical Informatics*, 44(6):1004–1019, December 2011. 155, 288

[178] S.N. Murphy, G. Weber, M. Mendis, V. Gainer, H.C. Chueh, S. Churchill, and I. Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2):124–130, 2010. 155, 233, 288

[179] T. Ganslandt, S. Mate, K. Helbing, U. Sax, and H.U. Prokosch. Unlocking data for clinical research–the german i2b2 experience. *Applied clinical informatics*, 2(1):116, 2011. 155

[180] S. Mate, T. Bürkle, F. Köpcke, B. Breil, B. Wullich, M. Dugas, H. Prokosch, and T. Ganslandt. Populating the i2b2 database with heterogeneous emr data: a semantic network approach. *Studies in health technology and informatics*, 169:502–506, 2010. 155

[181] K. Krishnan. *Data Warehousing in the Age of Big Data*. The Morgan Kaufmann Series on Business Intelligence. Elsevier Science, 2013. 156

[182] A. Sen and A.P. Sinha. A comparison of data warehousing methodologies. *Communications of the ACM*, 48(3):79–84, March 2005. 156, 159

[183] B. Boehm. A spiral model of software development and enhancement. *SIGSOFT Software Engineering Notes*, 11(4):14–24, August 1986. 158

[184] B. List, R.M. Bruckner, K. Machaczek, and J. Schiefer. A comparison of data warehouse development methodologies case study of the process warehouse. In Abdelkader Hameurlain, Rosine Cicchetti, and Roland Traunmuller, editors, *Database and Expert Systems Applications*, volume 2453 of

Lecture Notes in Computer Science, pages 203–215. Springer Berlin Heidelberg, 2002. 158, 159

[185] R. Kimball and J. Caserta. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Wiley, 2011. 158

[186] Y. Guo, S. Tang, Y. Tong, and D. Yang. Triple-driven data modeling methodology in data warehousing: a case study. In *DOLAP*, pages 59–66, 2006. 159

[187] D. Simoes. Conceptual framework for the construction and evaluation of data warehouses. In *Information Systems and Technologies (CISTI), 2010 5th Iberian Conference on*, pages 1–6. IEEE, 2010. 159

[188] P. Kruchten. *The Rational Unified Process: An Introduction*. The Addison-Wesley object technology series. Addison-Wesley, 2004. 159

[189] S. Asadullaev. Companion guidebook learning course: Data warehouse architectures and development strategy. Technical report, IBM, 2012. 160, 161

[190] P. Spyns, R. Meersman, and M. Jarrar. Data modelling versus ontology engineering. *SIGMOD Record*, 31(4):12–17, December 2002. 167, 186

[191] M. Uschold and M. King. Towards a methodology for building ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95*, 1995. 167

[192] A. Splendiani, A. Burger, A. Paschke, P. Romano, and S.M. Marshall. Biomedical semantics in the semantic web. *Journal of biomedical semantics*, 2(1):1–9, 2011. 167, 186

[193] B.H. Wixom and H.J. Watson. An empirical investigation of the factors affecting data warehousing success. *MIS quarterly*, 25(1):17–32, 2001. 180

[194] S. Bergamaschi, S. Castano, M. Vincini, and D. Beneventano. Semantic integration of heterogeneous information sources. *Data & Knowledge Engineering*, 36(3):215–249, 2001. 180

[195] N. Yates and C. Smith. Data quality strategy version 3. Technical report, Norfolk & Norwich University Hospital NHS Foundation Trust, Norwich, UK, 2009. 183, 184

[196] G. Simsion and G. Witt. *Data Modeling Essentials*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2004. 186

[197] A. Jacob. Generic design of web-based clinical databases. *Journal of Medical Internet Research*, 5(4), 2003. 186, 187

[198] P.M. Nadkarni. *Metadata-driven Software Systems in Biomedicine: Designing Systems that can adapt to Changing Knowledge*. Health Informatics. Springer, 2011. 186

[199] P. Nadkarni. An introduction to entity-attribute-value design for generic clinical study data management systems, center for medical informatics, yale university medical school, December 2013. 187

[200] K. Vanhaecht, M. Panella, R. Van Zelm, and W. Sermeus. An overview on the history and concept of care pathways as complex interventions. *International Journal of Care Pathways*, 14(3):117–123, 2010. 187, 188

[201] K. De Luc and J. Todd. *E-pathways: Computers and the patient's journey through care*. Radcliffe Publishing, 2003. 188

[202] S. Wakamiya and K. Yamauchi. What are the standard functions of electronic clinical pathways? *International Journal of Medical Informatics*, 78(8):543–550, 2009. 188

[203] A. Veselỳ, J. Zvárová, J. Peleška, D. Buchtela, and Z. Anger. Medical guidelines presentation and comparing with electronic health record. *International Journal of Medical Informatics*, 75(3):240–245, 2006. 188

[204] C. Spreckelsen, K. Spitzer, and W. Honekamp. Present situation and prospect of medical knowledge based systems in german-speaking countries. *Methods of Information in Medicine*, 49(3):207–218, 2010. 188

[205] Z. Huang, X. Lu, and H. Duan. On mining clinical pathway patterns from medical behaviors. *Artificial intelligence in medicine*, 56(1):35–50, 2012. 188, 189

[206] M. Stacey and C. McGregor. Temporal abstraction in intelligent clinical data analysis: A survey. *Artificial Intelligence in Medicine*, 39(1):1 – 24, 2007. 188

[207] M. Lang, T. Bürkle, S. Laumann, and H. Prokosch. Process mining for clinical workflows: challenges and current limitations. *Studies in health technology and informatics*, 136:229–234, 2007. 189, 191

[208] R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, and Piet P.J.M. Bakker. Application of process mining in healthcare - a case study in a dutch hospital. In *Biomedical Engineering Systems and Technologies*, pages 425–438. Springer, 2009. 189, 191, 219

[209] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, Cavallini A, S. Quaglini, and W. van der AALST. Process mining techniques: an application to stroke care. *Studies in health technology and informatics*, 136:573, 2008. 189, 191, 242

[210] S.T. Liaw, A. Rahimi, P. Ray, J. Taggart, S. Dennis, S. de Lusignan, B. Jalaludin, A.E.T. Yeo, and A. Talaei-Khoei. Towards an ontology for data quality in integrated chronic disease management: a realist review of

the literature. *International Journal of Medical Informatics*, 82(1):10–24, 2013. 189

[211] W. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, 2011. 190, 191, 220, 221, 224, 229

[212] W.M.P. Van Der Aalst, A.H.M. Ter Hofstede, and M. Weske. Business process management: A survey. In *Business process management*, pages 1–12. Springer, 2003. 190, 219

[213] M. Poulymenopoulou, F. Malamateniou, and G. Vassilacopoulos. Specifying workflow process requirements for an emergency medical service. *Journal of Medical Systems*, 27(4):325–335, August 2003. 190, 191

[214] Payam Homayounfar. Process mining challenges in hospital information systems. In *Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on*, pages 1135–1140. IEEE, 2012. 191

[215] S. Gupta. Workflow and process mining in healthcare. Master's thesis, Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, The Netherlands, 2007. 191, 219, 220, 221

[216] C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79–86, 2003. 191

[217] S. Stilou, P.D. Bamidis, N. Maglaveras, and C. Pappas. Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. *Studies in health technology and informatics*, 84(Pt 2):1399–1403, 2000. 191

[218] T. Imamura, S. Matsumoto, Y. Kanagawa, B. Tajima, S. Matsuya, M. Furue, and H. Oyama. A technique for identifying three diagnostic findings using association analysis. *Medical & Biological Engineering & Computing*, 45(1):51–59, 2007. 191

[219] S.M. Downs and M.Y. Wallace. Mining association rules from a pediatric primary care decision support system. In *Proceedings of the AMIA Symposium*, page 200. American Medical Informatics Association, 2000. 191

[220] J. Nahar, T. Imam, K.S. Tickle, and Y.P. Chen. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, 40(4):1086 – 1093, 2013. 192

[221] D. Sengupta, M. Sood, P. Vijayvargia, S. Hota, and P.K. Naik. Association rule mining based study for identification of clinical parameters akin to occurrence of brain tumor. *Bioinformation*, 9(11):555, 2013. 192

[222] S. Tsumoto and H. Abe. Mining clinical process in order histories using sequential pattern mining approach. In Jiuyong Li, Longbing Cao, Can Wang, KayChen Tan, Bo Liu, Jian Pei, and VincentS. Tseng, editors, *Trends and Applications in Knowledge Discovery and Data Mining*, volume 7867 of *Lecture Notes in Computer Science*, pages 234–246. Springer Berlin Heidelberg, 2013. 192

[223] T.Y. Lin and E. Louie. Association rules with additional semantics modeled by binary relations. In Masahiro Inuiguchi, Shoji Hirano, and Shusaku Tsumoto, editors, *Rough Set Theory and Granular Computing*, volume 125 of *Studies in Fuzziness and Soft Computing*, pages 147–156. Springer Berlin Heidelberg, 2003. 192

[224] GLOBOCAN. Prostate cancer incidence, mortality and prevalence worldwide in 2008, November 2013. 192

[225] National Institute for Health and Clinical Excellence. Prostate cancer overview, November 2013. 192

[226] R. Agrawal and R. Srikant. Fast algorithms for mining association rules.

In *Proceedings of the Twentieth International Conference on VLDB*, pages 487–499, 1994. 204, 205

[227] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007. 205

[228] J. Han, J. Pei, Y. Yin, and R. Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1):53–87, 2004. 205

[229] H. Mannila. Database methods for data mining. In *Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '98, 1998. 206

[230] B. Liu, W. Hsu, and Y. Ma. Mining association rules with multiple minimum supports. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 337–341, New York, NY, USA, 1999. ACM. 206

[231] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *Database Theory - ICDT'99*, pages 398–416. Springer, 1999. 206

[232] S.B. Yahia, T. Hamrouni, and E.M. Nguifo. Frequent closed itemset based algorithms: a thorough structural and analytical survey. *ACM SIGKDD Explorations Newsletter*, 8(1):93–104, 2006. 206

[233] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering, 1995*, pages 3–14, Mar 1995. 209

[234] S. Ramakrishnan and A. Rakesh. Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th In-*

ternational Conference on Extending Database Technology: Advances in Database Technology, EDBT '96, pages 3–17, London, UK, UK, 1996. Springer-Verlag. 209

[235] M.J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1-2):31–60, 2001. 210

[236] M. Abouelhoda and M. Ghanem. String mining in bioinformatics. In Mohamed Medhat Gaber, editor, *Scientific Data Mining and Knowledge Discovery*, pages 207–247. Springer Berlin Heidelberg, 2010. 210

[237] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu. Freespan: frequent pattern-projected sequential pattern mining. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 355–359. ACM, 2000. 210

[238] P. Fournier-Viger, A. Gomariz, A. Soltani, H. Lam, and T. Gueniche. Spmf: Open-source data mining platform, January 2014. 210

[239] P. Fournier-Viger, U. Faghihi, R. Nkambou, and E.M. Nguifo. Cmrules: Mining sequential rules common to several sequences. *Knowledge-Based Systems*, 25(1):63 – 76, 2012. Special Issue on New Trends in Data Mining. 210, 211

[240] Y. Hirate and H. Yamana. Generalized sequential pattern mining with item intervals. *Journal of computers*, 1(3):51–60, 2006. 214

[241] W. Van der Aalst, T. Weijters, and L. Maruster. Workflow mining: discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142, Sept 2004. 219

[242] W.J.M. Weijters, WMP van der Aalst, and A.K. Alves de Medeiros. Process mining with the heuristicsminer algorithm, wp 166. Technical report, Eindhoven University of Technology, 2006. 219, 220, 221

[243] G.M. Veiga and D. Ferreira. Understanding spaghetti models with sequence clustering for prom. In Stefanie Rinderle-Ma, Shazia Sadiq, and Frank Leymann, editors, *Business Process Management Workshops*, volume 43 of *Lecture Notes in Business Information Processing*, pages 92–103. Springer Berlin Heidelberg, 2010. 220

[244] B.F. van Dongen, A.K. de Medeiros, H.M.W. Verbeek, AJ.M.M. Weijters, and W.M.P Van Der Aals. The prom framework: A new era in process mining tool support. In *Applications and Theory of Petri Nets 2005*, pages 444–454. Springer, 2005. 220

[245] C. Friedman, G. Hripcsak, S.B. Johnson, J.J. Cimino, and P.D. Clayton. A generalized relational schema for an integrated clinical patient database. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 335. American Medical Informatics Association, 1990. 233

[246] M.J. O'connor, R.D. Shankar, D.B. Parrish, and A.K. Das. Knowledge-data integration for temporal reasoning in a clinical trial system. *International Journal of Medical Informatics*, 78:S77–S85, 2009. 235

[247] E.S. Berner. *Clinical Decision Support Systems: Theory and Practice*. Health Informatics. Springer, 2007. 242

[248] R. Soley and the OMG Staff Strategy Group. Model driven architecture. *Object Management Group (OMG) white paper*, 2000. 242

[249] F. Truyen. The fast guide to model driven architecture - the basics of model driven architecture. Technical report, Object Management Group, Cephas Consulting Corp, 2006. 243

[250] D. Frankel. *Model Driven Architecture: Applying MDA to Enterprise Computing*. John Wiley & Sons, Inc., New York, NY, USA, 2002. 243

[251] J. Martin. *Rapid Application Development*. Macmillan Publishing Co., Inc., Indianapolis, IN, USA, 1991. 243

[252] F. Buschmann, K. Henney, and D. Schimdt. *Pattern-oriented Software Architecture: On Patterns and Pattern Language*, volume 5. John Wiley & Sons, 2007. 245

[253] J. Nakashima, C. Ozu, T. Nishiyama, M. Oya, T. Ohigashi, H. Asakura, M. Tachibana, and M. Murai. Prognostic value of alkaline phosphatase flare in patients with metastatic prostate cancer treated with endocrine therapy. *Urology*, 56(5):843 – 847, 2000. 261

[254] S.J. Weinstein, K. Mackrain, R.Z. Stolzenberg-Solomon, J. Selhub, J. Virtamo, and D. Albanes. Serum creatinine and prostate cancer risk in a prospective study. *Cancer Epidemiology Biomarkers & Prevention*, 18(10):2643–2649, 2009. 263

[255] A.M. Hill, N. Philpott, J.D.S. Kay, J.C. Smith, G.J. Fellows, and S.H. Sacks. Prevalence and outcome of renal impairment at prostatectomy. *British Journal of Urology*, 71(4):464–468, 1993. 263

[256] W. van der Aalst, A. Adriansyah, A.K. de Medeiros, et al. Process mining manifesto. In *Business process management workshops*, pages 169–194. Springer, 2012. 275

[257] E. Geisler. *Managing the Aftermath of Radical Corporate Change: Reengineering, Restructuring, and Reinvention*. Greenwood Publishing Group, 1997. 278

[258] D. Glez-Pena, M. Reboiro-Jato, P. Maia, M. Rocha, F. Diaz, and F. Fdez-Riverola. Aibench: A rapid application development framework for translational research in biomedicine. *Computer Methods and Programs in Biomedicine*, 98(2):191 – 203, 2010. 288

Appendix A - Information Systems and Data Collection Details

1.1 Summary of Information Systems

The section below summarises the 15 different Information Systems or data sources investigated in this thesis.

- **PACS and RIS - Picture Archiving Communication System (PACS) and Radiology Information System (RIS)**
  The PACS is dedicated to radiological imaging storage and visualisation. Administration information and imaging text reports are also stored in the RIS and are retrieved using Business Objects (BO), a Business Intelligence software. PACS Room staff produces reports and datasets from the latter.

- **OPT - Operating Theatre Database (ORSOS)**
  The Operating Room Scheduling Office System (ORSOS) contains information and coding (OPCS) on any surgical procedures. Datasets can be requested from the ORSOS IT staff based on a particular OPCS code or free-text keyword search.

- **LAB - Histopathology and Biochemistry (ICE Database)**

This database contains all biochemistry test results as well as histopathology text reports. Biochemistry data is retrieved using the BO software. Histopathology data is retrieved based on a free-text keyword search with the assistance of a consultant histopathologist who has built a querying system to facilitate retrieval.

- **CRE - Hospital Cancer Register (Somerset Cancer Registry)**
  This system was introduced in 2007 and stores the dataset for the National Cancer Waiting Times Database. This database has the potential to store almost all information on cancer patients (from biochemistry to appointments and notes) and has more recently been fed data from other systems as well as outcomes of Multidisciplinary Team Meetings (MDTs). This is the most accurate database for counts of cancer patients. Datasets are retrieved by liaising with the Cancer Information Services team but the system includes built-in reporting tools.

- **PAS - Patient Administration Systems (McKesson PAS)**
  The PAS system contains general admissions and demographics information. Clinical coding for diagnoses (ICD) is only recorded for inpatients. Procedures codes (OPCS) and procedure dates are also stored in this system. Data is retrieved using the BO software but it is also possible to request PAS data from the Information Services.

- **ONC and ONT - Oncology Department Database (Varian ARIA/MedOncology) and the Notes Database System (ONT)**
  The oncology department database, which was originally created (2003) to store radiotherapy information, now (2007) also stores detailed information such as case notes and patient history from patients who are treated by the oncology department. The Notes System, which is now obsolete, contains free-text records such as case notes prior to 2008, from the Oncology department. Data can be retrieved from this system by liaising with the

Oncology staff and running back-end queries.

- **CR - Eastern Cancer Registration and Information Centre (ECRIC)**
  The ECRIC database is the local cancer register. It stores a core dataset which is agreed at national level and provides additional information, some of which is not stored in hospital systems, such as coding for cause of death. This system also stores tumour staging information in a canonical form, and provides information on cancer diagnoses and treatments, even if the patient receives treatment in other hospitals in the country.

- **NSTS - NHS Strategic Tracing Service (NSTS)**
  The national tracing system contains most accurate and up-to-date information on patients registered with the NHS. The information is essentially demographic, including addresses and date of death. The information services department at the NNUH liaises directly with the NSTS to process lists of patients. This system does not provide causes of death.

- **ORT - Orthopaedics System (Bluespier)**
  This system facilitates the retrieval of information on patients who have been treated for pathological fractures from advanced metastatic cancers. The orthopaedics IT staff can extract this data, or it can be obtained from the Bluespier system, which includes a comprehensive query builder for data retrieval.

- **OSR/NSR, TIA and CST - Stroke Databases**
  The Stroke Register Database was initially set up in 1996. The initial database, OSR (Old Stroke Register), now in Microsoft Access format, contains a set number of attributes and its data spans from 1996 to 2008. The new stroke register database (NSR) is an improved version of the OSR, still in Microsoft Access where further attributes were introduced, and includes strokes from 2008 to 2013. In 2013 the Capture Stroke system (CST) was

introduced as a replacement for the NSR database and is currently the preferred method for collection of stroke data and submission of data for national audits. The CST system allows and encourages multiple users (including nurses and clinicians) to add information directly onto the database. At the time of writing the NSR database was still being used together with CST. The TIA database is similar to the NSR database, also in Microsoft Access, except it records *mini-strokes* and further attributes such as blood pressure.

- **AE - Accident and Emergency Database (Symphony)**
  The sole use of the Accident and Emergency (AE) system in this thesis was the inspection of admissions regarding stroke patients. No data from this system was used becasue the OSR/NSR systems will already contain the most relevant information for stroke patients, however, this would still be an interesting system to explore in the future.

1.2 Research Stroke Register Data Warehouse

A complete list of the Research Stroke Register Data Warehouse (RSR) is given below, together with key figures of the cohort of stroke and TIA patients. The details and numbers pertain to the RSR version 3.

Key Figures

Strokes

Average age 76.65 (SD 12), range [0, 105]
Female 53.4%, Male 46.6%
Average number of patients per year: 919 (SD 159) (1997-2011)
Type of stroke: Ischemic 82%, Haemorrhagic + SAH 15%, other 3%

TIAs

Average age 71.74 (SD 12), range [4, 98]
Female 53.7%, Male 46.3%
Average number of patients per year: 847 (SD 217) (Jan 2003 - Nov 2013)

Tables

Patient Table
LAB Data Source
Records: 14,782
Unique Patients IDs: 14,782

List of Attributes:

Patient Attribute	Description
Anon_ID	Internal anonymised ID.
HospNo	Hospital Number.
Surname	Patient Last Name.
NHS Number	NHS Number.
Gender	Gender.
DOB_SR	Date of Birth in Stroke Register.
DOB_PAS	Date of Birth in PAS.
Notes	Notes on this patient, any linkage issues that need to be fixed.
NSTS Validated Date	Date when National Database was last checked for this patient.
Duplicated NHS Number	When there are patients with more than one NHS number, keep both until any linkage issues are resolved.
Flag_Use	Boolean value as to whether this record has been fully validated and can be used.

Co-morbidities and Follow-up Table
PAS Data Source
Records: 409,539
Unique Patient IDs: 10,562

List of Attributes:

Comorbid Attributes	Description
Cpkey	(database use) Internal primary key (unique number for each row).
anon_ID	Anonymised patient ID.
Episode End Date	Episode end date.
Episode Start Date	Episode start date.
Date of Birth	Patient's date of birth.
Date of Death	Patient's date of death.
Discharge Date	Discharge date (from episode).
Procedure Date	Procedure date (relating to episode).
Admission Date	Admission date (relating to episode).
Sex	Gender.
Alert Notes	PAS Alerts (e.g. hospital infection, particular allergies, etc.).
Ethnic Group	Patient ethnic group.
Alert Description	PAS Alerts - further description.
Diagnosis Codes	ICD codes (primary code followed by any other co-morbid codes).
Procedure Codes	OPCS codes for any procedures pertaining to the episode.

Table Strokes
Main Stroke Data Table, overlapping Data Sources NSR+OSR
Records: 16,232
Unique Patient IDs: 14,782

List of Attributes:

Stroke Attributes	Stroke Attributes (cont.)
HospNo	PRENDS
anon_ID	APHASIA1
Surname	Bamford
Source	COMMUN7
No	SWALPR7
Consultant	STSEEN
GPPC	BSCAN
Ward	CT48HR
DOB	CT24HR
Age	Type of Stroke
Gender	Date of d/c from hosp
Ethnic	WARDD
OnSetDate	STATD
OnSetTime	DESTD
AdmissionDate	COGFUND
AdmissionTime	URIBID
ArrivalDate	ESTBID
finalDate	Physio
Year	OT
First Ward	SLT
CONLEV0	Dietician
SIDE	PEG
ARM0	SLT48
LEG0	P24
PRELIV	AP/AD
PRERANK	AP/DC
PREMOB	IST
PREDISA	Rankin on discharge

Biochemistry Table
LAB Data Source
Records: 142,920 (after granularity was reduced, was previously 518,464)
Unique Patients IDs: 142,920

List of Attributes:

Biochemistry Attribute	Description
Bpkey	(database use) Internal primary key (unique number for each row).
hospNo	NNUH hospital number.
bloodName	Blood name and internal hospital code.
totalReadings	Total number of blood readings between admission and discharge.
Average	Average for a particular numeric blood value.
stDev	Standard Deviation for the above average.
Max	Maximum blood value found between admission and discharge.
Min	Minimum blood value found between admission and discharge.
admissionReading	Blood value found at admission.
dateAdmissionReading	Date of blood value closest to admission.
dischargeReading	Blood value found at discharge.
dateDischargeReading	Date of blood value closest to discharge.
linkageIssue	(validation use) Validation of date of birth (stroke register - PAS).

