Modelling of Cancer Patient Records: A Structured Approach to Data Mining and Visual Analytics

Jing Lu[1], Alan Hales[1,2] and David Rew[2]

[1] University of Winchester, Winchester UK, SO22 5HT
Jing.Lu@winchester.ac.uk
[2] University Hospital Southampton, Southampton UK, SO16 6YD
aahales@btinternet.com, D.Rew@soton.ac.uk

Abstract. This research presents a methodology for health data analytics through a case study for modelling cancer patient records. Timeline-structured clinical data systems represent a new approach to the understanding of the relationship between clinical activity, disease pathologies and health outcomes. The novel Southampton Breast Cancer Data System contains episode and timeline-structured records on >17,000 patients who have been treated in University Hospital Southampton and affiliated hospitals since the late 1970s. The system is under continuous development and validation. Modern data mining software and visual analytics tools permit new insights into temporally-structured clinical data. The challenges and outcomes of the application of such software-based systems to this complex data environment are reported here. The core data was anonymised and put through a series of pre-processing exercises to identify and exclude anomalous and erroneous data, before restructuring within a remote data warehouse. A range of approaches was tested on the resulting dataset including multi-dimensional modelling, sequential patterns mining and classification. Visual analytics software has enabled the comparison of survival times and surgical treatments. The systems tested proved to be powerful in identifying episode sequencing patterns which were consistent with real-world clinical outcomes. It is concluded that, subject to further refinement and selection, modern data mining techniques can be applied to large and heterogeneous clinical datasets to inform decision making.

Keywords: Clinical data environment, electronic patient records, health information systems, data mining, visual analytics, decision support

1 Introduction

The healthcare industry has many established systems being used for electronic patient records, hospital administration, resource management and to circulate clinical results. There is a growing need to be able to share large amounts of health data, perform complex analysis and visualise lifeline tracks of patients. One of the latest approaches is through the implementation of a digital strategy at various levels.

ITBAM 2017 Submission
© Springer-Verlag Berlin Heidelberg 2017

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

1.1 Background

After years of digitising patient records, the UK National Health Service (NHS) is acquiring a considerable repository of clinical information – hundreds of millions of test results and documents for tens of millions of patients. NHS England describes the role of health informatics as being fundamental to the transformational changes needed. As a result the NHS is investing in a number of initiatives to integrate data and provide insight. However, despite this, the NHS "has only just begun to exploit the potential of using data and technology at a national or local level" [17].

While the NHS collects huge amounts of data on patient care, it is nevertheless very difficult to establish ground truths in the relationships between multidisciplinary clinical inputs and therapeutic outcomes in a wide variety of chronic diseases of childhood and adulthood. This is not only because of the sheer complexity of human lives and populations, but also because of the number of variables which affect real-world health events. All human lives and health events play out over time, from conception to death, such that the passage of time is a central element in each and every disease process. Despite this truism, it is not easy to find commercial health informatics systems in which temporally-structured data presentation and analysis is a central element.

The electronic patient record (EPR) system is a digital compilation of every patient's healthcare information, unique identifiers and demographic data, and contains a range of documents from notes to test results [22]. One benefit to having EPRs is that paper records no longer have to be maintained, which supports the government's 2018 vision of a paperless NHS.

Emerging big data technologies mean it is now possible to share data from systems that previously could not communicate; this could potentially allow different parts of the health service to work together [18]. All of this stored data can be used for aiding decisions or to learn something new [15]. One approach which could provide doctor's with the ability to derive useful information from massive medical datasets is known as big data analytics.

1.2 Case Study – University Hospital Southampton

In 1996, Professor Ben Shneiderman and colleagues at the Human Computer Interaction Laboratory in Bethesda at the University of Maryland developed a conceptual structure for the visualisation of clinical records in which a complex and heterogeneous series of data could be displayed and easily understood on parallel timelines within a simple two-dimensional graphic. This nevertheless had powerful features and it formed the basis for a universal descriptor for human lives and medical histories as a tool for the overview of a complex dataset, into which individual documents and reports could be readily zoomed, and from which extraneous or unwanted detail could be temporarily filtered out [14].

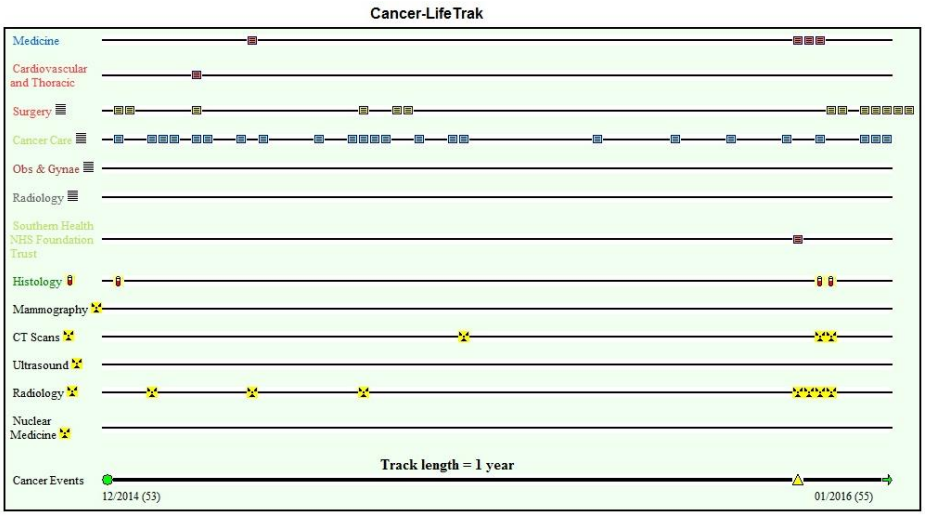
The Lifelines model has been adopted as a framework for a practical, dynamic and interactive EPR which now sits within the University Hospital Southampton (UHS) clinical data environment (CDE). It provides much faster access times and overviews

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

to the more than one million patient records compared with other available software. UHS Lifelines has proved to be a valuable testbed for timeline-structured EPRs and it remains in continuous, agile and iterative development.

Figure 1 is a screenshot of the UHS Lifelines timeline-structured EPR graphical interface, taken from a patient record within the Southampton Breast Cancer Data System. This figure illustrates a number of unique features of the system including separate clinical and reporting timelines, on which are displayed icons at the time of their generation in an intuitive manner. Clicking on each icon displays the underlying document or report. The lowermost timeline, labelled "Cancer Events", is the master timeline or "UHS Lifetrak". In this case, a patient developed a right-sided cancer in December 2014 (green circle) and overt metastases in January 2016 (yellow triangle).

Figure 1 Screenshot of the UHS Lifelines timeline-structured EPR graphical interface



The clinical document outputs and diagnostic test result events shown are those which are most frequently relevant and informative when considering a breast cancer case. The entire diagnostic test result history for a patient is frequently so large however that it would create visual noise and severely limit the effectiveness of the lifeline graphic.

The data visualisation and access tools within UHS Lifelines provide a methodology which can be adapted to research into the natural history and progression of a wide range of human diseases, and the opportunity for new approaches to temporally-orientated data mining and data analysis [22]. This paper presents a framework in this context and describes the application of several data mining tools to develop new insights into the data and the structure of the data system.

Since 2012 UHS have designed, built and populated the Southampton Breast Cancer Data System (SBCDS) to help understand the clinical inputs and outcomes for a substantial cohort of breast cancer patients who have been treated in one of the

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

largest specialist centres in the UK since the 1970s. A variety of data sources has been used including a card index of more than 12,000 patients, which has been maintained continuously since 1978, and a range of legacy concurrent datasets with information on breast cancer patients.

It soon became apparent that, by building links to elements of the UHS Lifelines data system into the individual cancer patient records, there would be a direct evidence base for each patient's clinical progression within the EPR system. This evidence base would accrue continuously to the patient record (when alive) at every contact with the hospital, in whatever clinical discipline. By integrating access to the electronic documents into the data system, the cost and time penalty of populating the individual datasets would be reduced significantly when compared with calling forward paper records.

The result has been that from 2012 onwards SBCDS has accrued more than 17,000 unique timeline-structured patient records, which continue to accrue at the rate of around 10 per week. These are undergoing continuous updating and validation, the results of which will be published separately in due course.

It then became apparent that the clinical progression of the disease could be described for each and every patient along a master timeline, which has been called the UHS Cancer-LifeTrak – because it resembles the single track railway line on a UK Ordnance Survey Land Ranger map. Each station on the line would represent a point of transition in the patient's care pathway – from the time of Primary Diagnosis through to Loco-Regional Recurrence to the appearance of overt distant Metastases to final outcome. In practice, the progression of disease is very variable and complex, and its representation is more challenging because patients may present with left, right or bilateral tumours in series or parallel. As Southampton is a maritime city, the right-sided tumours are represented by green icons (starboard buoys) and left-sided tumours by red icons (port buoys).