Table Strokes_OSR
Old Stroke Register (OSR) Data Source
Records: 10,769
Unique Patient IDs: 10,033

List of Attributes:

HospNo	ARM0	DEFDUR	ESTBID
Surname	LEG0	COGFUN7	Physio
No	PRELIV	COMMUN7	OccupationalTherapy
Proforma	PRERANK	SWALPR7	SLT
Consultant	PREMOB	BLAD7	Dietician
NHSNO	PREDISA	PTSEEN	PEG
GPPC	PRENDS	OTSEEN	Previous
Ward	CONLEV1	STSEEN	AF
DOB	FACE1	BSCAN	Diabetes
Age	ARM1	CT48HR	Smoker
Gender	LEG1	CT24HR	Hypertension
Ethnic	APHASIA1	FINDIAG	IHD
OnSetDate	DYSARTH1	DisChargeDate	Other Coronary Disease
OnSetTime	CONFUS1	WARDD	Periph Vasc Disease
ADMDate	CGP1	STATD	Hyperlipideamia
ADMTime	FIELD1	DESTD	SLT48
Delay	NEGLECT1	COGFUND	P24
ADMWard	BSTEM1	URIBID	APADM
CONLEV0	OTHDEF1	TRANSBID	APD/C
SIDE	CLASS1	MOBID	IST

Table Strokes_NSR
New Stroke Register (NSR) Data Source
Records: 5,464
Unique Patient IDs: 5,447

List of Attributes:

HospNo	First Ward	Time first contact st team	ADM Gluc taken	Time of OT	Primary Diagnosis	FINAL D/C DATE	Repatriated from	Included 3
Surname	ArrivalDate	FirstContactSTDateANDTime	ADM GLUC	OT DATE AND TIME	Date of d/c from hosp	FINAL D/C TIME	Repatriated to	Date consented 3
PCT	ArrivalTime	Clinically approp SU	Swallow test indicated	OT	Time of d/c from hosp	FINAL D/C DateANDTime	Date repatriation requested	Date randomised 3
No	ArrivalDATE AND TIME	ADM S U Date	Date of swallow	DATE SLT	DateANDTimeD/Cfrm HOSP	Any time spent in Neuro or CCU	Time repatriation requested	Date withdrawn 3
BF	DISCHARGE LOS	ADM S U Time	Time of swallow	TIME SLT	Date of d/c from stroke unit	Appropriate days off SU	Date of repatriation	Divert Patient?
NHS NO	WARD LOS	ADM SU Date and Time	SWALLOW DATE AND TIME	SLT DATE AND TIME	Time of d/c from stroke unit	receive THROMB	Time of repatriation	Repatriation Destination
Postcode	% GUNT	Reason Non Admit SU	PRELIV	SLT	DateANDTimeD/CfrmSU	Reason no thromb	Trial 1	Date of repatriation Pre-Alert to DGH
TRUST	AdmissionDATE	Imaging indicated	PRERANK	PEG DATE	Rankin on discharge	Date of Thrombolysis	Date screened/PIS given	Date of notification of estimated date of repatriation to DGH
NNUHID	AdmissionTIME	Date of imaging	PREDISA	PEG	DISCHARGE DESTINATION	Time of Thrombolysis	Included	Date & Time formal notification of repatriation
DOB	AdmissionDateANDTime	Time CT study	PRENDS	WEIGHED DATE	TRANSFERRED	Thrombolysis Date/Time	Date consented	Date & Time confirmation received at DGH

Age	Direct to Neuro or CCU	Scan Date and Time	Evidence of AF	WEIGHED	ESD	Consultant	Date randomised	Time transport arrnaged
Sex	RETURNadmissionDATE 1	Time CT approval	Anticoagulant?	Dietician	ESD Area?	Present	Date withdrawn	Date & Time of repatriation
Ward	1 FROM	CT or MRI	AP/AD	DIET Date	Eligible for joint care plan	Speciality	Trial 2	Comments Repatriation
OnSetDate	RETURNadmission DATE 2	URGENTscan	AP/DC	MOOD	Joint care plan	Is there an established infarct?	Date screened 2	
OnSetTime	2 FROM	ADM GCS	NOTE	MOOD Date	Telemedicine	NIHSS 2HRS	Included 2	
STROKE in hosp	FAST in A+E	SIDE	Date of physio	MOOD Score	D/C to BEECH DATE	NIHSS 24HRS	Date consented 2	
On waking	Alert stroke date	NIHSS on adm	Time of physio	SURGERY?	D/C to BEECH TIME	2ND SCAN DATE	Date randomised 2	
First contact	Alert Stroke time	NIHSS score	PHYSIO DATE AND TIME	Date of Surgery	DATE D/C REHAB	2ND SCAN CREAT TIME	Date withdrawn 2	
ArrivalPOINT	AlertStroke DateANDTime	Adm BP taken	Physio	Type of stroke	TIME D/C REHAB	HAEM?	Trial 3	
Ambulance	First contact st team	ADM BP	Date of OT	Bamford	DateANDTimeD/C REHAB	Repatriated?	Date screened 3	

Table TIAs
TIA Database Data Source
Records: 9,325

List of Attributes:

ID NO	ABCD taken 1st contact	Time of MRI	Notes
NNUHID	ABCD2	Date/Time MRI	AF
NHS	DATE OFFERED	Blood test done	GP SURGERY
GP POSTCODE	DATE SEEN	ECG done	MED AT-Time Refer Hrs
AGE	TIME SEEN	Aspirin indicated	MED AT-Time Seen Hrs
SEX	Date/Time Seen	Aspirin commenced	MED AT-Time CT Hrs
INPATIENT	DNA	Statin indicated	MED AT-Time Dop Hrs
ADMISSION DATE	DIAGNOSIS	Statin commenced	
DISCHARGE DATE	Detailed diagnosis	Warfarin indicated	
FORM COMPLETE	Doppler indicated	Warfarin taken	
DATE OF ONSET	DATE OF DOPPLER	Referred for Surgery	
TIME OF ONSET	TIME OF DOPPLER	Date referred for surgery	
Date/Time Onset	Date/Time Doppler	Time referred for surgery	
DATE sought med att 1st contact GP A+e	CT indicated	Date of surgery	
Time sought medical attention	Date of CT	Smoking	
Date/Time Med Att	Time of CT	Diet/Salt	
DATE REFERRED	Date/Time CT	Alcohol	
TIME REFERRED	CT or MRI	Exercise	
Date/Time Referral	MRI head scan indicated	Driving	
REFERRAL SOURCE	Date of MRI	Management plan given	

1.3 Prostate Cancer Operational Data Store

Additional data tables containing further biochemistry data or comorbidities were linked from the stroke databases (e.g. haemoglobin, full blood count, etc.) or directly from the systems but were not in the ODS data store.

CaP ODS - Key Figures & Tables

ODS Tables Overview

The ODS comprises a total of 13 Tables and 4,464,055 records.
A summary of the tables is given below.
Descriptions of the tables and their attributes are given in the next pages.

Table Name	Code	Data Source	Short Description	Records
Table PAS_C61	PAS	Patient Administration System.	Hospital episodes.	9,461
Table Lab_OT1	LAB	Biochemistry System.	Multiple blood tests.	4,016,561
Table Lab_OT2	LAB	Biochemistry System.	Testosterone readings.	15,524
Table Lab_OT3	LAB	Biochemistry System.	Vitamin D and Calcium.	224,952
Table Lab_PSA	LAB	Biochemistry System.	PSA tests.	176,472
Table ORT	ORT	Orthopaedics System.	Femoral neck fractures.	1,783
Table CRE	CRE	Hospital Cancer Register and Cancer Waiting Times System.	Cancer Waiting Times.	3,064
Table RAD	RAD	Radiotherapy Database.	Radiotherapy Sessions.	2,769
Table RIS	RIS	Radiology Information System.	Imaging.	1,281
Table OPT	OPT	Operating Theatre Information System.	Operations.	4,637
Table HIS	LAB	Histopathology System.	Biopsies.	5,083
Table ONT	ONT	Oncology NOTES System.	Oncology Treatment.	1,601
Table ONC	ONC	New Oncology Department System.	Oncology Treatment.	867
			Total Records	4,464,055

ODS Table Details & Attributes

Table PAS_C61 - Patient Administration System

Data Origin and Constraints	PAS Data Source (via Business Intelligence software) where Diagnosis Code 1 = "C61" and Episode Start Date between 01/01/2001 - 30/12/2010.	
Records	9,461	3.70 records per patient
Unique Patient IDs	2,552	
Description	This list of prostate cancers in this table was defined by any hospital discharge letters where the ICD code (C61) was coded.	
Retrieval Date (ODS)	11/08/2011	

List of Attributes:

Attribute Name	Description
Patient Number	Hospital Number.
Episode End Date	Episode end date.
Episode Start Date	Episode start date.
Consultant Code	Consultant code.
Consultant	Consultant name.
NHS Number	NHS Number.
Date of Birth	Date of Birth.
Date of Death	Date of Death.
Diagnosis Codes	List of all ICD diagnoses in discharge letter for episode (includes comorbidities).
Diagnosis Code 1	First ICD code associated with the episode.
Procedure Codes	OPCS 4.5 codes for any procedures undertaken during episode.
Discharge Date	Date patient was discharged.
Procedure Date	Date of the procedure (if any).
Discharge Destination Nat Code	Discharge destination code.
Discharge Destination	Discharge destination name.
Admission Date	Admission Date.
Sex	Gender (always Male).
Alert Notes	Alert Notes (allergies, transport, infection).
Alert Description	Further text description of alerts.

Table Lab_OT1 - Biochemistry System (multiple biochemistry tests)

Data Origin and Constraints	LAB Data Source (via Business Intelligence software) where Test Code = alkaline phosphatise or creatinine or urea or AST or GGT or total bilirubin, and Date of Entry between 01/01/2003 - 31/12/2007.	
Records	4,016,561	10.15 records per patient
Unique Patient IDs	395,531	
Retrieval Date (ODS)	23/06/2009	

List of Attributes:

Attribute Name	Description
Hospital Number	Hospital Number.
Age at Episode	Age recorded when blood sample was taken.
Clinical History	Free text clinical history manually entered when blood sample was taken.
Date of Entry	Date the blood sample was taken.
Time of Entry	Time the blood sample was taken.
Date of Authorisation	Date the blood sample was analysed.
Test Data	The blood reading.
Fasting Indicator	Indicator whether patient was fasting.

Table Lab_OT2 - Biochemistry System (Testosterone)

Data Origin and Constraints	LAB Data Source (via Business Intelligence software) where Test Code = testosterone and Date of Entry between 01/01/2003 - 31/12/2007.	
Records	15,524	1.46 records per patient
Unique Patient IDs	10,637	
Retrieval Date (ODS)	23/06/2009	

This table's list of attributes is the same as the previous table (Lab_OT2).

Table Lab_OT3 - Biochemistry System (Vitamin D and Calcium)

Data Origin and Constraints	LAB Data Source (via Business Intelligence software) where Test Code = Vitamin D or Calcium and Date of Entry between 01/01/2003 - 31/12/2007.	
Records	224,952	3.39 records per patient
Unique Patient IDs	66,455	
Retrieval Date (ODS)	23/06/2009	

This table's list of attributes is the same as the previous table (Lab_OT3).

Table Lab_PSA - Biochemistry System (Prostatic Specific Antigen)

Data Origin and Constraints	LAB Data Source (via Business Intelligence software) where Test Code = Vitamin D or Calcium and Date of Entry between 01/01/2003 - 31/12/2011.	
Records	176,472	2.90 records per patient
Unique Patient IDs	60,794	
Retrieval Date (ODS)	22/01/2012	

This table's list of attributes is the same as the previous table (Lab_OT3).

Table ORT - Orthopaedics System (Femoral Neck)

Data Origin and Constraints	ORT Data Source (via own reporting software) where Diagnosis is neck of femur fracture, and Date of Entry between 01/01/2008 - 31/10/2011.	
Records	1,783	1.10 records per patient
Unique Patient IDs	1,651	
Retrieval Date (ODS)	31/10/2011	

List of Attributes:

Attribute Name	Description
Hospital Number	Hospital Number.
Operation Date	Date of operation.
Procedure	Procedure Name, detail.
Procedure 1 Name	Procedure Name, general.
Consultant	Consultant code.
Surgeon	Surgeon code.
Primary Diagnosis Desc	Description of Primary Diagnosis.

Table CRE - Hospital Cancer Register and Cancer Waiting Times Database

Data Origin and Constraints	CRE Data Source (via own reporting software and integrated between old and new cancer waiting times databases) where Diagnosis is ICD C61, and Date of Entry between 01/01/2001 - 31/12/2010.	
Records	3,064	1.12 records per patient
Unique Patient IDs	2,714	
Retrieval Date (ODS)	10/10/2011	

List of Attributes:

Attribute Name	Description
NHS#	NHS Number.
NoEps	Number of episodes.
Source	Old Waiting Times Database: 2002 - 2008 New Waiting Times Database: 2008 - 2010
RefSource	Source of Referral (GP, A&E, other)
Prority	Urgent Referral.
RefDate	Referral Date.
BodySite	Body Site (prostate).
FirstSeen	Date first seen.
FirstSeenOrg	Organisation first seen.
DelayReasonCodeFirstSeen	Reason for delay (if any) in first seen date.
MDTDate	Date of Multidisciplinary Team Meeting
Laterality	Tumour laterality (if any).
EventType	Pathway treatment primary/recurrent/metastatic/etc.
DTTORG	Organisation responsible for decision to treat.
DTT	Decision to treat date.
Treat	Treatment date.
Modality	Type of Treatment.
ClinicalTrial	Clinical Trial indicator (if any).
CareSetting	Inpatient admission, day case, out-patient, other.
TreatORG	Organisation responsible for treatment.
RTIntent	Radiotherapy Intent (palliative, anti-cancer, other).

Table RAD - Radiotherapy Database

Data Origin and Constraints	RAD Data Source (via own reporting software and integrated between old and new radiotherapy databases) where Site is prostate and/or diagnosis is prostate cancer, and Date of Entry between 01/01/2003 - 31/12/2010.	
Records	2,769	1.40 records per patient
Unique Patient IDs	2,013	
Retrieval Date (ODS)	27/10/2011	

List of Attributes:

Attribute Name	Description
Hospital Number	Hospital Number.
Intent	Radiotherapy Intent (Palliative, Curative or other).
Site treated	Site treated.
Date of Treatment	Date Treatment Commenced.
NoOfFractions	Number of Fractions (Sessions).
Fractions of Photons	Fractions of Photons (subset of NoOfFractions).
Fractions of Superficial	Fractions of Superficial (subset of NoOfFractions).
Fractions of Electrons	Fractions of Electrons (subset of NoOfFractions).

Table RIS - Radiology Information System

Data Origin and Constraints	RIS Data Source (via business intelligence software) where word prostate appears in Report Text, and Date of Entry between 01/01/2003 - 31/12/2010.	
Records	1,281	1.50 records per patient
Unique Patient IDs	866	
Retrieval Date (ODS)	11/02/2012	

List of Attributes:

Attribute Name	Description
Hospital Number	Hospital Number.
Date of Birth	Date of Birth.
Report Date.	Date the radiological report was produced.
Exam Type	Type of Imaging / Imaging Modality.
Report Text Line	Line of text report where word prostate appears.

Table OPT - Operating Theatre Information System

Data Origin and Constraints	OPT Data Source (via own reporting software) where word prostate appears in Report Text, and Date of Entry between 01/01/2003 - 31/12/2010.	
Records	4,637	1.13 records per patient
Unique Patient IDs	4,121	
Retrieval Date (ODS)	24/09/2011	

List of Attributes:

Attribute Name	Description
ORSOS NO	Internal Database Number.
DATE OP	Date of Operation.
ADMITTING CONSULTANT	Consultant code.
1ST PROCEDURE_CODE	1st Procedure Code (OPCS).
1ST PROCEDURE	1st Procedure Name.
1ST PROCEDURE SITE	1st Procedure Site.
2ND PROCEDURE_CODE	2nd Procedure Code (OPCS).
2ND PROCEDURE	2nd Procedure Name.
2 ND PROCEDURE SITE	2nd Procedure Site.
3RD PROCEDURE_CODE	3rd Procedure Code (OPCS).
3RD PROCEDURE	3rd Procedure Name.
3RD PROCEDURE SITE	3rd Procedure Site.
4TH PROCEDURE_CODE	4th Procedure Code (OPCS).
4TH PROCEDURE	4th Procedure Name.
4TH PROCEDURE SITE	4th Procedure Site.
PT NO	Hospital Number.
BIRTHDATE	Date of Birth.
AGE	Age at Operation.
SEX	Gender.
Day of Week	Operation week day.

Table HIS - Histopathology System

Data Origin and Constraints	LAB Data Source (albeit different working space in the system; via own reporting software and business intelligence software, and integrated old and new attributes) where word prostate appears in Report Text, and Date of Entry between 01/01/2003 - 31/12/2010.	
Records	5,083	1.17 records per patient
Unique Patient IDs	4,342	
Retrieval Date (ODS)	29/11/2011	

List of Attributes:

Attribute Name	Description
HospNo	Hospital Number.
Date of Entry	Date of Histological Report was produced.
Age	Patient's age at the time of tissue sample taken.
FullReport	Full histopathology text report (integrated).
SpecType	Specimen Type.

Table ONT - Oncology NOTES System

Data Origin and Constraints	ONT Data Source (via own reporting software) where word prostate appears in primary site, and Date of Entry between 01/01/2000 - 25/04/2007.	
Records	1,601	1 record per patient
Unique Patient IDs	1,601	
Retrieval Date (ODS)	07/10/2010	

List of Attributes:

Attribute Name	Description
Hospital Number	Hospital Number.
Primary Site	Primary Site = Prostate.
Diagnosis	Text diagnosis (no coding).
Consultant	Consultant code.
Date Registered	Date patient registered in the system (oncology department).
Date of Birth	Date of Birth.

Table ONC - New Oncology Department System

Data Origin and Constraints	ONC Data Source (via own reporting software) where diagnosis code ICD C61, and Date of Entry between 01/01/2003 - 09/02/2012.	
Records	867	1.004 records per patient
Unique Patient IDs	863	
Retrieval Date (ODS)	09/02/2012	

List of Attributes:

Attribute Name	Description
Hospital Number	Hospital Number.
Date of Birth	Date of Birth.
Age	Age at Admission.
Date Entered onto Aria	Date patient was entered onto the system.
Diagnosis Date	Date of Diagnosis.
Age at Diagnosis	Age at Diagnosis.
ICD Code	ICD Diagnosis Code.
Consultant	Consultant Code.
Last Visit	Date Patient last visited Oncology department.
Description	Text line containing description of the diagnosis and tumour staging, Gleason grade, PSA or other information (not consistently recorded or coded).

Appendix B - Background on Prostate Cancer

B.1 Carcinoma of the Prostate

According to Cancer Research UK [b3], prostate cancer has overtaken lung cancer as the most common type of cancer in males (over 30,000 cases detected in 2001 in the UK). In the United States, the American Cancer Society estimated 30,200 deaths and 189,000 new detected cases for the year 2002 [b6]. This type of cancer lacks effective detection methods when the disease is on its early stages and screening tests are not entirely reliable. Prostate cancer may spread to other anatomical sites without signs or symptoms for many years. Hormone escape is the most advanced stage of this disease. In this situation patients receive treatment but the cancer recurs and there is no control over its growth. The causes for this disease (Etiology) are still not well understood. All these issues make prostate cancer a major research topic in modern medicine.

Cancer (malignant neoplasm) is a group of diseases characterized by uncontrolled growth of abnormal cells that have mutated from normal tissues [b1, b6]. Shall this uncontrolled growth affect vital organs or spread throughout the body, essential systems can be damaged, often leading to death.

Normal cells grow, divide, and die. Cancer cells, instead of dying, continue to grow new abnormal cells. Apoptosis is the programmed process of eliminating cells (cell death). Cancer cells avoid this process and continue to multiply. These abnormal (cancer) cells often travel to other parts of the body where they grow further and replace normal tissue. This spreading process, called metastasis, occurs as the cancer cells get into the bloodstream or lymphatic system.

The prostate is a gland in the male reproductive system that is responsible for the storage of seminal fluid and controlling urination. This seminal fluid is released to form part of semen [b2]. The prostate is located below the urinary bladder, in front of the rectum, as shown in figure 1.1. If the prostate grows too large, the flow of urine can be slowed or stopped. An enlarged prostate may be caused by a benign prostate hypertrophy (BPH). This is a common problem in elderly men with normally functioning testicles (producing testosterone). In fact, an enlarged prostate resulting of BPH can shrink in size if the testicles are surgically removed (orchidectomy) or by medication. BPH does not have any connection with cancer and it does not put patients at risk for prostate cancer [b1].

Similarly to BPH, prostate cancer in an advanced stage may result in an enlarged size of the prostate (due to the growth of the tumour, shown in figure 1.2). This may produce symptoms such as the inability to urinate. There are cases when prostate cancer can regress if testosterone levels are lowered (androgen ablation: orchidectomy or medication [b1]). The process of suppressing male hormones is called androgen ablation and focuses on blocking prostate cancer cells from getting dihydrotestosterone (DHT), a hormone produced in the prostate [b2].

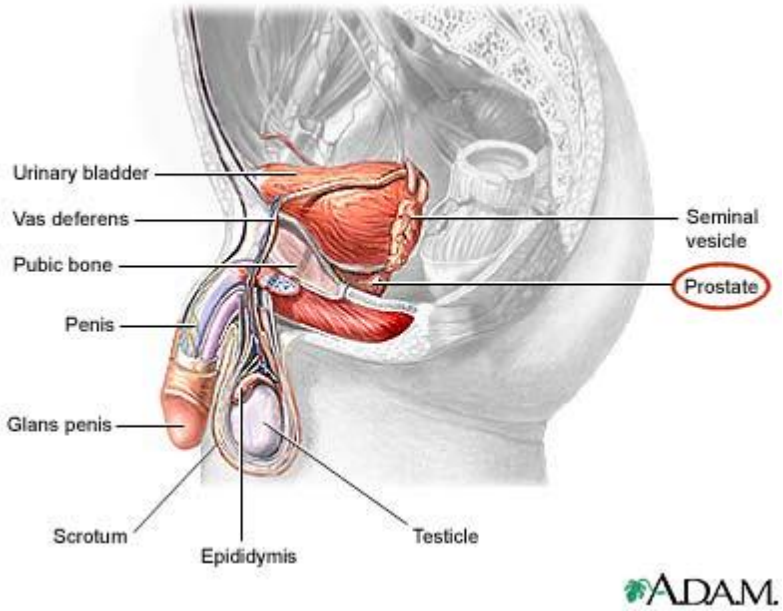


Figure 2.1. Basic male reproductive anatomy (MedLine, Adam).

'Prostate cancer is a malignant (cancerous) tumour (growth) that consists of cells from the prostate gland' [b5]. This tumour normally grows slowly and only in an advanced stage can produce symptoms, and reach other areas of the body such as bones, lungs, and liver [b5]. Figure 2.3 shows the different zones of the prostate.

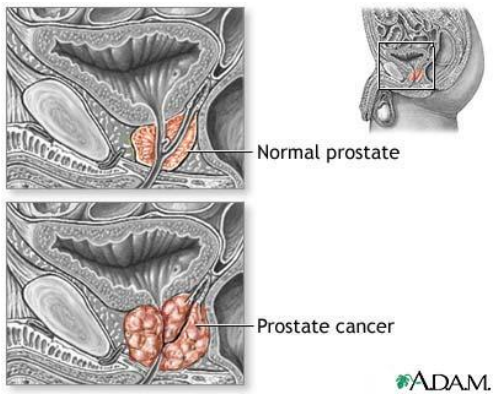


Figure 2.2. Difference between normal prostate and prostate cancer (MedLine, Adam).

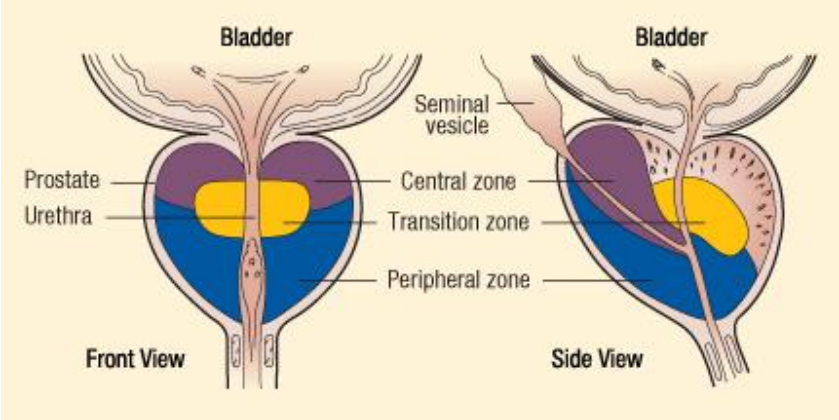


Figure 2.3. Front and side views of the prostate. In most cases, the peripheral zone (PZ) is where most malignant neoplasms develop [b8].

Carcinoma of the prostate is a type of malignant neoplasm (cancer) commonly known as prostate cancer. Carcinoma, however, is wrongly used as synonym for cancer. The definition of carcinoma is a malignant tumour (or malignant neoplasm) that arises in the epithelium (tissue that lines the skin and internal organs of the body [b5]). 'Carcinomas invade the surrounding tissues and tend to metastasize to other anatomic sites' [b8]. We shall use the term prostate cancer to refer to any type of cancer (including carcinomas) arising in the prostate. Figure 1.4 provides a clear representation of the different types of neoplasms, including carcinomas.

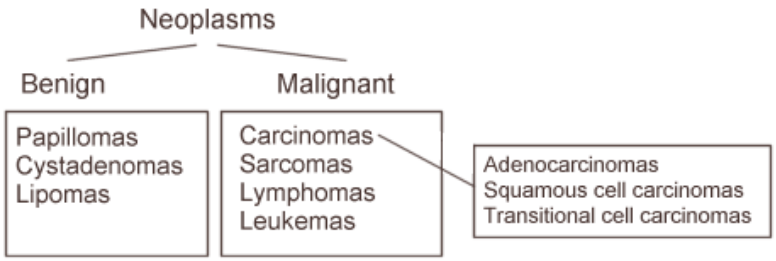


Figure 2.4. Types of neoplasms. Examples of benign and malignant neoplasms.

Advanced prostate cancer can be of two types. Locally advanced prostate cancer has spread from the prostate gland to neighbour tissues or glands. Metastatic prostate cancer has spread to the lymph nodes or other parts of the body (metastasized). Prostate cancer tends to spread to the bones and lymph nodes (glands involved in protecting the body against infection) [b10]. Bone metastases, therefore, are likely to occur in a patient with advanced prostatic cancer. 'The spine, pelvis, ribs, and bones of the arm and thigh are the most common sites of bone metastases' [b10]. Hormone escape happens in metastatic prostate cancer if the tumour grows back after a certain period of time when treatment is being administered. In this stage there is no control over the growth of the tumour, and therefore no effective treatment. This is the greatest clinical problem in men's cancer. Palliative care (treatment aimed at reducing the severity of the symptoms) is administered in this situation.

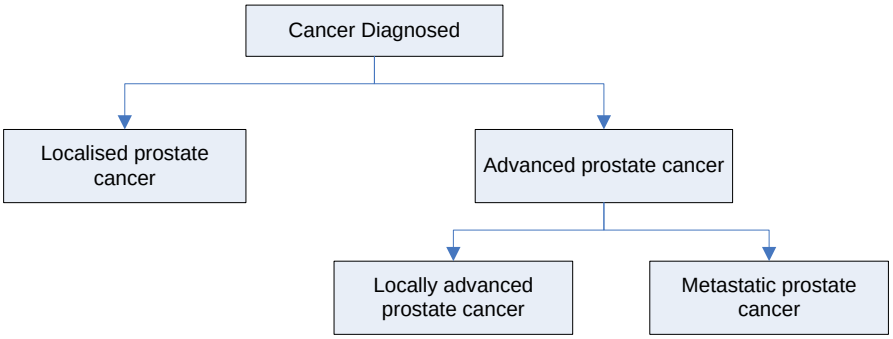


Figure 2.5. Types of prostate cancer: localised, locally advanced, and metastatic.

The standard international codes that are regularly used for this disease are listed in figure 1.6 (SNOMED is most common within the NHS).

UMLS (Unified Medical Language System)	**ICD-9** (International Statistical Classification of Diseases 9th edition)	**ICD-10** (International Statistical Classification of Diseases 10th edition)	**SNOMED Clinical Terms**
Term Name: CA – Carcinoma of prostate **Concept Unique Identifier:** C0600139 **Term Unique Identifier:** L0533491	**Term Name:** Malignant neoplasm of prostate **Code:** 185	**Term Name:** Malignant neoplasm of prostate **Code:** C61	**Term Name:** CA – Carcinoma of prostate **AUI:** A3002443 **TTY:** SY **ID:** 254900004

Figure 2.6. International standard codes for prostate cancer.

B.1.1 Etiology and Incidence

The exact causes of prostate cancer are unknown. However, several studies have been carried out and it is possible to set a number of risk factors that contribute (or not) to the incidence of prostate cancer in men:

- **Age:** The incidence of prostate cancer increases with age. More than 90% of men diagnosed with prostate cancer are older than 50 years [b10].

- **Race:** A large percentage of prostate cancer occurs in black elderly men. The lowest incidence occurs in vegetarians and Japanese men [b1].

- **Family History (Genetic Factor):** There is high incidence of prostate cancer in men whose father or brother has had the disease. There is also an increased risk for prostate cancer when breast cancer is or was present in close relatives. ‘Alteration of genes on chromosome 1, 17, and the X chromosome have been found in some patients with a family history of prostate

cancer. The hereditary prostate cancer 1 (HPC1) gene and the predisposing for cancer of the prostate (PCAP) gene are on chromosome 1, while the human prostate cancer gene is on the X chromosome' [b9].

- **Environmental:** In general, exposure to carcinogenic substances increases the risk for prostate cancer. 'Some studies show an increased chance for prostate cancer in men who are farmers, or those exposed to the metal cadmium while making batteries, welding, or electroplating' [b7].

- **Diet:** Men who eat a high-fat diet may have a greater chance of developing prostate cancer [b7]. The consumption of dietary fiber, soy protein, carotenoids containing lycopenes (found in tomato juice and tomato paste), and vitamin E with selenium, have been found to inhibit the growth of prostate cancer [b7].

The number of diagnosed cases of prostate cancer has risen significantly over the last two decades [b2]. Prostate cancer is rare in men under 50 years old. The median age of both diagnosis and mortality is 75 years old. Figure 2.7 shows in detail the number of new cases and the number of deaths from prostate cancer in the UK.

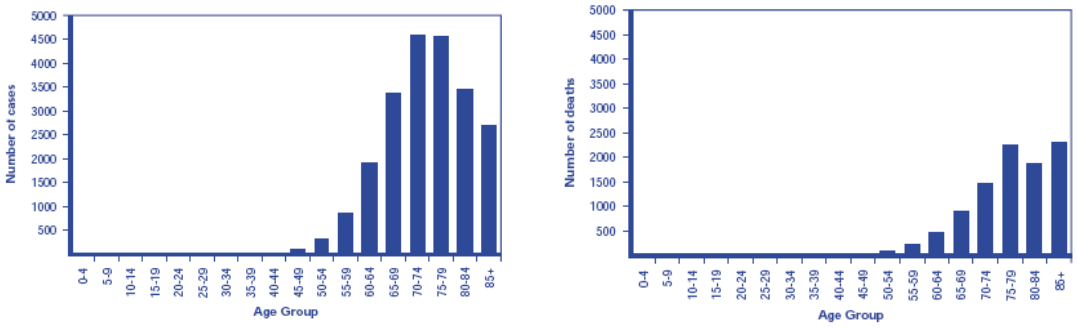


Figure 2.7 Number of new cases of prostate cancer by age in 1997 (top). Number of deaths from prostate cancer by age in 1999 (bottom) [b3].

B.1.2 Diagnosis, Differentials and Screening

Diagnosis is the process of identifying a disease by its signs, symptoms, and results of various diagnostic procedures. Screening may be described as the process of early diagnosis of a disease, although, it does not provide a conclusive diagnosis. A better definition of screening is the process of looking for evidence of a particular disease in people with no symptoms. Because there are usually no specific signs of early prostate cancer, screening is a very important test for men approaching andropause (age range from 40 to 55). It is therefore important for men on this age to screen for prostate cancer every two to three years.

The relevant screening tools that can be major indicators for the diagnosis and management of prostate cancer are:

- **Digital Rectal Examination (DRE)**

  This procedure consists of palpating the prostate from inside the rectum. The physician inserts a gloved finger into the rectum through the anus and palpates the prostate gland [b10]. The physician then tries to feel the size and shape of the prostate and searches for hard or lumpy areas, which may indicate cancer. The DRE procedure only allows the palpation of a certain area of the prostate, the one felt through the rectum (the peripheral zone). This is the area where most prostate cancers arise [b2].

- **Prostate-Specific Antigen (PSA) test**

  The PSA test measures the levels of a protein (PSA, secreted by luminal cells in the epithelium, in the prostate [b12]) in a blood sample. The PSA test can also be referred as: kallikrein III, seminin, semenogelase, γ-seminoprotein and P-30 antigen. 'PSA is produced for the ejaculate where it liquefies the semen and allows sperm to 'swim' freely' [b2].

  There are (arguably) three levels of PSA measurement to consider:
  Normal: from 0 to 4 nanograms per milliliter (ng/mL)
  Slightly elevated (intermediate): from 4 to 10 ng/mL
  Elevated: 10 ng/mL and above

  An age specific range may also be used. Patients under 60 years old have a threshold of 3 ng/mL, patients between 60 and 69 a threshold of 4 ng/mL, and patients over 70 have a threshold of 5 ng/mL. Values above these thresholds may indicate prostate cancer.

  Results above 4 ng/mL may indicate: BPH, prostate cancer, prostatitis (inflamation of the prostate), urinary tract infection (cystitis), recent urinary tract operation. A PSA over 4 ng/mL normally requires further evaluation with a prostate biopsy or other relevant medical tests (cystoscopy, transrectal ultrasound, etc). To maximize the accuracy of the results, patients are recommended to avoid ejaculation and physical activity affecting the prostate (e.g. bicycle riding) for two days before the test. PSA testing may not help a man with a cancer that has already spread to other parts of his body before being detected. Correlations between PSA levels and other biochemistry tests may provide better clinical information for the screening, diagnosis or staging of prostate cancer. For instance, suggestions have been made that men with low testosterone levels might have higher rates of prostate cancer [79], and this would imply a negative correlation (PSA-Testosterone).

Another test that can be carried out for screening purposes is a transrectal ultrasound (TRUS) [b11], further explored in section 3.2. DRE and PSA tests are not only important for screening purposes; they also help diagnosing prostate cancer. In order to diagnose prostate cancer, a physician should order further examinations, explored in section 3.2.

Screening

The basic purpose of screening for a given disease is to help separating large groups of healthy individuals from groups of individuals who have a high probability of having the disease [b11]. In the UK, the National Health Service (NHS) coordinates several national screening programmes but prostate cancer is not one of them. The NHS Cancer Screening states several reasons for not implementing a programme for prostate cancer [b1]: the natural history of the disease is not well understood; it is hard to

recognise at an early stage as even DRE and PSA examinations may throw false positives; it may introduce false alarms and anxiety; there is no effective examination to screen for prostate cancer. Therefore, instead of having a national screening programme, the NHS introduced information packs for general practitioners (GPs) so that they can assist their concerned patients. Even though the natural history is not yet well understood, Whitmore [b8] pioneered a study on this subject. He defines natural history of prostatic cancer as the 'evolving clinical and pathologic manifestations of the disease in the untreated host'. In 1973, Whitmore concluded that it was 'morally and ethically impossible' to 'obtain data regarding the natural history of the disease which might resolve some of the existing questions regarding therapy'. 'Screening for prostate cancer is controversial because it is not clear if the benefits of screening outweigh the risks of follow-up diagnostic tests and cancer treatment' [b2]. 'Concerned middle-aged patients, or in fact most men reaching andropause, should be screened for prostate cancer every two to three years' [b2]. Doctors may decide not to screen for prostate cancer in men older than 75 years old because of 'concern that prostate cancer therapy may do more harm than good as age progresses and life expectancy decreases' [b2]. Prostate cancer grows very slowly and in most cases is never detected in early stages. Most prostate cancers never grow to the point where they cause symptoms, and most men die of other causes than cancer [b2]. According to the counselling leaflet on the PSA test published by the NHS Cancer Screening Programmes, for every 1000 men (ages between 50 and 70) who have had the PSA test, about 100 men have a raised PSA level. Of these 100, about 74 have a negative result in biopsy. Figure 2.8 illustrates these numbers.

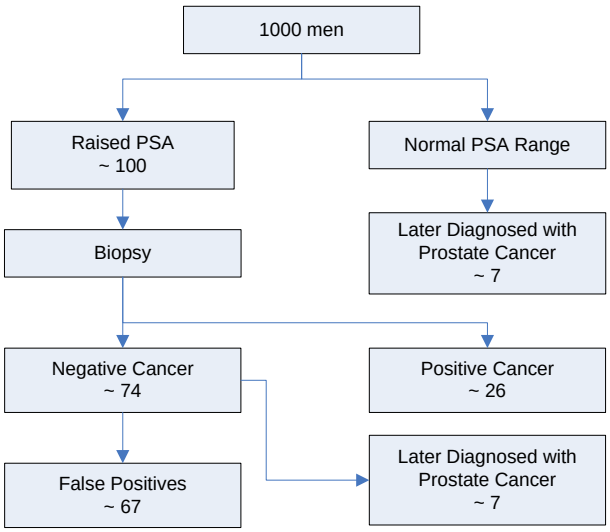


Figure 2.8. The PSA Test in 1000 men, from the counselling leaflet by the NHS Cancer Screening Programmes.

Differential diagnosis

Differential diagnosis can be defined as the diagnosis of a condition whose signs and/or symptoms are shared by various other conditions. The most common differential diagnoses for prostate cancer are:

- Benign prostatic hypertrophy (BPH)
- Calculi (stones that may form in either kidney or bladder)
- Prostatic cysts (closed sac, normal or abnormal, that contains liquid or semi-solid substance)
- Prostatic tuberculosis (rare, contagious bacterial disease of the prostate)

- Prostatitis (inflammation of the prostate gland)

Differential diagnoses from radiological findings are [b4]:

- Paget Disease

From hypoechoic (darker regions in ultrasound images) area in prostate peripheral zone (PZ) on Transrectal ultrasonography:

- Cancer
- Prostatitis
- Prostatic infarct
- Prostatic intraepithelial neoplasia

From multiple sclerotic lesions (scar tissue) within bone:

- Developmental: Bone islands, osteopoikilosis, Voorhoeve disease (osteopathia striata), and tuberose sclerosis (or tuberous sclerosis)
- Neoplastic (abnormal tissues): Metastases, lymphoma, osteomata, and myeloma
- Paget disease
- Vascular: Bone infarcts

B.1.3 Signs and Symptoms

The following local symptoms are reported in a great majority of patients with one or more of the following:

- Urinary frequency (especially at night)
- Weak or interrupted flow of urine
- Difficulty urinating or holding back urine
- Inability to urinate
- Pain or burning when urinating
- Blood in the urine (hematuria) or semen
- Painful ejaculation
- Difficulty having an erection
- Nagging pain in the lower back, hips, or pelvis

The following symptoms are reported in patients with metastatic prostate cancer:

- bone pain, with or without pathologic fracture (because prostate cancer when metastatic, has a strong predilection for bone [b9])
- lower extremity pain (mainly legs and feet)
- weight loss and loss of appetite

B.1.4 Tests

DRE and PSA tests are useful as part of the initial investigations for prostate cancer. When these tests suggest the disease, further tests need to be carried out:

- **Transrectal ultrasonography (TRUS)**

  Ultrasound are released into the rectum to produce a picture of the prostate (these pictures are called sonograms). TRUS plays an important role in the early diagnosis of prostate cancer, especially when more and more patients have tumours that are not palpable with DRE exams. Studies have been carried out and determined that TRUS is also helpful for a guided biopsy of the prostate [33].

- **Intravenous pyelogram (IVP)**

  A series of X-Rays of the organs of the urinary tract. The radiologist injects a contrast agent (isotope) that once taken will travel through the urinary tract. The x-rays capture the flow of the contrast agent, determining any obstructions in the urinary tract. This test is performed to detect enlarged prostate, kidney tumours, and kidney or bladder stones.

- **Cystoscopy**

  This is a procedure where the urologist can look into the urethra and bladder using a thin lighted tube. This test is recommended to patients with one of the following conditions: urinary tract infections, hematuria (blood in urine), incontinence or over active bladder, enlarged prostate, and tumours.

- **Biopsy**
  Tissue samples are removed from the prostate to be further analysed by a pathologist. The pathologist will search for cancer cells and grade the tumour when cancer is present. This grade can suggest how fast the tumour is likely to grow, and appropriate treatment can then be recommended. The most common grading method is called the Gleason system, rating the tumours in a range from 2 to 10. The Gleason score combines two grades that range from 1 to 5. These grades are given from the two most common patterns observed in the biopsy results. A Gleason score from 7 to 10 indicates the worst prognosis: fast tumour growth [34]. Furthermore, a Gleason score of 7 or a PSA value of 10 ng/mL also puts patients at risk for progressing to advanced cancer.

The following tests can help detecting bone metastases:

- **Bone scan (radionuclide scan)**

  Bone scans can help physicians looking for the (metastatic) spread of cancer in bones. A bone scan is indicated in patients with prostate cancer who have symptoms suggesting bony metastases [b10].

  Similarly to an IVP, a bone scan requires the injection of a harmless radioactive substance (radionuclide) into the blood stream. This substance (contrast agent) collects in bones, especially in areas where bones are repairing or fractured. Activity in the bone scan may not be observed until 5 years after metastases has occurred. Therefore, bone scans with negative results do not prove the absence of metastasis [b10]. Figure 2.9 shows a normal whole body scan on

the left. The image on the right shows the spread of prostate cancer to the spine.

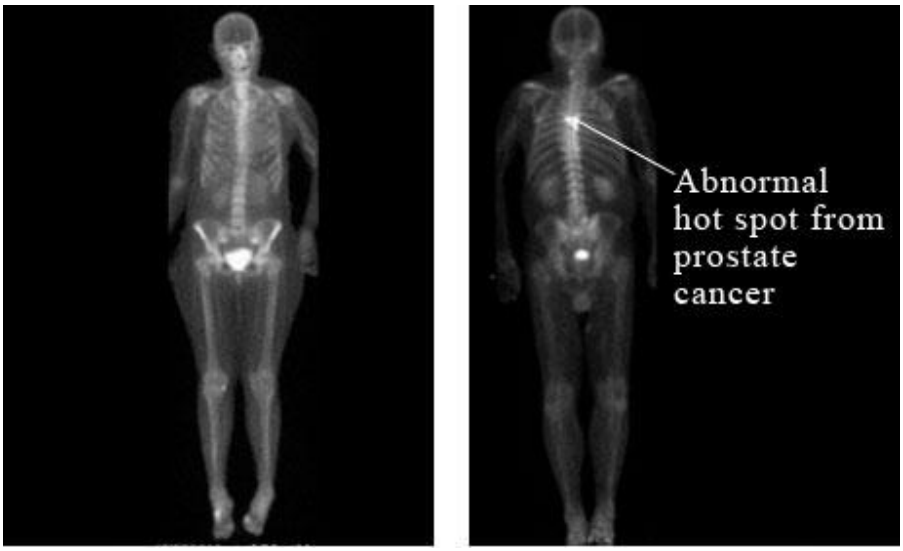


Figure 2.9. Normal body bone scan (on the left), and abnormal hot spot from prostate cancer in the spine (on the right). Intermountain Medical Imaging, Boise, Idaho.

- **Computed tomography (CT) or Computed axial tomography (CAT)**

  A CAT scan of the abdomen and pelvis in patients suggested to have locally advanced disease may give an indication of extracapsular extension (tumour grows outside the prostate capsule), seminal vesical involvement, pelvic lymph node enlargement and liver metastases [b10]. A CAT scan takes several x-rays from different angles, compiling a final 3D image. The image data derived from a CAT scan is particularly useful as it can show different types of tissue: soft tissue, bone, and blood vessels. CT scans are unable to detect small bone metastases and therefore are not often used to assess metastatic bone disease [b7].

- **Magnetic Resonance Imaging (MRI)**

  MRI uses a large magnet, radio waves, and a computer to produce three-dimensional images of internal body structures. In prostate cancer, an MRI is used to examine the prostate and its surroundings (including lymph nodes) to differentiate benign and malignant areas.

- **Bone Biopsy**

  Similarly to a biopsy of the prostate tissues, a bone biopsy can be carried out to establish a diagnosis of bone metastases.

- **Calcium and Alkaline Phosphatase**

  Bony metastases can lead to hypercalcemia (elevated calcium in the blood) as they destroy the bone, however, calcium levels may be raised for other reasons and a supplementary imaging test would still be needed to confirm the diagnosis. Similarly, alkaline phosphatase may be

increased due to the presence of metastases in the bone, however, this enzyme is produced in the liver and it may also indicate liver conditions.

B.1.5 Staging and Planning Treatment

The most commonly used staging method is the international tumour, node, metastasis (TNM) staging system. When interpreting the stages, it is important to appreciate that prostate cancer may not progress in a sequential manner. In the primary tumour (T) stage, the numbers indicate the size of the tumour.