The fact of these episode markers now allows the duration of episodes to be measured between the various phase transitions of the disease. These time intervals can be related back to the original pathological descriptors of the tumours and to the multidisciplinary treatment inputs, which variously include surgery, systemic chemotherapy, radiotherapy and anti-oestrogenic therapy. The time points are measured in months and years (mm/yyyy), as measuring by days would afford spurious accuracy to the data when diagnoses and treatments occur over days, weeks and months.

The temporal structure to the data system presented an opportunity to explore a range of different tools for data mining and analytics. This motivated a collaborative research project between UHS and Southampton Solent University, which began in 2014 with the following objectives: enhancement of the SBCDS user interface; expansion of its data mining capability; and exploitation of large-scale patient databases. Anonymised breast cancer data from UHS was pre-processed and several data mining techniques were implemented to discover frequent patterns from disease event sequence profiles [12].

The remainder of this paper proceeds as follows: a process-driven framework for health data analytics is proposed in section 2 which comprises a data layer, functional

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

layer and user layer. The data warehousing, data mining and visualisation components of the methodology are discussed further. In section 3, following a description of the data sources, the emphasis is on the pre-processing for data mining and multi-dimensional modelling. A series of results is given in section 4 which covers the visual analytics and data mining techniques, highlighting sequential patterns graph and decision tree classification. The discussion then proceeds to include an evaluation by domain experts before the concluding remarks.

2 Methodology

Informed by the collaborative work using a complex anonymised real-life dataset, a process-driven framework for health data analytics is described below.

2.1 Process-Driven Framework

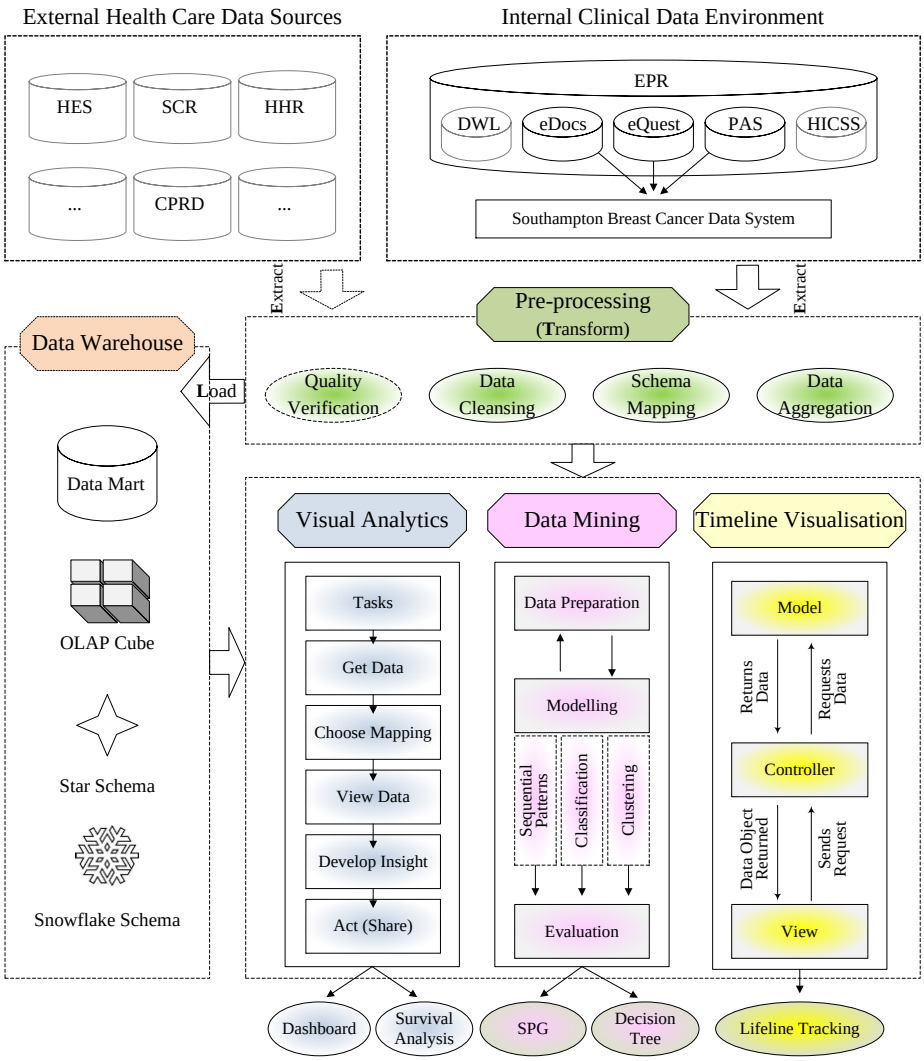
Healthcare generates a vast amount of complex data with which to support decision making through information processing and knowledge extraction. The growing amount of data challenges traditional methods of data analysis and this has led to the increasing use of emerging technologies – a conceptual architecture is shown (see Figure 2) which integrates the pre-processing, data warehousing, data mining and visualisation aspects. In the healthcare context, the data may include clinical records, medical and health research records, administrative and financial data – for example Hospital Episode Statistics, Summary Care Records, the Doctor's Work List and the Patient Administration System.