The TNM system has the following stages:

Primary Tumour (T)		
TX	**Primary tumour cannot be assessed**	
T0	**No evidence of primary tumour**	
T1	**Not palpable or visible by imaging**	
	T1a	Tumour incidental (irrelevant) histologic (microscopic) finding in 5% or less of tissue resected (in prostatectomy)
	T1b	Tumour incidental histologic finding in more than 5% of tissue resected
	T1c	Tumour identified by needle biopsy (that was carried out due to elevated PSA)
	Common treatment: Prostatectomy or radiation therapy may be chosen. Hormonal therapy may also be used before, during, or after prostatectomy or radiation. In some cases, the physician may decide to wait and monitor mild symptoms (watchful waiting).	
T2	**Confined within prostate**	

	T2a	Tumour involves one-half of 1 lobe or less
	T2b	Tumour involves more than one-half of 1 lobe but not both lobes
	T2c	Tumour involves both lobes
	Common treatment: Prostatectomy and radiation therapy. Hormonal therapy may be used before, during, or after prostatectomy or radiation.	
T3	**Through prostatic capsule**	
	T3a	Extracapsular extension (unilateral or bilateral)
	T3b	Tumour invades seminal vesicle(s)
	Common treatment: Prostatectomy, radiation therapy, or both. Hormonal therapy may be used before, during, or after prostatectomy or radiation.	
T4	Tumour is fixed or invades adjacent structures other than seminal vesicles: bladder neck, external sphincter, rectum, levator muscles, and/or pelvic wall	
	Common treatment: Same as in T3. However, prostatectomy is less frequent.	

Lymph node involvement (N)		
NX	Regional lymph nodes not assessed	
N0	No regional lymph node metastasis (lymph nodes confined to the true pelvis)	
N1	Metastasis in regional lymph node(s)	
Common treatment: Hormonal therapy is generally used. Prostatectomy or radiation may be used with hormonal therapy. Chemotherapy may be used later if hormonal therapy is no longer working.		
Metastases (M)		
MX	Distant metastasis cannot be assessed (not evaluated by any modality)	
M0	No distant metastasis	
M1	Distant metastasis	
	M1a	Non-regional lymph node(s)
	M1b	Spread to Bone(s)
	M1c	Other site(s) with or without bone disease
Common treatment: Hormonal therapy is generally used. Chemotherapy may be used later if hormonal therapy is no longer working.		

The following excerpt from Dan Theodorescu *et al.* [b9] seems important as it states the current natural history of the prostate cancer.

> *'The natural history is still relatively unknown, and many aspects of progression are poorly understood. Symptoms or abnormal DRE findings in the pre-PSA era only brought 40-50% of patients with prostate cancer to medical attention, and these patients usually had locally advanced disease. The advent of PSA testing has helped identify patients with less-advanced, organ-confined disease.*
>
> *Evidence suggests that most prostate cancers are multifocal and heterogeneous. Cancers can start in the transitional zone or, more commonly, the peripheral zone. When these cancers are locally invasive, the transitional zone tumours spread to the bladder neck, while the peripheral zone tumours extend into the ejaculatory ducts and seminal vesicles. Penetration through the prostatic capsule and along the perineural or vascular spaces is a relatively late event.*
>
> *The mechanism for distant metastasis is poorly understood. The cancer spreads to bone early, occasionally without significant lymphadenopathy* [b9].

B.1.6 Treatment Options

The common treatments for each tumour stage are stated above, in section 4. However, it is important to understand them in more detail. Treatment options that may be used alone or in combinations are:

- **Orchidectomy**

  The removal of both testicles to reduce the amount of testosterone produced.

- **Prostatectomy**

  The surgical removal of the prostate, along with the tumour. If the tumour has not spread outside of the prostate surgery may be the best option to treat the disease [b37].

- **Hormone treatments (androgen ablation or withdrawal)**

  A female hormone (estrogen) may be prescribed to help treat prostate cancer. The British National Formulary (BNF 45) states that 'metastatic cancer of the prostate usually responds to hormonal treatment aimed at androgen depletion' [b35]. The BNF 45 also states that occasionally patients respond to other hormone manipulation such as anti-androgen (cyproterone acetate, flutamide, bicalutamide). The BNF 45 also states that gonadorelin analogues are as effective as orchidectomy. However, they have a large list of side effects, including: sexual dysfunction, sweating, rashes, headaches, visual disturbances, dizziness, hair loss, gastro-intestinal disturbances, weight changes, sleep disorders, and mood changes amongst others.

- **Chemotherapy**

  This type of therapy can be defined as 'the prevention or treatment of a disease by the use of chemical substances'. Essentially, chemotherapy is the use of drugs to kill or reduce the growth of cancer cells. The drugs used are typically antimetabolites [b35] that prevent cells from growing.

- **Radiotherapy (therapeutic radiology)**

  This involves the exposure of the cancerous area of the metastasized bone to radiation (gamma-rays). This technique may shrink the tumour or destroy some of the cancer cells. It often decreases pain associated with the spread to the bone [36]. Many forms of cancer are destroyed using this method.

- **Surveillance (watchful waiting or active monitoring)**

  Instead of active treatment, the physician will monitor the patient's condition by performing regular tests such as PSA, DRE, complete blood count (CBC) to look for signs of blood in urine (anemia), and general signs that may indicate a progression or regression of the disease. This process may be recommended in early stages of the disease or whenever the patient is not expected to tolerate other treatments.

- **Brachytherapy**

  This procedure follows the same reasoning as general radiation therapy where ionising radiation is used to kill prostate cancer cells. However, this specific method involves the insertion of radioactive seeds (implants) into the prostate. The long term outcomes and effectiveness of this treatment are not yet known [b3].

- **Cryosurgery**

  Not as effective as surgery or radiation [b5], this method uses liquid nitrogen to freeze and kill cells in the prostate. 'However, cryosurgery is potentially better than radical prostatectomy for recurrent cancer following radiation therapy. Cryosurgery is reserved for localized cancer within the prostate and in cases where conventional therapies like surgery or radiation could not be applied. Advantages of cryosurgery over general surgery include less blood loss, less pain and shorter recovery time' [b2].

- **Bone surgery**

  This type of surgery is aimed at stabilizing or fixing the bone, generally prior to radiotherapy [b7].

- **Palliation**
  Palliative care and support consists of reducing the severity of the symptoms, improving quality of life. This method does not cure the disease but it may be used in conjunction with curative therapies [b6].

B.1.7 Treatment of Metastatic Prostate Cancer

Metastatic prostate cancer is the most advanced stage of prostate cancer. The most common site for the cancer cells to spread is the bone, eventually leading to fractures. Prostate cancer may also spread to the lymph nodes [b7]. The bones that are most frequently involved are those of the 'pelvic girdle, lumbar spine, upper femurs, dorsal spine, and ribs' [b10]. Metastases in spinal regions may lead to spinal cord compression (the spine is compressed by fragments originating from a tumour) and eventually paraplegia. According to the British Association of Urological Surgeons (BAUS) when patients reach this stage they should receive 'ongoing help and support', and further diagnostic tests may be recommended to assess the spread of the cancer. Because prostate cancer commonly spreads to the bone, imaging techniques that target the bones are often used (X-rays, radionuclide bone scanning, MRI). Hormonal therapy (androgen ablation) is the standard treatment for metastatic prostate cancer. This type of treatment may control prostate cancer for many years. If the cancer stops responding to this treatment it becomes hormone resistant (hormone escape) and only palliative care can help sustaining quality of life. Figure 2.10 shows some results of an interesting study on the effects of hormonal ablation on prostate cancer [b9]. This study shows that hormonal ablation caused a decrease in size of the tumour. Apart from painkillers and hormone ablation, there are three other important (palliative) treatments for metastatic bone disease (especially used on hormone escape) [b7]:

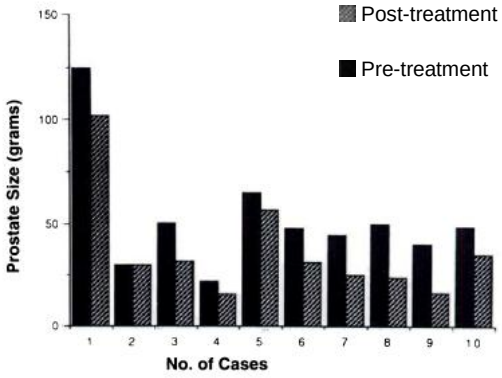


Figure 2.10. Volume of prostate gland before and after hormonal ablation [51]. Reduction was seen in nine of 10 patients. In case number two, hormone escape may justify the invariability.

- **Bisphosponates**
  This type of treatment therapy aims at preventing skeletal complications such as fractures and hypercalcaemia (high levels of calcium in the blood). Bisphosponates work by reducing the breakdown of bone caused by the cancer. Zoledronic acid is currently the only bisphosponate licensed in the UK that has proven effective [b7]. This treatment therapy can help sustaining quality of life. It is administered via a 15 minute intravenous infusion (injection into a vein) once every 28 days [b7]. Side effects may occur and are similar flu symptoms. Specialists may recommend this therapy prior the occurrence of any symptoms of metastatic bone disease.

- **Radiotherapy**
  Local radiotherapy can be recommended as palliative care to alleviate bone pain. There are two ways of administering radiotherapy: by using an X-ray machine on the painful area; or by injecting radionuclides (e.g. strontium-89) into the veins. In the latter process the injected radionuclide will find its way to the bones where it releases radiation directly onto the cancer cells [b7].

- **Surgery (orthopaedic interventions)**

  Prior to radiotherapy, orthopaedic interventions may be recommended to stabilise or fix a bone fracture. Patients can be mobile as soon as one day after an operation [b7]. This method may not be considered palliation in the sense that it fixes or stabilises fractures, but at the same time it only alleviates the symptoms of metastatic cancer bone disease, it does not cure this disease.

The BAUS guidelines for the management of metastatic prostate cancer reveal a flow diagram illustrating the recommended approach to managing this disease (Figure 2.11).

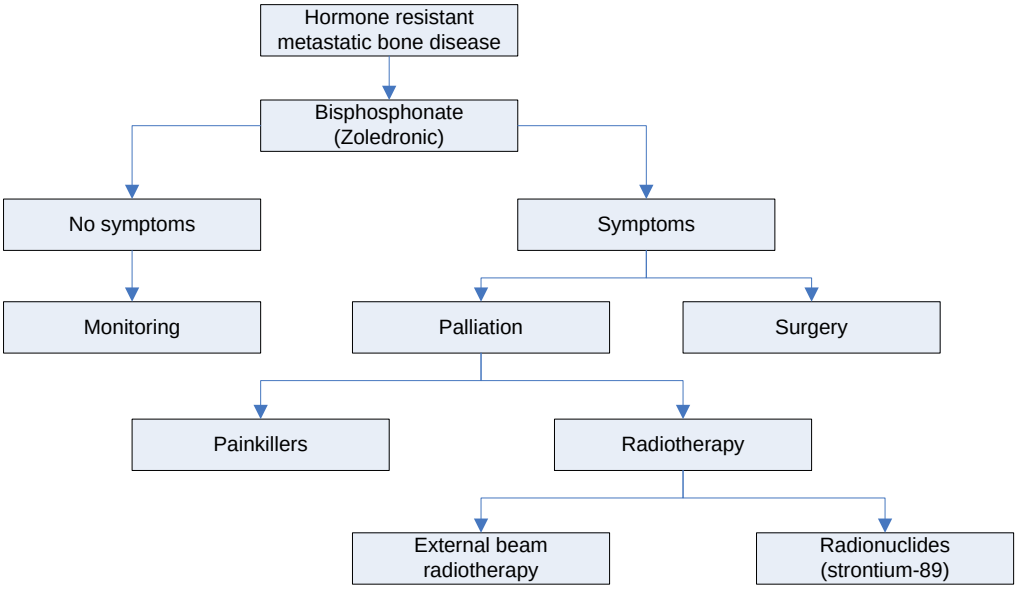


Figure 2.11. BAUS recommendations for managing hormone resistant metastatic bone disease [b7].

B.1.8 Conclusion

Prostate cancer is one of the very few types of cancer that may grow so slowly that it never produces any symptoms. Unfortunately in some other cases the tumour grows, metastases may spread, and spinal fractures may occur possibly leading to paraplegia. Androgen ablation delays the progression of this disease, but hormone resistance can soon develop in which case only palliative care can sustain quality of life.

There are many treatment options and diagnostic tests. Some may be poor to effectively assess cancer spread, some may stop the growth of the tumour for many years, and some may produce unpleasant side effects. Specialists have many choices in their hands but lack key knowledge (such as the natural history) for decision making. The use of information and knowledge present in today's electronic archives of imaging modalities, histological data, and patient record datasets may boost the understanding of this disease.

Scientists are researching other ways of detecting prostate cancer in early stages by using chemical methods and chemical information. Treatment approaches are also being perfected and new approaches such as cryosurgery may introduce more efficient and last longing control over the disease.

In Chapter 5, correlation hypotheses are briefly investigated, but they merely indicate ways in which to use the prostate cancer data collected from multiple hospital sources. This is the case of life expectancy and the most common treatment types, PSA and testosterone (as mentioned before, there is a hypothesis that high grade prostate cancer and hence PSA correlate well with low levels of testosterone), and a study on the numbers of prostate cancers metastasising to bone.

References

[b1] Prostate Cancer Risk Management, NHS Cancer Screening, (2006) http://www.cancerscreening.nhs.uk/prostate/index.html

[b2] Cancer Research UK. Prostate CancerStats 2002. London, Cancer Research UK

[b3] Watson E. et al. (2002) Cancer Research UK Booklet, NHS Cancer Screening Programmes

[b4] Clements R. (2006) Prostate Carcinoma, http://www.emedicine.com/med/topic574.htm

[b5] Bahn D.K. et al. (2002) "Targeted cryoablation of the prostate: 7-year outcomes in the primary treatment of prostate cancer". Urology 60 (2 Suppl 1): 3-11

[b6] Palliative Care, Wikipedia (2006) http://en.wikipedia.org/wiki/Palliative_care

[b7] The BAUS Guidelines for the Management of Metastatic Prostate Cancer, British Association of Urological Surgeons, (2006) http://www.prostate-cancer.org.uk/news/features/baus_1.asp

[b8] Whitmore, W.F. (1973) The natural history of prostatic cancer. Cancer; 32: 1104

[b9] Chen, M., Hricak, H., Kalbhen, C.L., et al. (1996) Hormonal ablation of prostatic cancer: effects on prostate morphology, tumor detection, and staging by endorectal coil MR imaging. AJR Am J Roentgenol; 166:1157-1163

[b10] Pollen, J.J., Shlaer, W.J. (1979) Osteoblastic response to successful treatment of metastatic cancer of the prostate. AJR Am J Roentgenol; 132(6): 927-31

[b11] Kirkels W., Rietbergen J. (1996) Screening for prostate cancer, Journal of Urological Research 25(2): 53-56

[b12] Dorkin T., Neal D. (1997) Basic science aspects of prostate cancer. Seminars in Cancer Biology 8:21–7

Appendix C - Additional Data on Pathways and Mining

1.4 System-level Paths

This section shows additional details on the system paths including the full list of the sequential pairs, the results of the Apriori, PrefixSpan and CMRules algorithms, and a detailed list of average times between systems. This section also shows the results and interpretation of the application of process mining techniques to the system paths.

System Paths: Supplementary Information

Detailed List of Sequential Pairs

The first table shows (in two parts) the complete list of all possible interactions between the systems (permutations of any two systems in the path sequence); it gives the number of paths where the sequence occurs, the overall frequency across all paths, and the ratio (overall:paths). The second table (far right) shows the first 32 (from a total of 42) most common sequence pairs at the start of a path.

Sequence	Paths	Overall	Ratio
LAB;LAB	3508	24246	6.9
LAB;HIS	2103	2134	1.0
HIS;LAB	2031	2136	1.1
RAD;LAB	1032	1194	1.2
LAB;PAS	983	2005	2.0
PAS;LAB	809	1655	2.0
PAS;HIS	692	727	1.1
LAB;RAD	680	820	1.2
OPT;HIS	632	658	1.0
OPT;PAS	625	664	1.1
ONC;LAB	580	587	1.0
LAB;ONC	549	553	1.0
LAB;OPT	504	571	1.1
PAS;OPT	496	520	1.0
ONC;RAD	484	484	1.0
HIS;ONC	437	438	1.0
HIS;OPT	404	409	1.0
LAB;IMG	350	394	1.1
IMG;LAB	344	382	1.1
OPT;LAB	324	346	1.1
HIS;PAS	291	306	1.1
PAS;PAS	241	501	2.1
HIS;RAD	237	237	1.0
HIS;IMG	207	207	1.0
PAS;RAD	151	181	1.2
RAD;PAS	143	163	1.1
IMG;RAD	105	113	1.1
ONC;HIS	104	104	1.0
RAD;RAD	91	117	1.3
PAS;IMG	81	88	1.1
HIS;HIS	75	77	1.0
IMG;PAS	60	63	1.1

Sequence (cont.)	Paths	Overall	Ratio
ONC;PAS	55	55	1.0
RAD;IMG	42	44	1.0
IMG;ONC	39	39	1.0
PAS;ONC	38	38	1.0
ONC;IMG	33	33	1.0
IMG;OPT	33	33	1.0
IMG;IMG	28	35	1.3
ONC;OPT	26	26	1.0
LAB;ORT	19	19	1.0
OPT;OPT	17	20	1.2
ORT;LAB	10	10	1.0
RAD;ONC	9	9	1.0
IMG;HIS	8	8	1.0
RAD;HIS	7	7	1.0
OPT;ONC	7	7	1.0
IMG;ORT	4	5	1.3
ONC;ONC	4	4	1.0
RAD;OPT	3	3	1.0
OPT;RAD	3	3	1.0
RAD;ORT	3	3	1.0
PAS;ORT	2	2	1.0
ORT;RAD	2	2	1.0
OPT;IMG	1	1	1.0
ORT;PAS	1	1	1.0
ORT;IMG	1	1	1.0
HIS;ORT	0	0	-
ORT;HIS	0	0	-
ONC;ORT	0	0	-
OPT;ORT	0	0	-
ORT;ONC	0	0	-
ORT;OPT	0	0	-
ORT;ORT	0	0	-

Start Seq.	Paths
LAB, LAB	1743
LAB, HIS	1111
LAB, PAS	201
HIS, LAB	153
ONC, RAD	138
LAB, ONC	116
LAB, OPT	116
ONC, LAB	101
PAS, LAB	93
HIS, RAD	90
OPT, PAS	71
PAS, OPT	71
HIS, ONC	66
HIS, OPT	52
OPT, HIS	48
PAS, HIS	37
HIS, PAS	37
HIS, HIS	30
OPT, LAB	24
ONC, HIS	21
HIS, IMG	17
PAS, PAS	13
LAB, IMG	11
PAS, RAD	10
RAD, PAS	8
LAB, RAD	8
ONC, PAS	7
PAS, ONC	7
ONC, OPT	5
IMG, LAB	4
RAD, RAD	4
IMG, ONC	4

Results of Apriori Algorithm for Mining Frequently Closed Association Rules in System Paths
The table shows the top 108 association rules with respective Support and Confidence.

Rule	Supp.	Conf.
PAS, IMG, HIS → LAB	203	99.51%
OPT, IMG → LAB	166	99.40%
OPT, IMG, HIS → LAB	159	99.38%
PAS, OPT, IMG → LAB	145	99.32%
PAS, OPT, IMG, HIS → LAB	138	99.28%
PAS, RAD, IMG → LAB	186	98.41%
PAS, IMG → LAB	341	97.99%
PAS, RAD, ONC → LAB	328	97.91%
PAS, RAD, ONC, HIS → LAB	186	97.89%
RAD, ONC, IMG → LAB	171	97.71%
RAD, IMG, HIS → LAB	238	97.54%
IMG, HIS → LAB	485	97.39%
RAD, ONC, IMG, HIS → LAB	111	97.37%
RAD, ONC, OPT, HIS → LAB	145	97.32%
RAD, IMG → LAB	345	97.18%
ONC, IMG → LAB	232	97.07%
IMG → LAB	650	97.01%
PAS, RAD, ONC, OPT, HIS → LAB	130	97.01%
RAD, ONC, OPT → LAB	161	96.99%
PAS, ONC → LAB	466	96.88%
PAS, RAD, HIS → LAB	347	96.66%
PAS, RAD, ONC, OPT → LAB	144	96.64%
PAS, ONC, HIS → LAB	286	96.62%
ONC, IMG, HIS → LAB	155	96.27%
ONC, OPT, HIS → LAB	246	96.09%
ONC, OPT → LAB	265	96.01%
OPT, IMG → HIS	160	95.81%
PAS, RAD, OPT, HIS → LAB	250	95.79%
LAB, OPT, IMG → HIS	159	95.78%
PAS, RAD, OPT → LAB	271	95.76%
RAD, OPT → LAB	303	95.58%
PAS, ONC, OPT, HIS → LAB	216	95.58%
RAD, OPT, HIS → LAB	280	95.56%
PAS, ONC, OPT → LAB	231	95.45%
OPT, IMG → LAB, HIS	159	95.21%
PAS, OPT, IMG → HIS	139	95.21%

Rule (cont.)	Supp.	Conf.
PAS, LAB, OPT, IMG → HIS	138	95.17%
PAS, RAD → LAB	590	95.16%
PAS, OPT, IMG → LAB, HIS	138	94.52%
LAB, OPT → HIS	1291	94.30%
OPT → HIS	1394	94.25%
RAD, ONC, HIS → LAB	482	94.14%
PAS → LAB	1869	93.87%
PAS, LAB, OPT → HIS	1073	93.55%
PAS, LAB, ONC, OPT → HIS	216	93.51%
ONC, HIS → LAB	762	93.50%
PAS, OPT → HIS	1162	93.48%
PAS, HIS → LAB	1291	93.42%
PAS, ONC, OPT → HIS	226	93.39%
HIS → LAB	3235	92.91%
LAB, ONC, OPT → HIS	246	92.83%
ONC, OPT → HIS	256	92.75%
OPT, HIS → LAB	1291	92.61%
OPT → LAB	1369	92.56%
RAD, OPT → HIS	293	92.43%
LAB, RAD, OPT → HIS	280	92.41%
PAS, OPT, HIS → LAB	1073	92.34%
PAS, OPT → LAB	1147	92.28%
PAS, LAB, RAD, OPT → HIS	250	92.25%
PAS, RAD, OPT → HIS	261	92.23%
PAS, LAB, RAD, ONC, OPT → HIS	130	90.28%
LAB, RAD, ONC, OPT → HIS	145	90.06%
RAD, ONC, OPT, HIS → PAS	134	89.93%
PAS, RAD, ONC, OPT → HIS	134	89.93%
RAD, ONC, OPT → PAS	149	89.76%
RAD, ONC, OPT → HIS	149	89.76%
LAB, RAD, ONC, OPT, HIS → PAS	130	89.66%
LAB, RAD, ONC, OPT → PAS	144	89.44%
LAB, RAD, OPT → PAS	271	89.44%
LAB, RAD, OPT, HIS → PAS	250	89.29%
RAD, OPT → PAS	283	89.27%
PAS, ONC, OPT → LAB, HIS	216	89.26%

Rule (cont.)	Supp.	Conf.
ONC, OPT → LAB, HIS	246	89.13%
RAD, OPT, HIS → PAS	261	89.08%
RAD, HIS → LAB	914	88.57%
PAS, RAD, OPT → LAB, HIS	250	88.34%
RAD, OPT → LAB, HIS	280	88.33%
ONC, OPT, HIS → PAS	226	88.28%
LAB, ONC, OPT, HIS → PAS	216	87.80%
ONC, OPT → PAS	242	87.68%
OPT, IMG → PAS	146	87.43%
LAB, OPT, IMG → PAS	145	87.35%
RAD, ONC, OPT → LAB, HIS	145	87.35%
OPT → LAB, HIS	1291	87.29%
PAS, RAD, ONC, OPT → LAB, HIS	130	87.25%
RAD, ONC, OPT, HIS → PAS, LAB	130	87.25%
LAB, ONC, OPT → PAS	231	87.17%
OPT, IMG, HIS → PAS	139	86.88%
OPT, IMG → PAS, LAB	145	86.83%
LAB, OPT, IMG, HIS → PAS	138	86.79%
RAD, ONC, OPT → PAS, LAB	144	86.75%
ONC → LAB	1166	86.63%
PAS, OPT → LAB, HIS	1073	86.32%
OPT, IMG, HIS → PAS, LAB	138	86.25%
RAD, OPT → PAS, LAB	271	85.49%
RAD, OPT, HIS → PAS, LAB	250	85.32%
ONC, OPT, HIS → PAS, LAB	216	84.38%
PAS, HIS → OPT	1162	84.08%
OPT → PAS	1243	84.04%
LAB, OPT → PAS	1147	83.78%
ONC, OPT → PAS, LAB	231	83.70%
OPT, HIS → PAS	1162	83.36%
OPT, IMG → PAS, HIS	139	83.23%
RAD → LAB	1275	83.17%
LAB, OPT, IMG → PAS, HIS	138	83.13%
PAS, LAB, HIS → OPT	1073	83.11%
LAB, OPT, HIS → PAS	1073	83.11%
RAD, ONC → LAB	742	82.72%

Results of PrefixSpan for Mining Frequently Closed Sequential Patterns in System Paths
The table shows the top 72 sequential patterns with itemset length (L) > 2 and the number of unique systems (S), ordered by support.

Sequence	Supp.	L	S
‹LAB, LAB, LAB›	3390	3	1
‹LAB, LAB, LAB, LAB›	3068	4	1
‹LAB, LAB, LAB, LAB, LAB›	2769	5	1
‹HIS, LAB, LAB›	2545	3	2
‹LAB, HIS, LAB›	2485	3	2
‹LAB, LAB, LAB, LAB, LAB, LAB›	2483	6	1
‹HIS, LAB, LAB, LAB›	2232	4	2
‹LAB, HIS, LAB, LAB›	2209	4	2
‹LAB, LAB, LAB, LAB, LAB, LAB, LAB›	2204	7	1
‹HIS, LAB, LAB, LAB, LAB›	1966	5	2
‹LAB, HIS, LAB, LAB, LAB›	1956	5	2
‹LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	1948	8	1
‹LAB, HIS, LAB, LAB, LAB, LAB›	1717	6	2
‹HIS, LAB, LAB, LAB, LAB, LAB›	1690	6	2
‹LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	1683	9	1
‹LAB, LAB, HIS›	1498	3	2
‹LAB, HIS, LAB, LAB, LAB, LAB, LAB›	1464	7	2
‹LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	1459	10	1
‹HIS, LAB, LAB, LAB, LAB, LAB, LAB›	1443	7	2
‹LAB, LAB, HIS, LAB›	1304	4	2
‹LAB, HIS, LAB, LAB, LAB, LAB, LAB, LAB›	1245	8	2
‹LAB, LAB, PAS›	1238	3	2
‹HIS, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	1236	8	2
‹LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	1232	11	1
‹LAB, LAB, HIS, LAB, LAB›	1140	5	2
‹PAS, LAB, LAB›	1074	3	2
‹LAB, HIS, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	1064	9	2
‹LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	1048	12	1
‹LAB, PAS, LAB›	1045	3	2
‹HIS, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	1037	9	2
‹LAB, RAD, LAB›	1020	3	2
‹LAB, LAB, HIS, LAB, LAB, LAB›	999	6	2
‹LAB, LAB, LAB, PAS›	978	4	2
‹OPT, LAB, LAB›	955	3	2
‹LAB, LAB, RAD›	951	3	2
‹OPT, HIS, LAB›	935	3	3

Sequence (cont.)	Supp.	L	S
‹PAS, LAB, LAB, LAB›	929	4	2
‹LAB, OPT, LAB›	929	3	2
‹RAD, LAB, LAB›	914	3	2
‹LAB, LAB, LAB, HIS›	907	4	2
‹LAB, PAS, LAB, LAB›	891	4	2
‹LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	890	13	1
‹LAB, HIS, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	880	10	2
‹LAB, OPT, HIS›	874	3	3
‹LAB, LAB, HIS, LAB, LAB, LAB, LAB›	869	7	2
‹LAB, RAD, LAB, LAB›	868	4	2
‹HIS, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	861	10	2
‹ONC, LAB, LAB›	840	3	2
‹PAS, LAB, LAB, LAB, LAB›	822	5	2
‹OPT, LAB, LAB, LAB›	816	4	2
‹LAB, LAB, RAD, LAB›	811	4	2
‹LAB, OPT, LAB, LAB›	810	4	2
‹OPT, HIS, LAB, LAB›	808	4	3
‹HIS, RAD, LAB›	805	3	3
‹LAB, LAB, OPT›	805	3	2
‹LAB, LAB, PAS, LAB›	789	4	2
‹LAB, LAB, LAB, HIS, LAB›	784	5	2
‹LAB, PAS, LAB, LAB, LAB›	784	5	2
‹LAB, OPT, HIS, LAB›	784	4	3
‹LAB, HIS, RAD›	782	3	3
‹LAB, LAB, LAB, LAB, PAS›	780	5	2
‹RAD, LAB, LAB, LAB›	780	4	2
‹LAB, ONC, LAB›	765	3	2
‹LAB, LAB, LAB, RAD›	762	4	2
‹ONC, LAB, LAB, LAB›	749	4	2
‹LAB, RAD, LAB, LAB, LAB›	746	5	2
‹LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	735	14	1
‹LAB, LAB, HIS, LAB, LAB, LAB, LAB, LAB›	733	8	2
‹LAB, HIS, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	731	11	2
‹PAS, LAB, LAB, LAB, LAB, LAB›	718	6	2
‹LAB, HIS, PAS›	718	3	3
‹HIS, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB, LAB›	715	11	2

Results of CMRules Algorithm for Mining Sequential Patterns in System Paths
The table shows all 67 sequential patterns (minsup = 10%, confidence = 10%) and it is ordered by confidence.

Rule	Supp.	Conf.
{RAD,HIS} → {LAB}	702	96.34%
{RAD,ONC} → {LAB}	569	95.58%
{PAS,HIS} → {LAB}	1059	94.15%
{RAD} → {LAB}	959	94.13%
{ONC,HIS} → {LAB}	596	93.89%
{OPT,HIS} → {LAB}	1058	93.58%
{PAS,OPT,HIS} → {LAB}	895	93.55%
{HIS} → {LAB}	2598	91.83%
{PAS} → {LAB}	1464	91.83%
{PAS,LAB,OPT} → {HIS}	688	91.77%
{LAB,OPT} → {HIS}	831	91.39%
{PAS,OPT} → {LAB}	950	91.03%
{OPT} → {LAB}	1139	90.75%
{PAS,OPT} → {HIS}	799	89.30%
{OPT} → {HIS}	981	88.62%
{ONC} → {LAB}	921	84.98%
{OPT,HIS} → {PAS}	515	83.91%
{LAB,OPT,HIS} → {PAS}	447	83.88%
{PAS,OPT} → {LAB,HIS}	688	83.54%
{LAB,OPT} → {PAS}	634	83.53%
{OPT} → {PAS}	768	83.28%
{OPT} → {LAB,HIS}	848	82.93%
{LAB,OPT} → {PAS,HIS}	475	76.66%
{OPT} → {PAS,LAB}	572	75.81%
{LAB} → {HIS}	2621	75.12%
{OPT} → {PAS,HIS}	542	74.37%
{PAS,LAB} → {HIS}	828	69.83%
{OPT} → {PAS,LAB,HIS}	455	69.57%
{PAS} → {HIS}	989	68.10%
{PAS} → {OPT}	721	64.49%
{PAS} → {LAB,HIS}	872	64.12%
{PAS,LAB} → {OPT}	584	63.93%
{LAB,ONC} → {RAD}	614	59.67%
{PAS} → {LAB,OPT}	657	58.70%
{PAS,LAB} → {OPT,HIS}	518	58.67%
{PAS} → {OPT,HIS}	629	57.59%

Rule (cont.)	Supp.	Conf.
{PAS} → {LAB,OPT,HIS}	590	53.87%
{ONC} → {RAD}	668	53.05%
{ONC} → {LAB,RAD}	592	50.70%
{LAB} → {PAS}	1523	43.96%
{LAB,HIS} → {OPT}	599	40.93%
{LAB,HIS} → {PAS}	754	40.87%
{HIS} → {OPT}	725	40.17%
{HIS} → {PAS}	858	39.86%
{HIS} → {LAB,OPT}	615	37.59%
{HIS} → {PAS,LAB}	750	37.53%
{LAB,HIS} → {PAS,OPT}	507	34.34%
{HIS} → {PAS,OPT}	600	33.71%
{LAB} → {OPT}	1142	33.65%
{HIS} → {PAS,LAB,OPT}	509	31.53%
{LAB} → {OPT,HIS}	1029	30.75%
{LAB} → {PAS,HIS}	1040	30.70%
{LAB} → {RAD}	1080	28.54%
{LAB} → {PAS,OPT}	958	28.11%
{LAB,HIS} → {RAD}	784	27.60%
{LAB} → {ONC}	915	27.60%
{HIS} → {RAD}	839	26.31%
{LAB} → {PAS,OPT,HIS}	873	25.79%
{HIS} → {LAB,RAD}	769	25.35%
{LAB,HIS} → {ONC}	545	23.41%
{HIS} → {ONC}	622	22.90%
{HIS} → {LAB,ONC}	549	21.50%
{LAB} → {RAD,HIS}	722	20.74%
{LAB} → {ONC,HIS}	596	17.59%
{LAB} → {RAD,ONC}	568	16.47%
{LAB} → {IMG}	537	14.03%
{LAB} → {PAS,RAD}	450	12.78%

Detailed List of Average Times Between Systems

This table shows (in two parts) the complete list of all possible interactions between two systems and the computed average time between them in days, the standard deviation, support, and 90th percentile.

Sequence	Time (avg.)	St. Dev.	Support	90th %ile
PAS;HIS	1.7	1.9	727	3
ORT;IMG	3	0	1	3
PAS;OPT	5.5	22	520	2.1
OPT;PAS	6	52.6	664	0
OPT;HIS	6.7	62.8	658	3
PAS;IMG	11	18.1	88	22.8
PAS;ORT	12	9	2	19.2
ORT;RAD	20	6	2	24.8
ONC;OPT	20.2	60.4	26	54.5
IMG;IMG	20.4	26.6	35	61
ORT;PAS	22	0	1	22
ONC;HIS	24.2	83.2	104	63
PAS;PAS	26	28.6	501	50
OPT;IMG	29	0	1	29
PAS;RAD	31.4	74.9	181	91
RAD;ONC	31.4	45.5	9	86.6
IMG;PAS	32.9	43.6	63	71.8
ONC;PAS	33.4	46	55	71.6
HIS;IMG	35.5	57.7	207	48
HIS;OPT	35.7	112.5	409	99.2
PAS;LAB	36.9	69.5	1655	91
IMG;ONC	39.1	33	39	71.6
RAD;PAS	41.4	51.8	163	96.6
ONC;IMG	44.8	103.3	33	61.2
IMG;HIS	48.6	34.9	8	88
PAS;ONC	49	154.7	38	68.2
LAB;PAS	51	125.2	2005	130
HIS;ONC	54.6	71.3	438	111
ONC;ONC	56.3	25.1	4	84.6
LAB;ONC	58.2	98.1	553	111
LAB;RAD	62	110.6	820	134
OPT;RAD	63.7	52.5	3	117

Sequence (cont.)	Time (avg.)	St. Dev.	Support	90th %ile
LAB;HIS	64.1	95.5	2134	104
LAB;IMG	64.3	107.5	394	153.4
IMG;OPT	67.9	34.2	33	102.2
IMG;LAB	68	73.1	382	157.6
IMG;RAD	71.6	69.8	113	160
HIS;PAS	71.7	72.3	306	139
ORT;LAB	72.4	76.8	10	180.1
OPT;ONC	77.7	33.2	7	121
ONC;LAB	79.8	182.2	587	179
RAD;LAB	82	95.1	1194	163.7
RAD;ORT	101.7	131.9	3	233.6
ONC;RAD	105.2	199.5	484	160
RAD;RAD	111	201.6	117	228.2
RAD;IMG	112.8	263.1	44	296.9
LAB;OPT	117.5	180.3	571	252
HIS;LAB	119.7	138.8	2136	210
HIS;HIS	154.8	275.5	77	414.2
LAB;LAB	167.3	173.7	24246	295
OPT;LAB	181.1	318.1	346	281
HIS;RAD	209.4	173.1	237	288.6
LAB;ORT	214.3	306.1	19	289.2
IMG;ORT	247.8	280.7	5	555
OPT;OPT	384.1	566.3	20	1331
RAD;OPT	780	577.6	3	1362
RAD;HIS	819.6	971	7	2255
HIS;ORT	-	-	0	-
ORT;HIS	-	-	0	-
ONC;ORT	-	-	0	-
OPT;ORT	-	-	0	-
ORT;ONC	-	-	0	-
ORT;OPT	-	-	0	-
ORT;ORT	-	-	0	-

Process Models built for the System Paths with the HeuristicsMiner Algorithm

Pre-processing

The ProM 5.2 framework was used with the algorithm's default input parameters. The System Paths Dataset was converted from comma separated values format to the required MXML format by ProM 5.2 using the Prom Import Framework software. By default a start and a complete audit trail entries are added but in the case of the System Paths Dataset they would contain the same information. Filters were applied so that only start event types are present. An excerpt of the final MXML format is given below, where process instances refer to a given patient path and audit trail entries encode the footprints (tasks) and respective timestamp. The excerpt shows a full path, ‹HIS, OPT, HIS, RAD›.

```
<Process id="UnifiedSystemPaths" description="Unified single process">
      <ProcessInstance id="123456">
            <AuditTrailEntry>
                  <WorkflowModelElement>HIS</WorkflowModelElement>
                  <EventType>start</EventType>
                  <Timestamp>2001-01-01T09:00:00.000+00:00</Timestamp>
            </AuditTrailEntry>
            <AuditTrailEntry>
                  <WorkflowModelElement>OPT</WorkflowModelElement>
                  <EventType>start</EventType>
                  <Timestamp>2001-04-01T09:00:00.000+00:00</Timestamp>
            </AuditTrailEntry>
            <AuditTrailEntry>
                  <WorkflowModelElement>HIS</WorkflowModelElement>
                  <EventType>start</EventType>
                  <Timestamp>2001-04-02T09:00:00.000+00:00</Timestamp>
            </AuditTrailEntry>
            <AuditTrailEntry>
                  <WorkflowModelElement>RAD</WorkflowModelElement>
                  <EventType>start</EventType>
                  <Timestamp>2001-07-16T09:00:00.000+00:00</Timestamp>
            </AuditTrailEntry>
      </ProcessInstance>
      ...
```

Process Models: Datasets

The System Paths Dataset was split into two datasets. A training set (I) containing 2212 paths and a test set (II) with 2225 paths. The purpose of this is to verify whether the models generated using the training set still hold true with data from the test set, in which case the models can be considered representative of the underlying processes. Furthermore, for each dataset, several experiments were carried out, each resulting in a different process model. The experiments included here are:

A. Default input parameters and data were used.
B. Adjusted input parameters that maximize the model's fitness were used.
C. Default input parameters with artificial start and end tasks (added to each path).
D. Adjusted input parameters (same as B) and artificial start and end tasks.

The HeuristicsMiner algorithm available in ProM 5.2 was used as it is a robust algorithm that copes with noisy data and has been previously applied to healthcare domains, albeit with mixed results. The output of the algorithm is a process model graph.

Process Models: Graph Description

The process models were built based on the system paths, where each task (represented in the graph as a box) is a hospital information system visited by a patient and the arches represent the dependencies between the tasks. The arches' labels indicate the dependency relation value (from 0 to 1, where 1 indicates 100% certainty that the dependency relationship exists between the connected tasks).

Generated Process Models: Summary

	I. Training Set (n= 2212)				II. Test Set (n= 2225)			
Model	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
Inputs								
Positive Observations (N)*	10	300	10	300	10	300	10	300
Results								
Fitness	-.390	.248	.910	.876	-.226	.306	.916	.904
Wrong Observations	1673	1266	10	3	1462	1464	9	9
Connections	22	34	24	24	24	37	21	22
Connections (Artificial Start)			5	5			6	5
Connections (Artificial End)			8	7			8	7

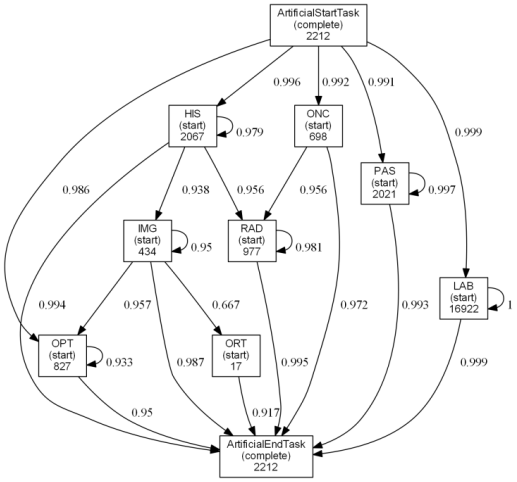
* The default number of positive observations is 10. The number of positive observations that maximised the fitness of models I B and II B was found to be 300. This value was used again in I D and II D.

Interpretation

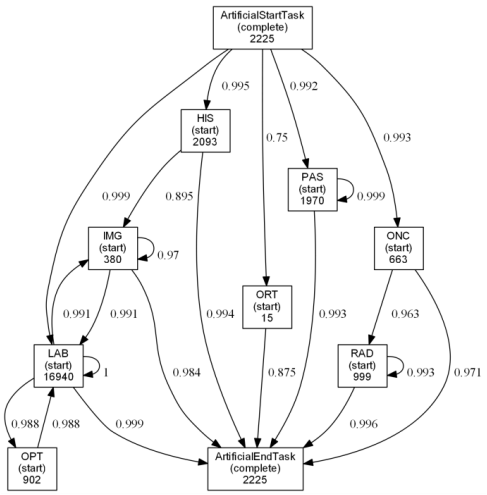
When compared to using the default parameters (A), marginal improvements were observed in models where the parameters of the algorithm were adjusted (in B). However, this was not true for models where artificial start and end tasks were used (C and D). The use of artificial start and end tasks improved the models significantly in respect to fitness (i.e. how the observed process complies with the control flow specified by the process model). Introducing artificial bounds has a normalisation effect, producing models that are easier to read from a single start task.

When comparing the training set to the test set, similar results were observed. In both cases model C outperformed all other models. However, fewer connection were observed in the test set in models C and D. An inspection of the best fitting model, C, revealed further differences between the models produces using the training and test set (model graphs are shown below and in larger formats in the next pages). For example, OPT does not appear after the artificial start in the test set. Instead, this task is exclusively connected to LAB in the test set. Similarly, ORT appears after the start task in the test set but not in the same position in the training set. Instead, ORT appears between IMG and the end task in a full sequence of connections ‹Start, HIS, IMG, ORT, End›. Note that in both cases, the tasks have similar frequencies in the overall training and test sets. The differences can be attributed to the second step of the heuristics miner algorithm, where splits and joins are learnt as well as the third step, where long distance relations are found. Nevertheless, such differences have a significant impact on interpretation of the models. The safer interpretation of the produced models would be to only take those connections that are in agreement between the training and test sets. As such, connection sequences ‹Start, ONC, RAD, End› and ‹Start, HIS, IMG, End› would be the only ones with an overall length larger than 3 (i.e. more than one system visit).