The process-driven framework comprises a data layer, a functional layer and a user layer. Pre-processing is used as part of the data layer in order to clean raw data and prepare the final dataset for use in later stages. Data cleansing and preparation stages include basic operations such as removing or reducing noise, handling outliers or missing values, and collecting the necessary information to model. Extraction, Transformation and Loading (ETL) is well known throughout the database industry: extracting data from various sources then transforming it through certain integration processes before finally loading the integrated data into a data warehouse. The data from the warehouse is held in a structured form and available for data mining, visualisation or analytics.

The functional layer includes (1) Data Warehousing – integrating data from multiple healthcare systems to provide population-based views of health information; (2) Visual Analytics – applying data visualisation techniques to healthcare data, transforming clinical information into insight through interactive visual interfaces; (3) Data Mining – bringing a set of tools and techniques that can be applied to large-scale patient data to discover underlying patterns, providing healthcare professionals an additional source of knowledge for making decisions; and (4) Timeline Visualisation – comprising a patient lifeline system with application to chronic diseases, enabling tracking of clinical/patient events over time.

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

Figure 2 Process-driven framework for health data analytics



Finally the user layer shows the possible results which can be derived, including for example graphical charts for survival analysis, representing output from visual analytics; as well as sequential patterns graphs and decision trees from data mining. These will be illustrated in the Results section of this paper.

2.2 Data Warehousing

Healthcare datasets come from various sources while health information systems are generally optimised for high speed continuous updating of individual patient data and

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

patient queries in small transactions. Using data warehousing can integrate data from multiple operational systems to provide population-based views of health information.

A data warehouse can be defined as "a copy of transaction data specifically structured for query and analysis" [9]. In order to facilitate decision-makers, complex information systems are assigned with the task of integrating heterogeneous data deriving from operational activities. Case studies include one from Stolba and Tjoa [23], who used a clinical evidence-based process model for the generation of treatment rules. Another example, the Data Warehouse for Translational Research (DW4TR), has been developed to support breast cancer translational research in the USA and this has been extended to support a gynaecological disease programme [7].

A data warehouse is often a collection of data marts: a sub-set of a data warehouse containing data from just one subject area. There are several ways a data warehouse or data mart can be structured, for example multi-dimensional, star or snowflake. However the underlying concept used by all the models is that of a *dimension*, representing the different ways information can be summarised such as by geography, time intervals, age groups and patients. Common to the star and snowflake models is the *fact* table, which contains data (factual history) such as cost or quantity.

On-Line Analytical Processing (OLAP) is an approach to answering multi-dimensional analytical queries. An OLAP cube is a term that typically refers to multi-dimensional arrays of data. OLAP tools enable users to analyse data interactively from multiple perspectives and consist of analytical operations such as roll-up, drill-down, and slicing and dicing.

2.3 Data Mining and Modelling

Data mining is the essential part of *knowledge discovery in databases* – the overall process of converting raw data into useful information and derived knowledge – one definition being "the science of extracting useful information from large datasets or databases" [5]. Data mining techniques could be particularly useful in healthcare and personalised medicine through the following areas of activity: drug development and research, forecasting treatment costs and demand of resources, anticipating patients' future behaviour given their history and the usage of data mining for diagnosis [6].