Model I C (Training)



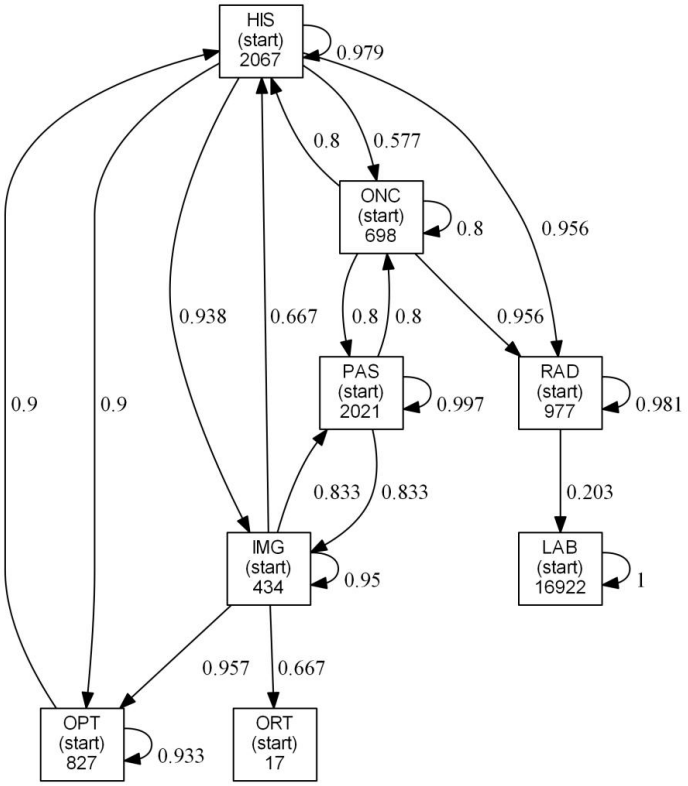
Model II C (Test)



Process Models

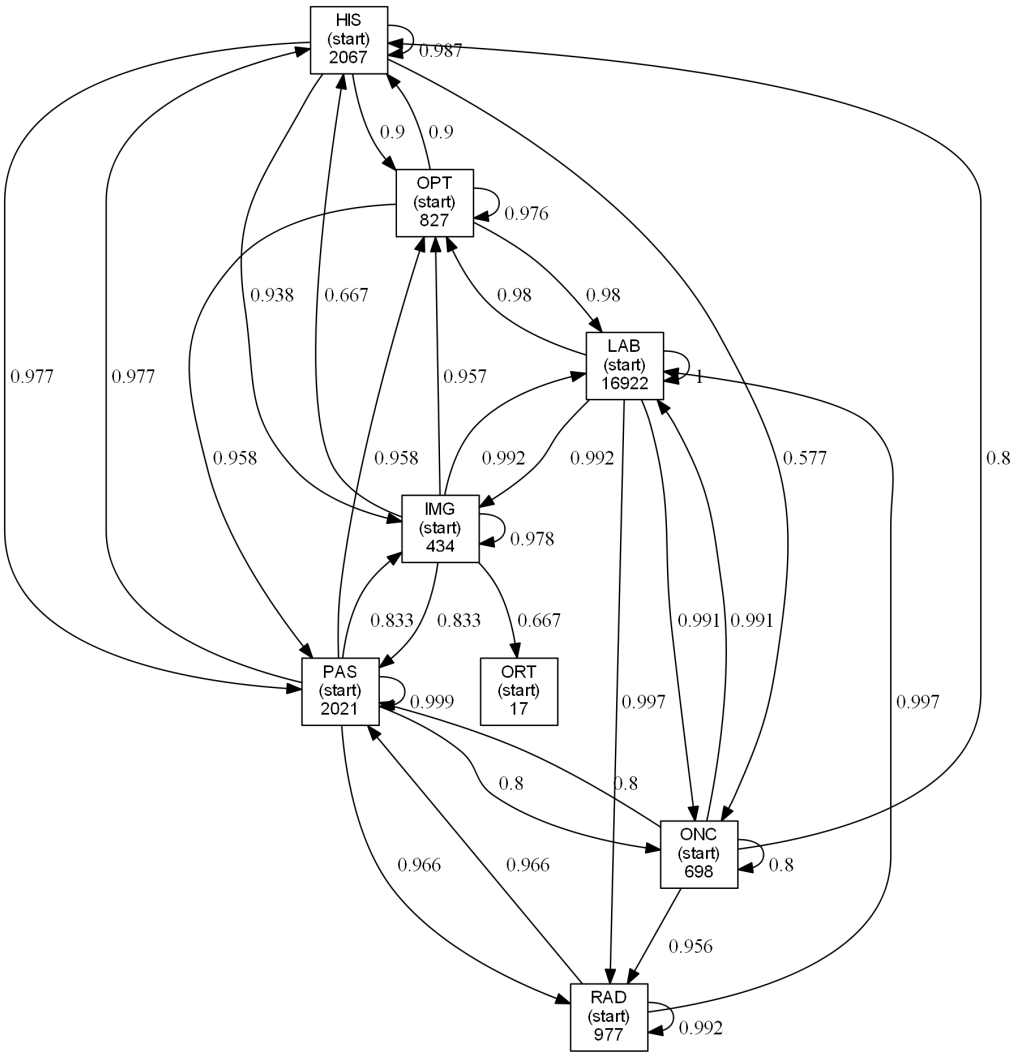
Model I. A - Training Set with Default Parameters

Fitness: -.390 (improved continuous semantics fitness)
Wrong observations: 1673
Connections: 22



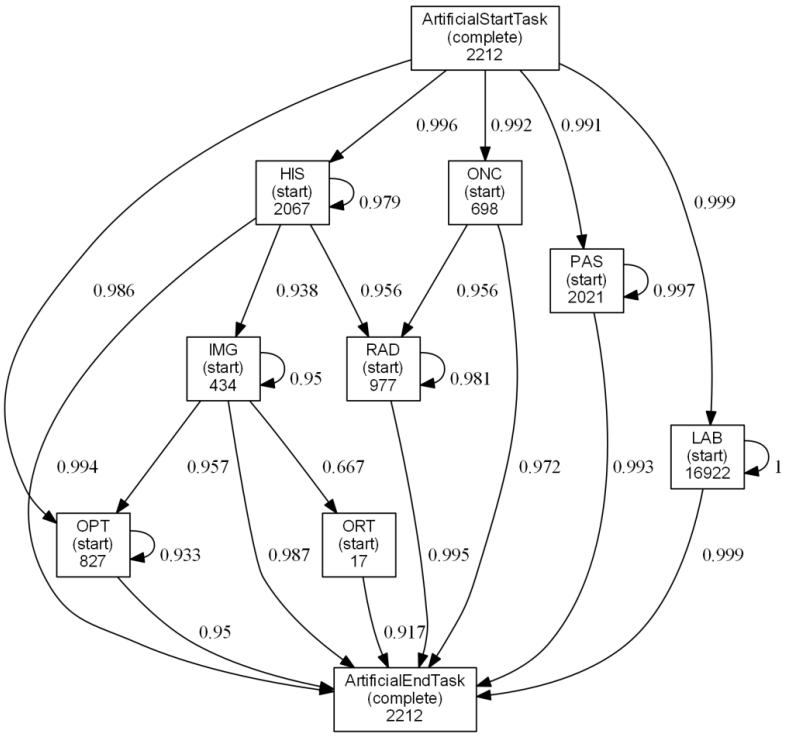
Model I. B - Training Set with Adjusted Parameters

Fitness: .248 (improved continuous semantics fitness)
Wrong observations: 1266
Connections: 34



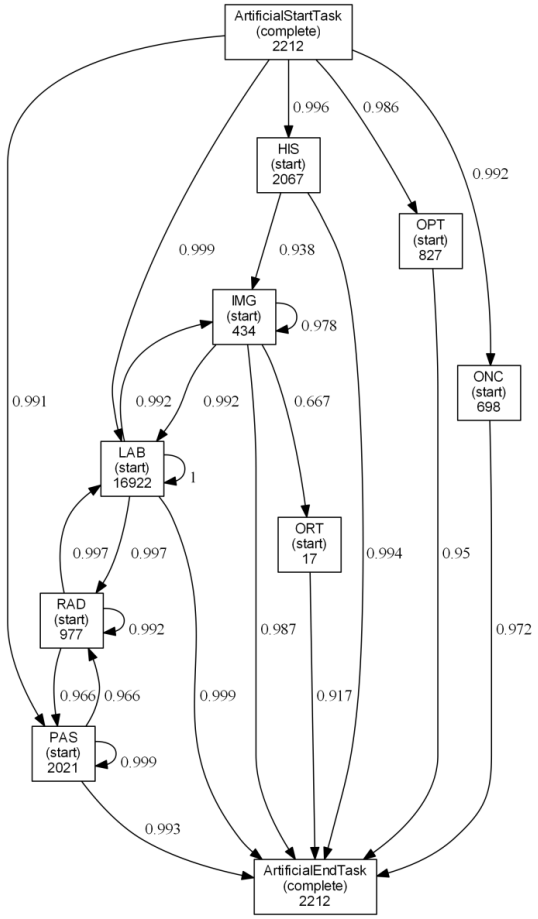
Model I. C - Training Set with Default Parameters and Artificial Start and End Tasks

Fitness: .910 (improved continuous semantics fitness)
Wrong observations: 10
Connections: 24



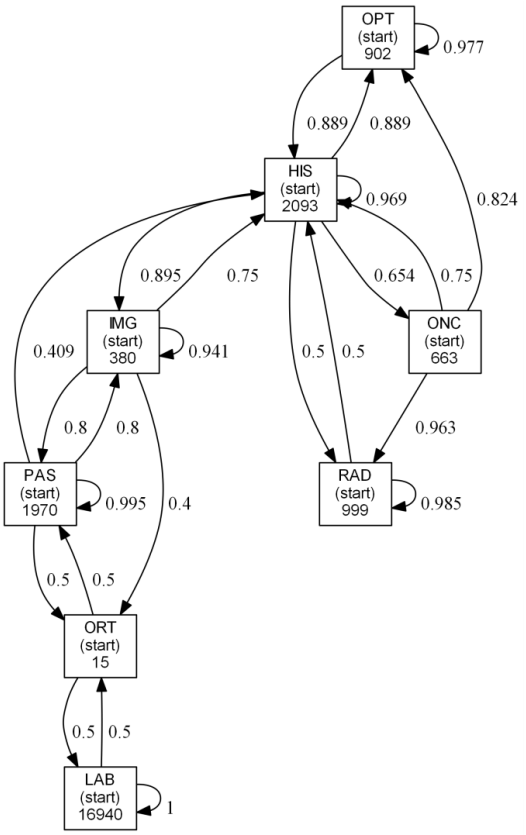
Model I. D - Training Set with Adjusted Parameters and Artificial Start and End Tasks

Fitness: .876 (improved continuous semantics fitness)
Wrong observations: 3
Connections: 24



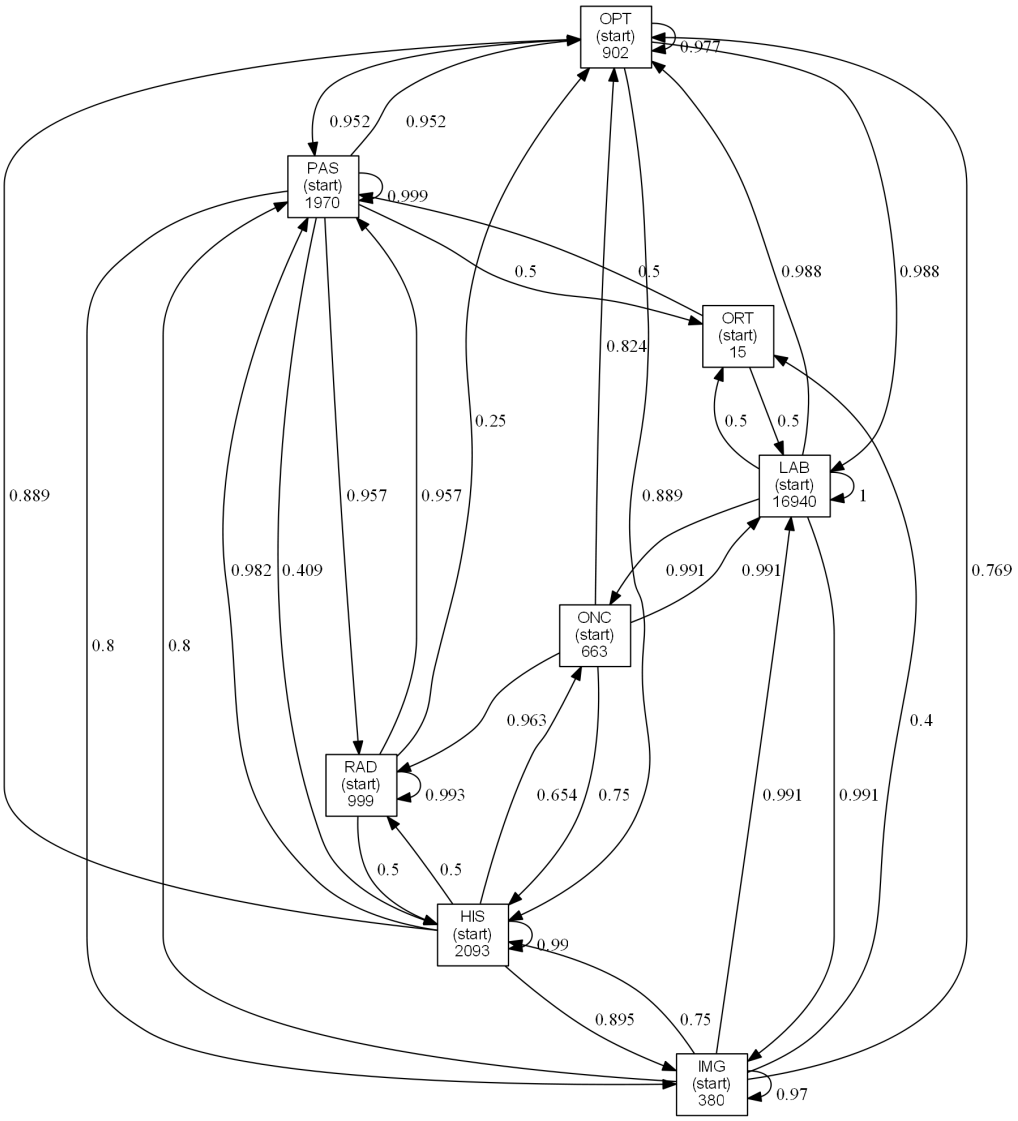
Model II. A - Test Set with Default Parameters

Fitness: -.226 (improved continuous semantics fitness)
Wrong observations: 1462
Connections: 24



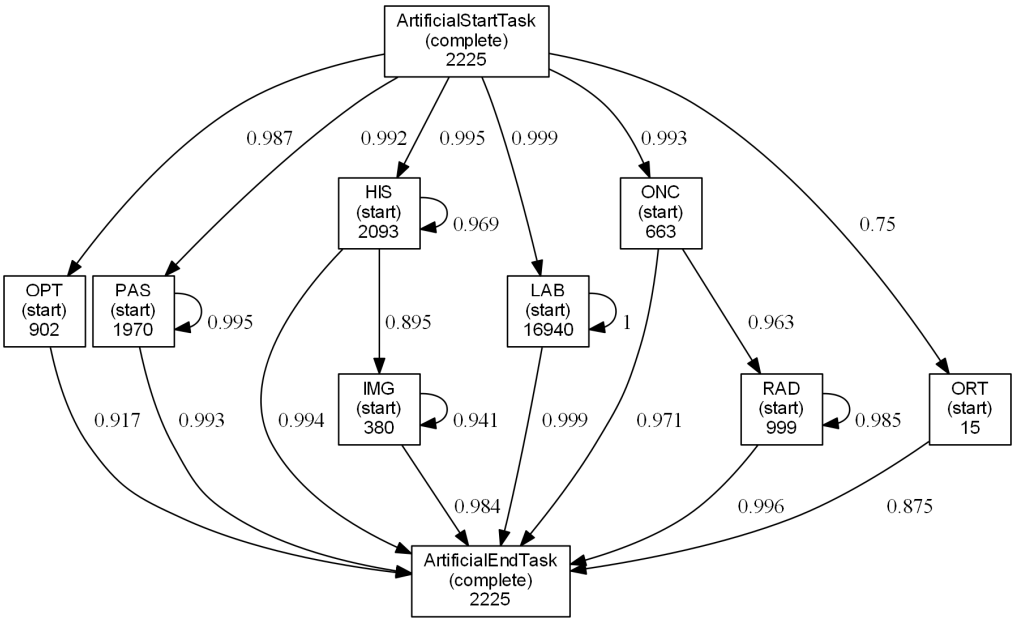
Model II. B - Test Set with Adjusted Parameters

Fitness: .306 (improved continuous semantics fitness)
Wrong observations: 1464
Connections: 37



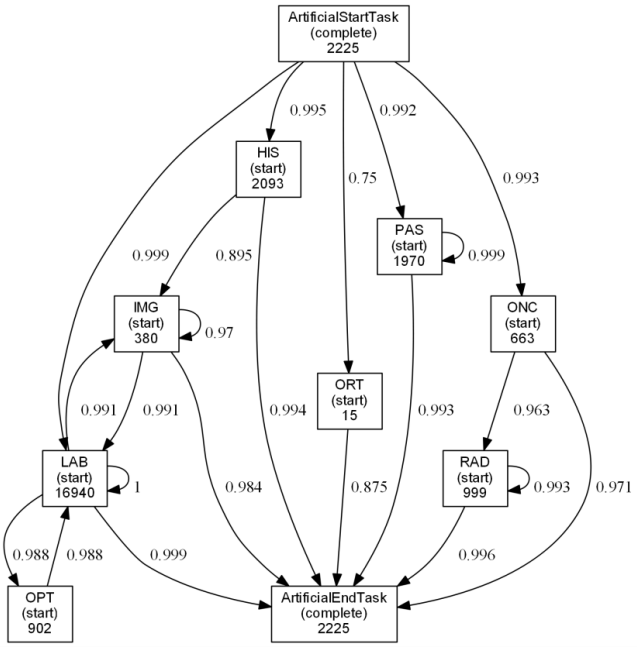
Model II. C - Test Set with Default Parameters and Artificial Start and End Tasks

Fitness: .916 (improved continuous semantics fitness)
Wrong observations: 9
Connections: 21



Model II. D - Test Set with Adjusted Parameters and Artificial Start and End Tasks

Fitness: .904 (improved continuous semantics fitness)
Wrong observations: 9
Connections: 22



Dependency Matrices for System Paths

Similar to the previous exercise, the system paths dataset was split into a training and test set and a dependency matrix was computed for each.

Training Set

	HIS	IMG	LAB	ONC	OPT	ORT	PAS	RAD
HIS	0.99	0.92	0.00	0.62	-0.23	0.00	-0.41	0.94
IMG	-0.92	0.97	-0.02	0.08	0.91	0.57	-0.16	0.44
LAB	0.00	0.02	1.00	-0.03	0.25	0.30	0.10	-0.19
ONC	-0.62	-0.08	0.03	0.80	0.56	0.00	0.18	0.96
OPT	0.23	-0.91	-0.25	-0.56	0.95	0.00	0.12	0.00
ORT	0.00	-0.57	-0.30	0.00	0.00	0.00	-0.25	-0.17
PAS	0.41	0.16	-0.10	-0.18	-0.12	0.25	1.00	0.05
RAD	-0.94	-0.44	0.19	-0.96	0.00	0.17	-0.05	0.99

Test Set

	HIS	IMG	LAB	ONC	OPT	ORT	PAS	RAD
HIS	0.97	0.90	0.01	0.65	-0.25	0.00	-0.41	0.92
IMG	-0.90	0.94	0.00	0.22	0.77	0.40	-0.20	0.46
LAB	-0.01	0.00	1.00	-0.07	0.20	0.13	0.11	-0.17
ONC	-0.65	-0.22	0.07	0.50	0.82	0.00	0.24	0.96
OPT	0.25	-0.77	-0.20	-0.82	0.86	0.00	0.08	-0.25
ORT	0.00	-0.40	-0.13	0.00	0.00	0.00	0.00	0.00
PAS	0.41	0.20	-0.11	-0.24	-0.08	0.00	1.00	0.01
RAD	-0.92	-0.46	0.17	-0.96	0.25	0.00	-0.01	0.99

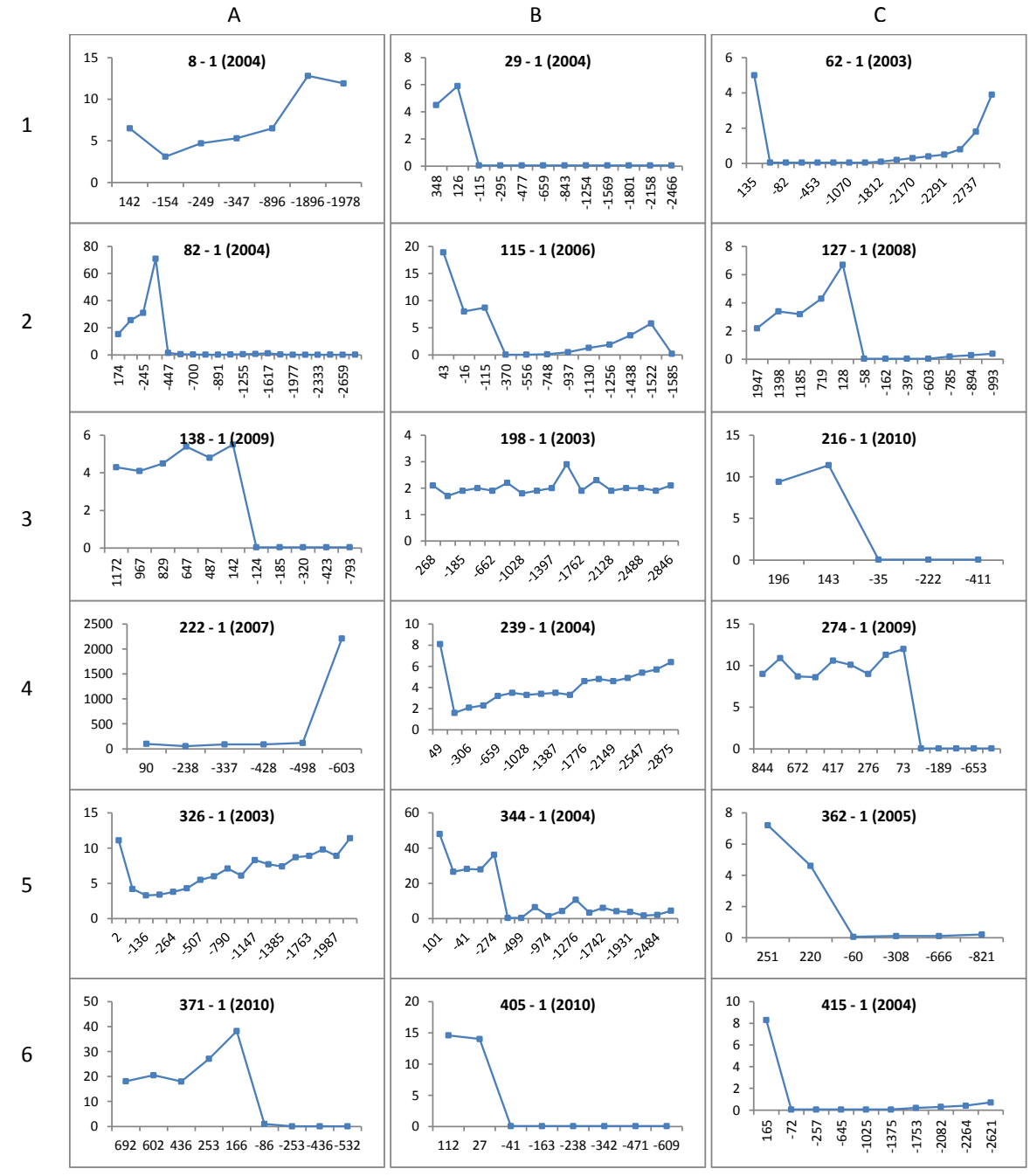
Difference between Training Set and Test Set (only computed for values >.5 in both training and test sets)

	HIS	IMG	LAB	ONC	OPT	ORT	PAS	RAD
HIS	0.02	0.03		-0.04				0.02
IMG		0.03			0.15			
LAB			0.00					
ONC					-0.26			0.00
OPT					0.10			
ORT								
PAS							0.00	
RAD								0.01

1.5 CaP VIS: Plotting PSA Curves

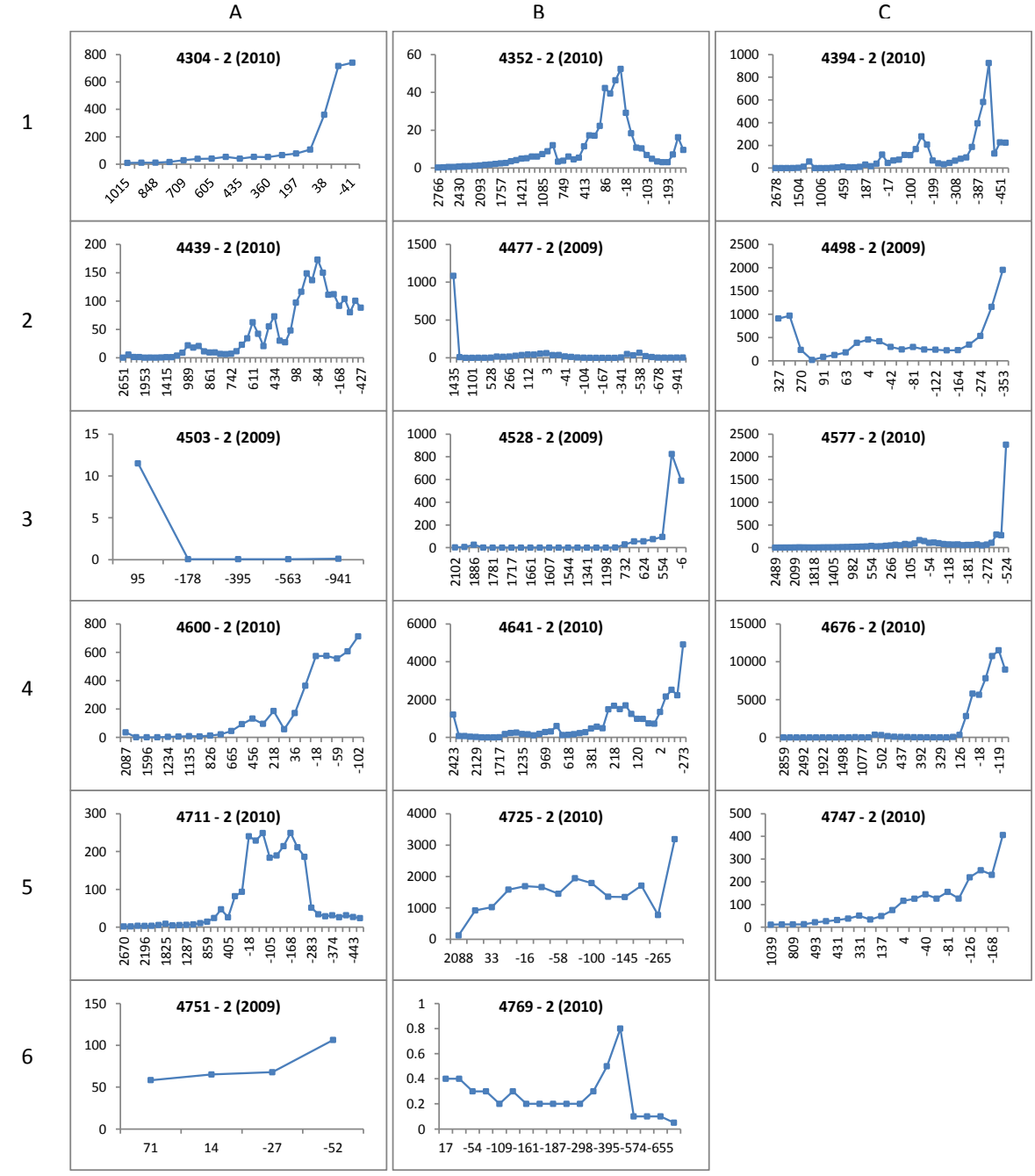
This section shows the first PSA plots created automatically from the pathway data model and the first version of the developed software. The plots are grouped by first treatment type as defined by the National Cancer Waiting Times Dataset.