While data preparation will be discussed further in section 3.2, three data mining methods have been considered: Sequential Patterns Mining aims to find sub-sequences that appear *frequently* in a sequence database; Classification maps each data element to one of a set of pre-determined classes based on the *differences* among data elements; Clustering divides data elements into different groups based on the *similarity* between elements within a single group. Once a model is built from a data analysis perspective, it is important to evaluate the results and review the steps executed to construct the model.

Regarding classification applications in breast cancer studies, Jerez-Aragones et al [8] presented a decision support tool for the prognosis of breast cancer relapse. It combined an algorithm for selecting the most relevant factors for prognosis of breast cancer with a system composed of different neural network topologies. The identification of breast cancer patients for whom chemotherapy could prolong

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

survival has been treated as a data mining problem as well [10]. Martin et al [16] examined factors related to the type of surgical treatment for breast cancer using a classification approach and Razavi et al [19] discuss a decision tree model to predict recurrence of breast cancer.

Clustering techniques have also been applied in breast cancer diagnosis with either benign or malignant tumours [2]. The comparison from their results showed that the k-means algorithm gave an optimum outcome due to better accuracy. In addition, sequential patterns mining has been explored to show the applicability of an alternative data mining technique, e.g. its application to a General Practice database to find rules involving patients' age, gender and medical history [21].

2.4 Visualisation

Visual Analytics

Visual analytics is an integrated approach that combines data analysis with data visualisation and human interaction. There are four separate stages in the process – data, visualisation, knowledge and models. Data mining methods are often used to generate models of the original data. Visual analytics normally commences with a pre-determined task – then goes through an iterative process to get the required data, choose appropriate visual structure (e.g. chart/graph), view the data, formulate insight and then act. This process involves users moving around between different steps as new data insights (and new questions) are revealed.

Tableau Software (https://www.tableau.com) supports this iterative process and provides a collection of interactive data visualisation products focused on business intelligence. After connecting to the data warehouse and adding the tables needed for the analysis, Tableau identifies the fact and dimension tables then sets up every dimension for immediate analysis. Sample results will be demonstrated in section 4.1.

Timeline Visualisation

It has been recognised that the timeline-based data visualisation model can be used as a generic tool with application in the study of all chronic diseases of childhood and adulthood, and as a template for other forms of health informatics research. The concept of the Lifelines EPR can thus been extended to the development of an integrated data system within the UHS-CDE using breast cancer as an exemplar [22].

The model–view–controller (MVC) architectural pattern and a timeline visualisation library can be applied to implement user interfaces. As shown in Figure 2, within the Timeline Visualisation box, MVC divides a software application into three interconnected parts and each of them are built to handle specific development aspects [20]. Model represents the data structure and typically contains functions to retrieve, insert and update the information in the database. View is the information that is being presented to a user and it is normally via a web page. Controller serves as an intermediary between model and view – it handles user requests and retrieves data.

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

MVC is one of the most frequently used industry-standard web development frameworks to create scalable and extensible projects. Considering two of the models in this case study – the patient model and the events model – based on patient ID, the controller will open up the patient model and subsequently retrieve all the relevant events. The events model will prepare the data for the template and return it back to the patient model, which will then return that prepared data back to the controller. The controller will then pass that data to the view to visualise the timeline.

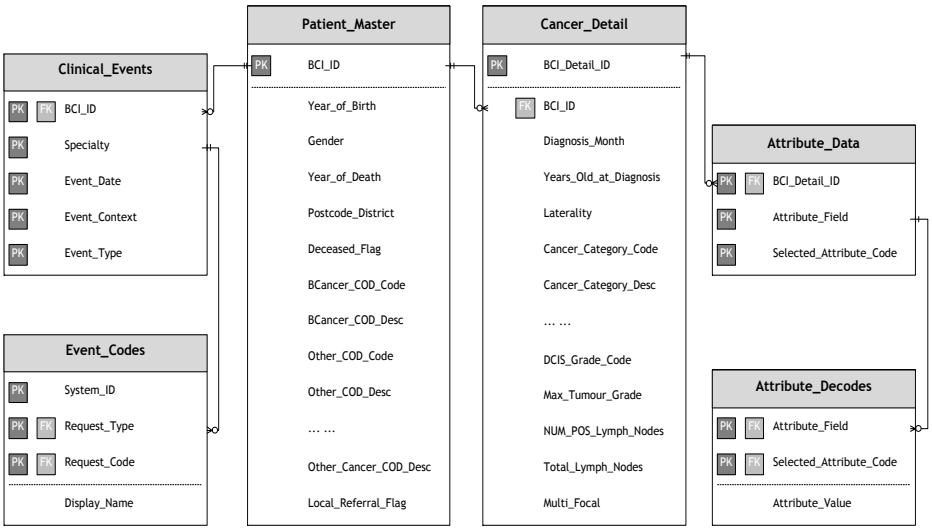
3 Clinical Data

The data sources from the Southampton case study are described in this section along with the pre-processing and multi-dimensional modelling techniques.