Surgery



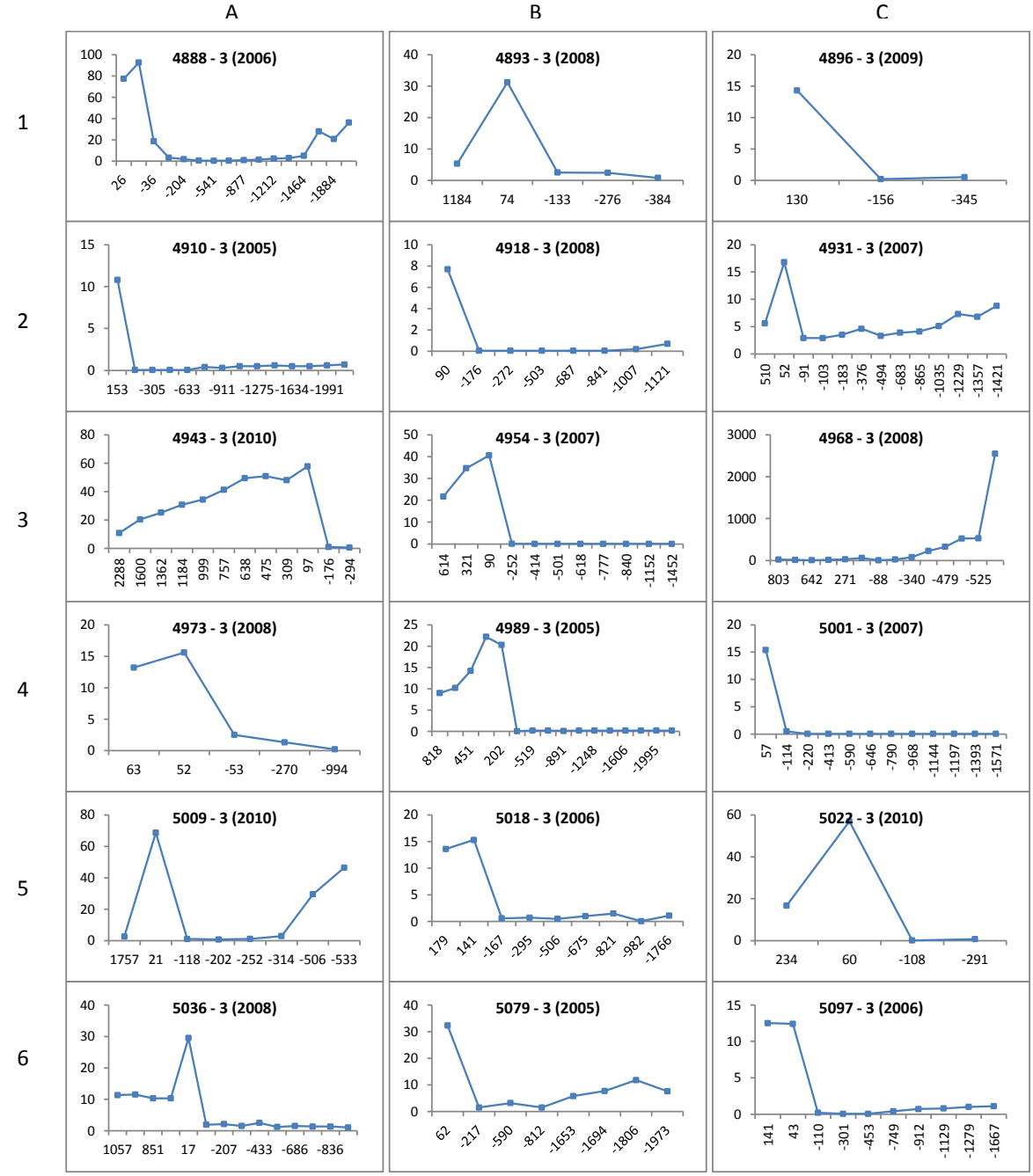
(X-axis: time in days before treatment, Y-axis: PSA value)

Anti-Cancer Drug Regimen – Cytotoxic Chemotherapy



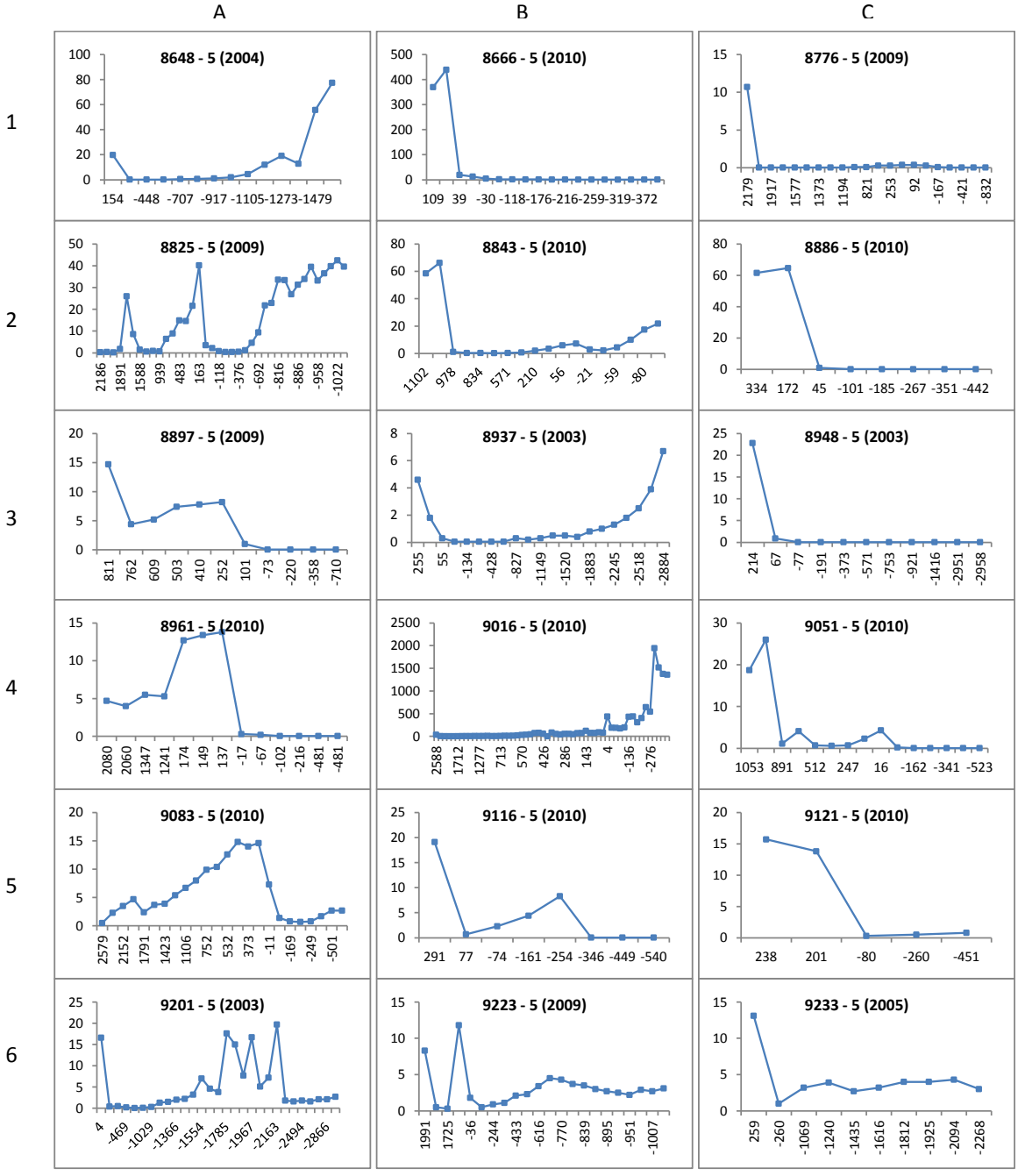
(X-axis: time in days before treatment, Y-axis: PSA value)

Anti-Cancer Drug Regimen – Hormone Therapy



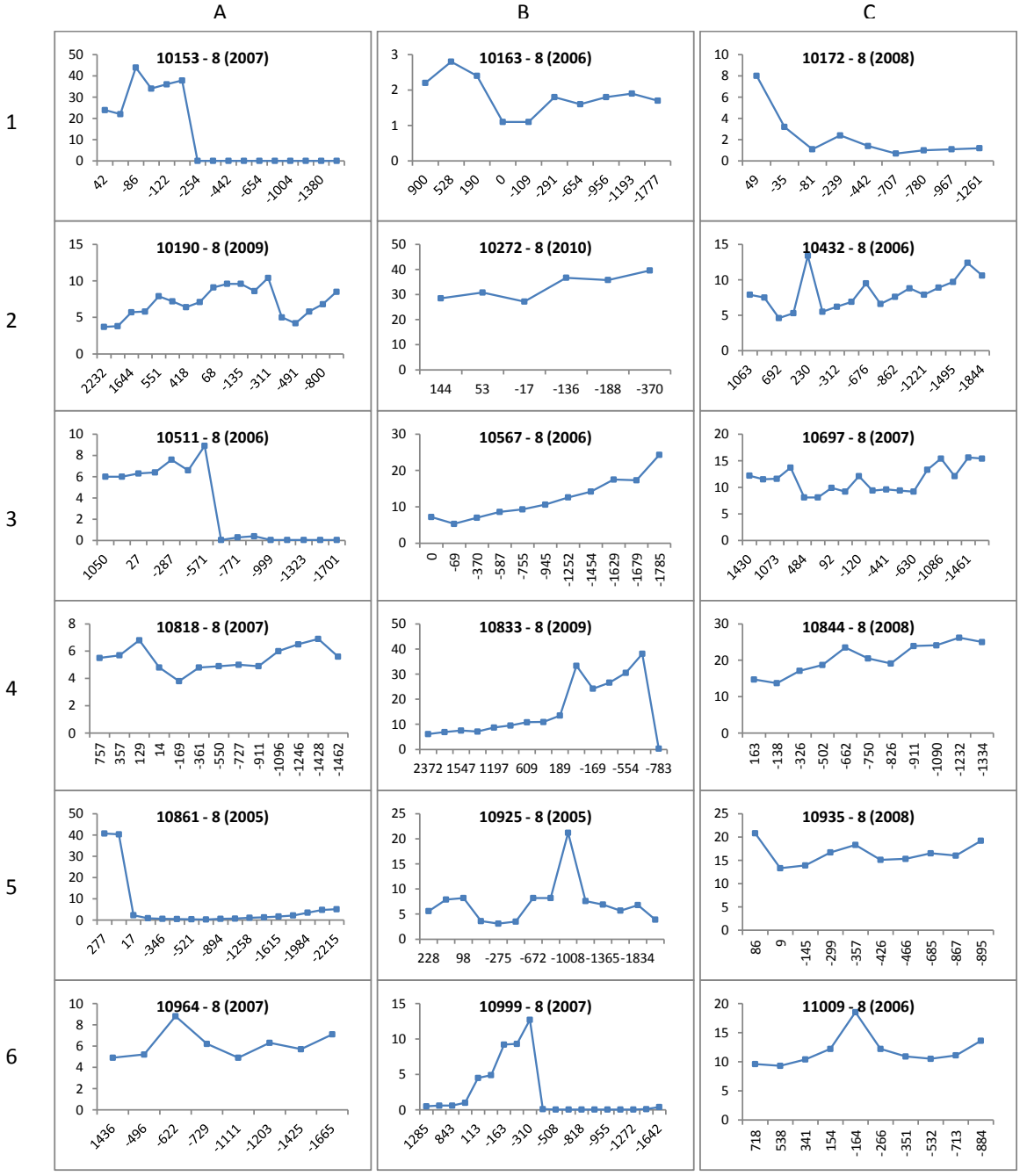
(X-axis: time in days before treatment, Y-axis: PSA value)

Teletherapy (Beam Radiation excluding Proton Therapy)



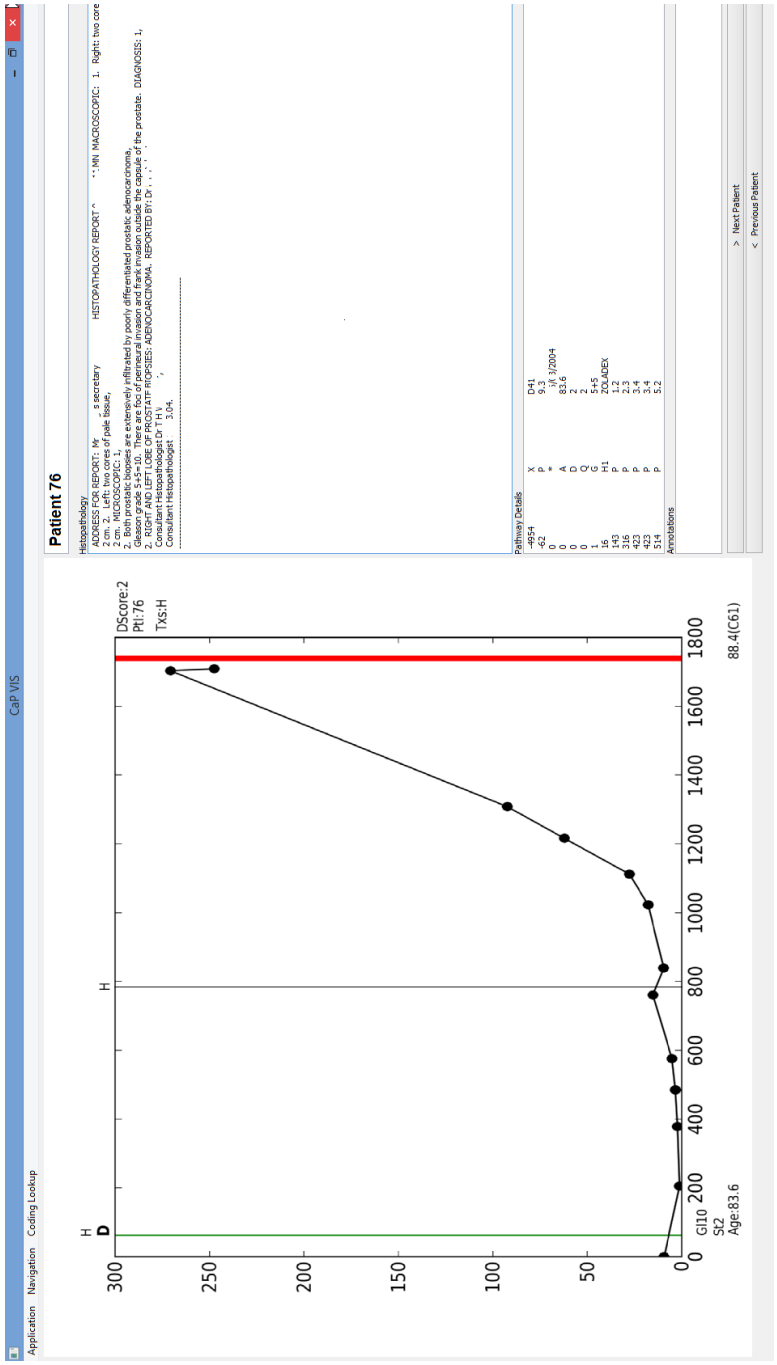
(X-axis: time in days before treatment, Y-axis: PSA value)

Active Monitoring (excluding non-specialist Palliative Care)



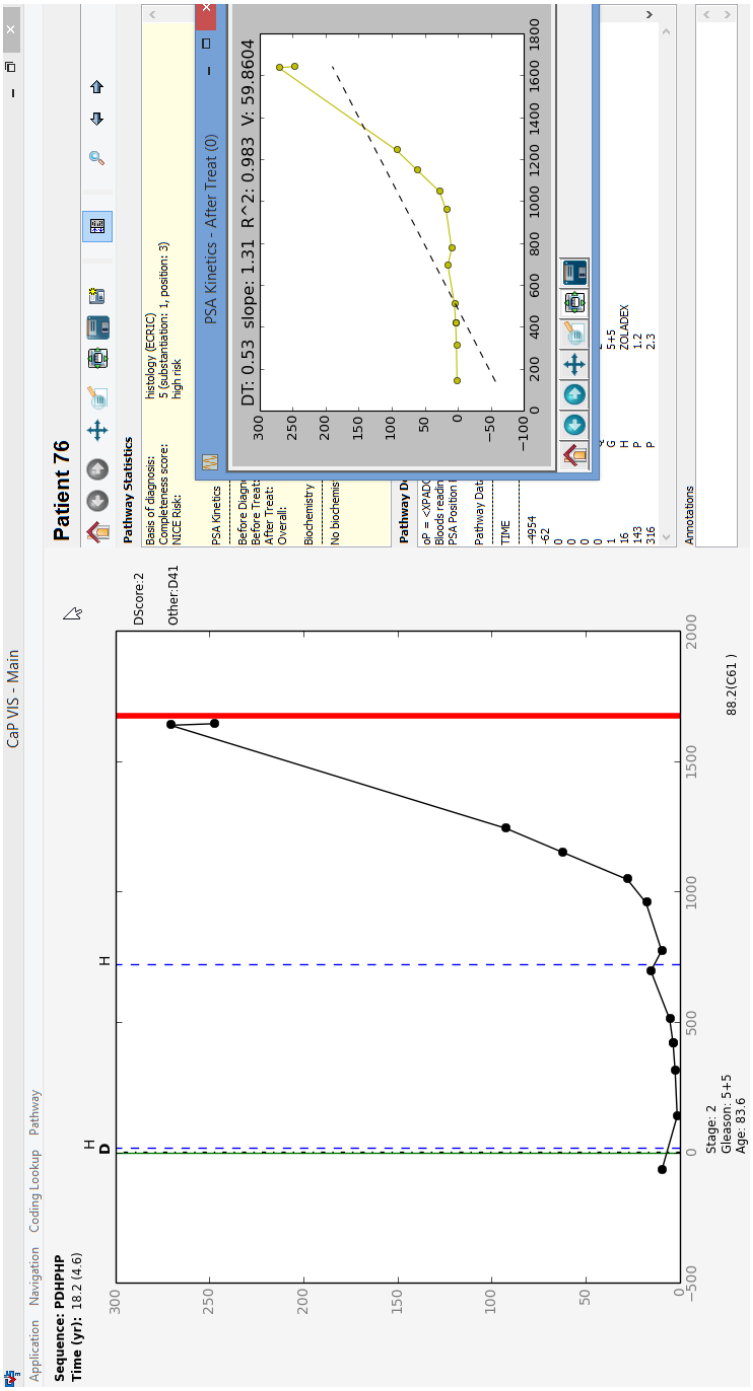
(X-axis: time in days before treatment, Y-axis: PSA value)

1.6 CaP VIS v.2: Main Window Screenshot



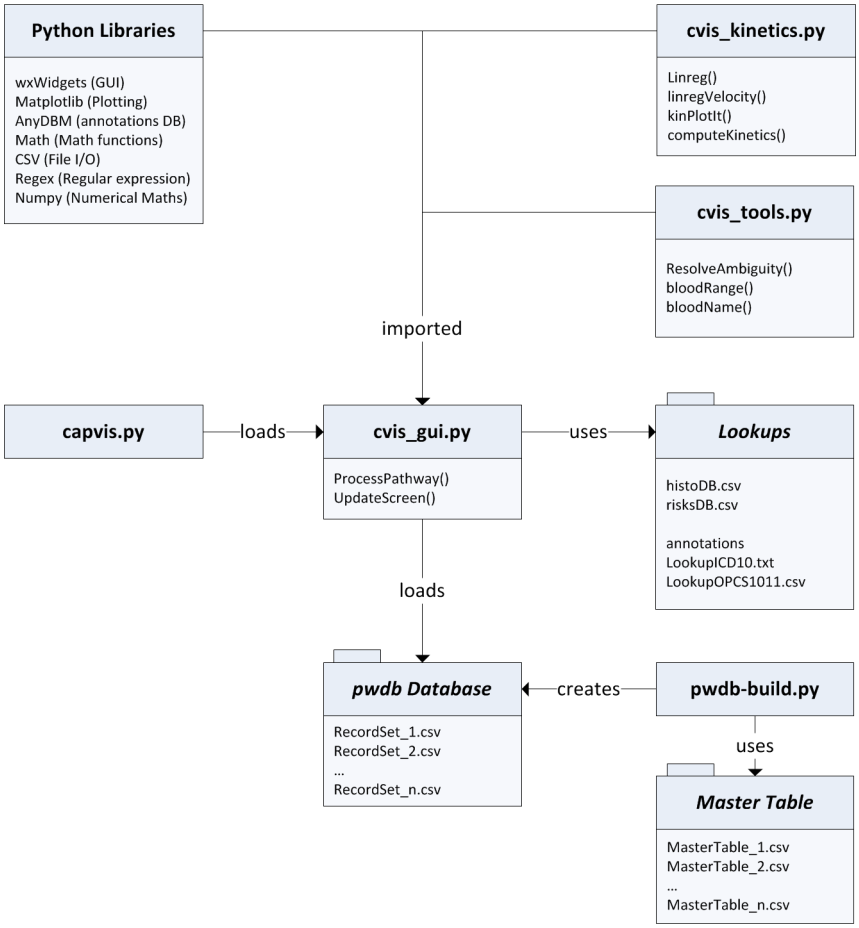
Main Screen for patient (id 76) in CaP VIS version 2.

1.7 CaP VIS v.3: Screenshots and Details



Main Screen for patient (id 76) in CaP VIS version 3.

1.7.1 File System and Interactions



This figure shows the CaP VIS file system and interactions; it shows the main program file (*cvis_gui.py*) and how it interact with Python libraries, other files and the pathway database (*pwdb*).

1.7.2 Console Screenshot

```
C:\Users\J\capvis3>python capvis.py

            --- CaP VIS 3 ---
----------------------------------------
-- Started on Sun Feb 23 11:54:26 2014 --
----------------------------------------

Print detailed Pathways in console? [y/N]

Anchor & Summary pre-load
-------------------------
Do you want the anchor window to load a complete summary? [y/N] y
OK. Reading pathways in pwdb/ and checking annotations + summary, this may take a
while (Ctrl+C aborts)...
Done.

Frame created. Program Ready.

Activity log:

Time            PatientID      Action
11:55:24        10001          > Loading first screen
11:55:24        10001          > Display
11:55:47        20022          > Display
11:55:51         8808          > Display
11:57:46        20022          > Display (search)
11:57:46        20022          > Plot Kinetics Before Treatment
11:58:00        20022          > Display
11:58:09        20022          > Save note
11:58:09         8808          > Display
11:58:12         8808          > Saved plot snapshot as id2-213451.png
11:58:32          122          > Display (anchor window)
11:58:39          122          > Save note

CaP VIS closed on Sun Feb 23 11:59:15 2014

C:\Users\J\capvis3>
```

CaP VIS v3 console output showing user and system actions.

1.7.3 Program Menus and Options

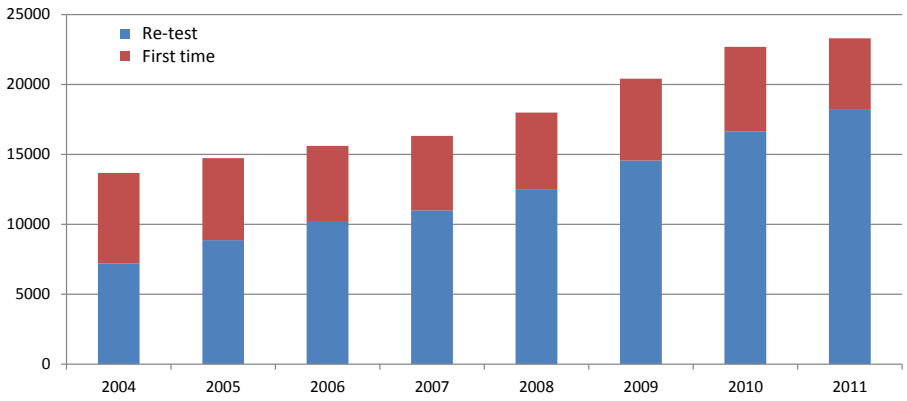
MENU	SUB–MENU / OPTIONS
Application	–> Information
	–> Open ExploraTree
	–> CaP VIS RECON
	–> Delete Annotations DB
	–> Text Size
Navigation	–> Next Patient
	–> Previous Patient
	–> Navigation Anchor
	–> Type in Number
Coding Lookup	–> Diagnoses ICD 10
	–> Procedure OPCS 4.5
Pathway	–> Plot Kinetics
	–> Before Diagnosis
	–> Before Treatment
	–> After Treatment
	–> Overall
	–> Plot additional curve
	–> Navigation Anchor
	–> Options
	–> Show bloods in pathway details
	–> Hide ECRIC treatments
	–> Show Hospital Episodes
	–> Translate coding (slow)

CaP VIS v3 Main Window Menus and Options.

Appendix D - Prostate Cancer Trends

Trends of PSA testing in Norfolk

Overall



- The total number of PSA tests is steadily increasing (8% a year SD 3.6) although the increase in 2011 was the lowest recorded (2.7%).
- Average number of new patients having their first PSA test is constant or on a marginal decline (-3.72% (SD 8.23)).
- Re-testing seems to be driving the overall increase.

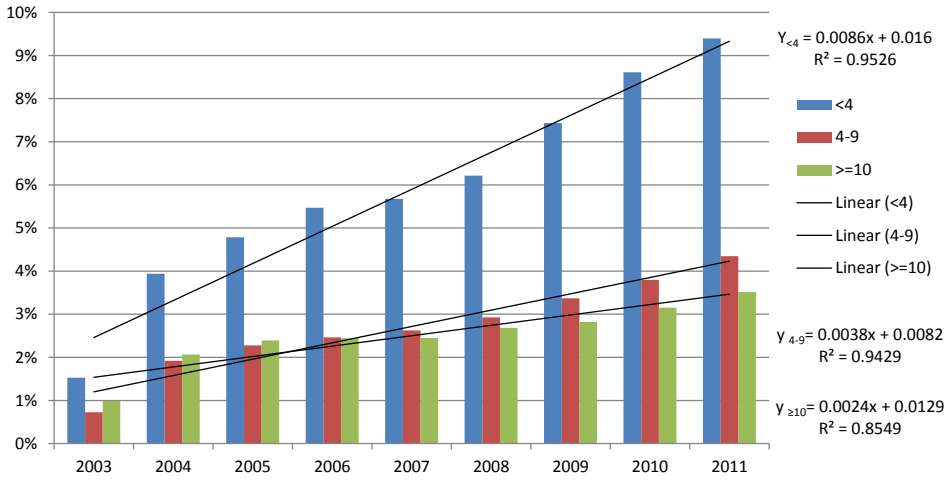
Trends of PSA testing in Norfolk
Overall

- Each patient has, on average, 3.21 (SD .7) tests.
- Every year, 3% (SD .19) of the Norfolk male population (aged 45-84) have a PSA test for the first time.
- And, on average, 6.7% (SD 1.7) have a re-test.
- The average time to re-test is 7 months (SD 3) within two years.

  8.2 months (SD 5) for PSAs <4 ng/ml
  6.5 months (SD 5) for PSA 4-9 ng/ml
  4.7 months (SD 7) for PSA >=10 ng/ml

Trends of PSA testing in Norfolk

PSA Test Results



PSA not age-adjusted.

- The number of normal PSA tests (<4 ng/ml) has increased the most. It is expected to continue to increase by 1% a year, R^2 .95.
- Abnormal tests are expected to grow at 0.4% (4-9 ng/ml, R^2 .94) and 0.2% (>=10 ng/ml, R^2 .85) a year.

Trends of PSA testing in Norfolk

Diagnosed Cancers

When looking at a cohort of 1,460 diagnosed prostate cancers (2005-2009) at the NNUH:

- The number of PSA tests per patient is 5.37 (SD 2.79).
- The rates of testing can be divided as follows:

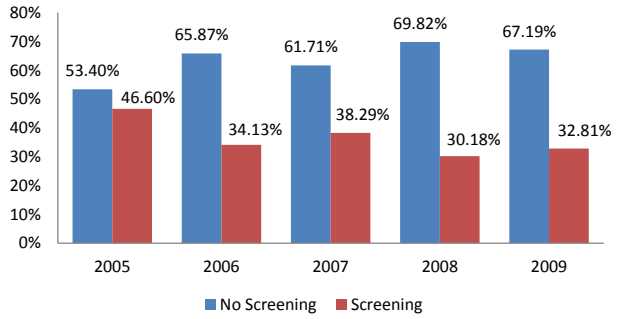
'Screening' (2 year period)	**Before Treatment** (incl. diagnosis)	**After Treatment** (2 year period)	**Total**
0.71 (SD 1.21)	0.88 (SD .73)	3.77 (SD 2.29)	5.37 (SD 2.79)

Screening was defined as the 2-year period up to 3 months before diagnosis.
Follow-up was defined as the 2-year period after treatment.

Trends of PSA testing in Norfolk

Diagnosed Cancers: Screening

- 35% of diagnosed prostate cancers were 'screening' before they were diagnosed (i.e. had 1 or more tests in the screening time window).
- The number of patients on 'screening' declined marginally over the period 2005-2009.



Trends of PSA testing in Norfolk

Diagnosed Cancers: Screening

- Characteristics between 'Screening' and 'No Screening'

No significant difference in age.

Those with a deprivation score of 2 or 3 are 'screening' more actively.

53% of those on screening have a low PSA at diagnosis, compared to 30% no screening.

No screening were more often diagnosed clinically than those that were screening.

		No Screening	Screening
Age Group	<45	0.11%	0.19%
	45-54	2.22%	2.14%
	55-69	38.05%	32.49%
	70-74	19.13%	23.54%
	75-84	31.50%	32.68%
	84+	8.99%	8.95%
Deprivation	1 (Least Deprived)	12.90%	9.34%
	2	22.73%	30.74%
	3	29.70%	34.05%
	4	25.05%	18.87%
	5 (Most Deprived)	9.62%	7.00%
PSA at Diagnosis	N/A	6.34%	0.00%
	High (>10)	41.86%	17.90%
	Med (>4-10)	22.30%	29.38%
	Low (<=4)	29.49%	52.72%
Basis of Diagnosis	clinical	17.34%	3.70%
	histology	82.56%	96.11%
	unknown	0.11%	0.19%

Trends of PSA testing in Norfolk

Diagnosed Cancers: Screening

- Overall 9% of patients in the cohort died of prostate cancer within 3 years.
- Of those Screened, 7% died of prostate cancer and of those not screened, 10% died.

Status at 3 years	Screening	No Screening	Total
Alive	476	853	1,329
Dead	38	93	131
% deaths	7.4%	10%	9%

However, p-value = 0.11 indicates a weak correlation, not statistically significant at .05.

Trends of PSA testing in Norfolk

Diagnosed Cancers: NICE Risk

- Those with a low NICE risk of progression at diagnosis have a marginally higher number of tests in the screening period, 1.13 (SD 1.44).
- Is screening bringing in unnecessary low risk cancers?

		Rates of Testing				
NICE Risk	**N**	**Screening**	**Before Treat**	**After Treat**	**Overall**	**Deaths (3 yrs)**
High	817	0.51 (SD 1.06)	0.92 (SD 0.72)	3.95 (SD 2.54)	5.38 (SD 2.9)	8.5% (n=124)
Intermediate	601	0.95 (SD 1.33)	0.86 (SD 0.72)	3.56 (SD 1.9)	5.37 (SD 2.66)	0.4% (n=6)
Low	42	**1.13** (SD 1.44)	0.58 (SD 0.97)	3.43 (SD 1.94)	5.13 (SD 2.64)	0.1% (n=1)

- A High NICE Risk of progression (PSA+Gleason+TNM) correlates with death within 3 years ($p < .05$, OR 16, CI 7.5-35).

Trends of PSA testing in Norfolk

Diagnosed Cancers: Treatment Types

- No particular treatment type associated with an increased rate of testing.

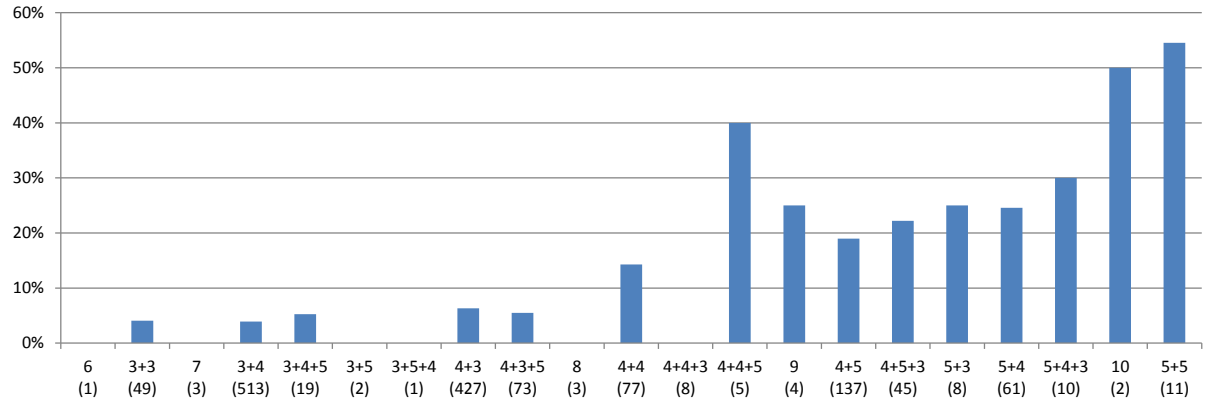
Treatment Package	%	N	Overall Rate of Testing	% Deaths Overall (3yrs)	% Deaths Relative to Tx (3yrs)
Hormone Therapy	52.7%	770	5.6 (SD 2.8)	6.6% (n=97)	12.6%
Hormone + Radio	0.9%	13	5.3 (SD 1.3)	nil	nil
Hormone + Surgery	0.3%	5	4.8 (SD 2.1)	0.1% (n=1)	*20%*
Radiotherapy	0.2%	3	7.7 (SD 1.3)	nil	nil
Surgery	28.6%	417	4.7 (SD 2.8)	1% (n=15)	3.6%
Surgery + Hormone	1.1%	16	5 (SD 1.8)	0.3% (n=4)	*25%*
Watchful Wait	16.2%	236	5.8 (SD 2.7)	1% (n=14)	6%
	1,460		5.37 (SD 2.79)	9% (n=131)	

- Watchful wait showed a similar rate of testing (overall or after decision to treat) to other treatment packages.

Trends of PSA testing in Norfolk

Diagnosed Cancers: Gleason Grades

% Deaths (3 yrs) by Gleason grade



- A high Gleason grade (>=8) shows increased mortality (OR 5, CI 3.4-7.2, p <.05)

Appendix E - Publications

On creating a patient-centric database from multiple Hospital Information Systems

J Bettencourt-Silva, B De La Iglesia, S Donell, V Rayward-Smith

Status: Published in *Methods of Information in Medicine,* 2012; 51(3): 210-20.

Abstract

Background: The information present in Hospital Information Systems (HIS) is heterogeneous and is used primarily by health practitioners to support and improve patient care. Conducting clinical research, data analyses or knowledge discovery projects using electronic patient data in secondary care centres relies on accurate data collection, which is often an ad-hoc process poorly described in the literature.

Objectives: This paper aims at facilitating and expanding on the process of retrieving and collating patient-centric data from multiple HIS for the purpose of creating a research database. The development of a process roadmap for this purpose illustrates and exposes the constraints and drawbacks of undertaking such work in secondary care centres.

Methods: A data collection exercise was carried using a combined approach based on segments of well established data mining and knowledge discovery methodologies, previous work on clinical data integration and local expert consultation. A case study on prostate cancer was carried out at an English regional National Health Service (NHS) hospital.

Results: The process for data retrieval described in this paper allowed patient-centric data, pertaining to the case study on prostate cancer, to be successfully collected from multiple heterogeneous hospital sources, and collated in a format suitable for further clinical research.

Conclusions: The data collection exercise described in this paper exposes the lengthy and difficult journey of retrieving and collating patient-centric, multi-source data from a hospital, which is indeed a non-trivial task, and one which will greatly benefit from further attention from researchers and hospital IT management.

Changes in antiplatelet use prior to incident ischaemic stroke over 7 years in a UK centre and the association with stroke subtype

JR White, JH Bettencourt-Silva, JF Potter, YK Loke, PK Myint

Status: Published in *Age and Ageing,* 2013; 42(5):594-598.

Abstract

Background: guidelines have changed in relation to the indication of antiplatelet therapy for the primary and secondary prevention of stroke. Of interest is how the proportion of patients who had or had not taken antiplatelet agents prior to an incident stroke has changed over time, whether the type of antiplatelet agents used has altered and whether prior antiplatelet use is associated with a particular ischaemic stroke subtype.

Methods: a stroke register was retrospectively examined. All ischaemic stroke patients admitted between January 2004 and March 2011 to a single University Hospital with a catchment population of ~750,000 were included. We excluded those who were on anticoagulants prior to the ischaemic stroke.

Results: a total of 4,307 ischaemic stroke patients [male 47.5%, mean age 77.6 (SD 11.7) years] were included. Of them, 54.7% (SD 2.2%) were not on any antiplatelet therapy prior to their incident stroke. The type and pattern of antiplatelet use prior to stroke did not change significantly during the 7-year study period, and there were no statistically significant differences between different ischaemic stroke subtypes with regards to prior antiplatelet use.

Conclusions: our findings highlight the requirement to improve currently available risk prediction scores as well as the potential clinical impact of antiplatelet resistance within the at risk population who are already on antiplatelets. These findings also indicate that targeting of multiple risk factors may be very important in stroke prevention.

Building data-driven pathways from routinely collected hospital data: a case study on prostate cancer

JH Bettencourt-Silva, J Clark, CS Cooper, R Mills, VJ Rayward-Smith, B De La Iglesia

Status: Submitted.

Abstract

Objectives: Routinely collected data in hospitals is complex, typically heterogeneous and scattered across multiple Hospital Information Systems (HIS). This paper explores how such data can be used to develop pathways and their potential uses in biomedical research.

Methods: We describe a framework for the construction, visualisation and quality assessment of pathways using a case study on prostate cancer. Data pertaining to prostate cancer patients was extracted from a large UK hospital from eight different HIS, validated and complemented with information from the local cancer registry.

Results: Data-driven pathways were built for all patients and an expert knowledge base, containing rules on the prostate cancer biomarker, was used to assess the completeness and utility of the pathways for specific clinical studies.

The proposed framework enables the summarisation, visualisation and querying of complex clinical information as well as the computation of quality indicators. A novel graphical representation of the data-driven pathways allows the synthesis of such information.

Conclusions: This work has enabled further research on prostate cancer and its biomarkers, and on the development and application of methods to mine, compare, analyse and visualise pathways constructed from routinely collected data.

Determinants of in-hospital mortality and length of stay in Total Anterior Circulation Stroke

ND Gollop, JH Bettencourt-Silva, AB Clark, AK Metcalf, KM Bowles, MD Flather, JF Potter, PK Myint

Status: Submitted.

Abstract

Background and Purpose: While the poor prognosis of total anterior circulatory stroke (TACS) is well documented, less is known about the factors associated with in-hospital mortality and prolonged length of hospital stay (LOS) following a TACS.

Methods: We assessed 2977 consecutive cases of TACS admitted to a UK-based teaching hospital between 18/11/1996 to 07/06/2012. Data collected included age, sex, stroke sub-type, pre-stroke functional status (pre-stroke modified Rankin score), pre-stroke residence, significant co-morbidities and outcomes including in-hospital mortality and LOS.

Results: 2445 (82%) of patients had an ischemic stroke. Pre-stroke, 1480 (58%) of individuals were asymptomatic and fully independent. The median age was 81 years with an interquartile range of 74 to 86. Following multivariate analysis, male sex, advanced age, hemorrhagic stroke (without lateralisation), a history of congestive heart failure, and requirement of percutaneous endoscopic gastrostomy (PEG) were significantly associated with in-hospital mortality. Following multivariate analysis the LOS was extended in ischaemic stroke types compared to haemorrhagic, with maximum stay observed in those with the right sided lateralisation (left hemispheric TACS). PEG insertion to support feeding was statistically significantly associated with a shorter stay. A 6-point scoring system was developed, based on the results of multiple logistic regression modelling, to predict inpatient death from risk score and inpatient death rates by year.

Conclusion: We report that male sex, advanced age, haemorrhagic stroke (without lateralisation), congestive heart failure and PEG are all indicators of in-hospital mortality risk and prolonged LOS following a TACS. Furthermore we propose a 6-point prognostic scoring system for prospective validation.

Age but not ABCD^2 score predicts any level of carotid stenosis in either symptomatic or asymptomatic side in Transient Ischaemic Attack

G Mannu, M Kyu, J Bettencourt-Silva, Y Loke, A Clark, A Metcalf, J Potter, P Myint

Status: In Press. International Journal of Clinical Practice

Abstract

Background: The ABCD2 score is routinely used in assessment of transient ischaemic attack (TIA) to assess the risk of developing stroke. There remains uncertainty regarding whether the ABCD2 score could be used to help predict extent of carotid artery stenosis (CAS).

Objectives: We aimed to (i) collate and analyse all available published literature on this topic and (ii) compare the data from our local population to the existing evidence-base.

Materials and Methods:-We conducted a retrospective-observational study over a 6-month period using our East of England hospital-based TIA clinic data with a catchment population of ~750,000. We also searched the literature on studies reporting the association between ABCD2 score and CAS.

Results: We included 341 patients in our observational study. The mean age in our cohort was 72.86 years (SD 10.91) with 52% male participants. ABCD2 score was not significantly associated with CAS ($p = 0.78$). Only age >60 years was significantly associated with - ipsilateral (>50%) and contralateral CAS (>50% and >70%) bilateral CAS ($p<0.01$) after controlling for other confounders. The systematic review identified 4 studies for inclusion and no significant association between ABCD2 score and CAS was reported, confirming our findings.

Conclusion: Our systematic review and observational study confirm that the ABCD2 score does not predict CAS. However, our observational study has examined a larger number of possible predictors and demonstrates that age appears to be the single best predictor of CAS in patients presenting with a TIA. Selection of urgent carotid ultrasound scan thus should be based on individual patient's age and potential benefit of carotid intervention rather than ABCD2 score.

Rates of prostate-specific antigen testing in Norfolk: a retrospective study using secondary care databases

JH Bettencourt-Silva, *et al.*

Status: In preparation.

Abstract

Objective: To assess the overall rate of prostate-specific antigen (PSA) testing for prostate cancer in the Norfolk population and, following a smaller cohort of linked patient data within hospital information systems, to determine the rates of testing at key clinical events for patients diagnosed with prostate cancer.

Methodological steps for the development and maintenance of domain-specific clinical data warehouses from routine hospital data: the Norwich Stroke register experience

JH Bettencourt-Silva, *et al.*

Status: In preparation.

Abstract

Purpose: We propose a simple methodological framework for the design and management of clinical data warehouses using routine data collected from hospital information systems. This method can be used alongside other software development frameworks and we argue that it improves on existing methodologies as it copes with volatile data environments. The methodology was used to develop and maintain the Norfolk & Norwich Research Stroke Register.