3.1 Data Sources and Understanding

Extracting data from the clinical environment requires knowledge of the database model and data dictionary as well as domain expertise. In this study, UHS data has been extracted in November 2015 from SBCDS which contains information for 17,000+ breast cancer patients, with a total of 23,200+ records (instances) showing their cancer details. The principle to extract data for this research project is to strictly avoid providing sufficient information that an individual might reasonably be identified by a correlation of seemingly anonymous individual items. Based on this principle, four tables have been exported that reflect SBCDS structure and were loaded into an Oracle database: Patient_Master, Cancel_Detail, Attribute_Data and Attribute_Decodes.

Figure 3 Entity Relationship Diagram (ERD) for Tables extracted from SBCDS



The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

In addition, two more tables have been added to help demonstrate the event clustering challenge through overloading of the graphical interface: Clinical_Events and Event_Codes. Figure 3 shows the ERD for the six tables extracted from SBCDS – note that Request_Type and Request_Code in the Event_Codes table correspond to Specialty and Event_Context in Clinical_Events. This model does not represent the original system which also includes data entry, system management and data analysis.

It can be helpful to show the other types of clinical data that breast cancer patients often have, and how that data relates temporally to the cancer data. This data would be deliberately thin in terms of attributes to avoid any confidentiality issues for patients. A set of decode values has been used to make sense of the coded values that accompany the patient ID and the event date. The events include a mixture of pathology, radiology and clinical documentation. This data is for a small number of patients only – if the same set of data was extracted for all breast cancer patients, it would come to in excess of a million rows of event data.

3.2 Data Pre-processing

Real-world data is occasionally incomplete (with no values for some variables), noisy (with errors and outliers) and inconsistent (with different categorisation or scale). Data pre-processing aims to prepare the data for data mining, modelling and analytics [4]. Good quality data can be achieved by cleansing which is the process of detecting, correcting or removing corrupt or inaccurate values from a record, table or database. This refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying or deleting the dirty data [13].

General Data Issues

Table 1 shows some general sample issues and solutions at the data pre-processing stage. As examples of outliers in the datasets, some patients were born before 1900 but are still considered to be alive and other patients do not have an initial record of DoB. Some deceased patients have been recorded with their year of death before 1920. A solution here is to consider the Year_of_Birth (e.g.) after 1920 as a cut off, thus to remove those patients and their corresponding cancer records.

Incomplete data is an unavoidable problem in dealing with most real-world data sources. Simply ignoring the instances is not the best way to handle the missing data and the method which has been used here is to identify the most similar attributes to the one with a missing value and deduce the case for it. For example the Deceased_Flag and Year_of_Death will together indicate if the patient has died – sometimes it is known that a patient has died without knowing the date of death with certainty – however, the Deceased_Flag should be "Y" if there is a valid value for Year_of_Death.

Laterality identifies the breast side diagnosed with cancer, with values of: L = Left; R = Right; B = Bilateral (i.e. both sides); 9 = Unknown (suggesting that the definitive diagnosis report is not available, as any valid diagnosis would identify the side); Null = No value was selected during data entry. There are 78 records with tumour size

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

Table 1 Anomaly properties and detection methods

Table Name	Data Issues	Action Required
Patient_Master	Year_of_Birth is unknown	Exclude the records for analysis which need patient's age (group)
	Year_of_Birth range	Exclude the patients (e.g.) born before 1920
	Deceased_Flag='N' but Year_of_Death is certain	Replace Deceased_Flag by 'Y'
Cancer_Detail	Laterality	Can be determined by further checking the value of tumour size on "L" or "R" side
	Missing Cancer_Category	Exclude the records for analysis which need cancer type
Attribute_Data	Attribute_Field= 'Cancer_Surgery_L/R'	Exclude the records for the surgery analysis which have both simple mastectomy and wide local excision on the same date
	Attribute_Value= 'Simple mastectomy' OR 'Wide local excision'	

values for both sides yet the Laterality is not "B" – these have been corrected. In addition, there are 26 instances where the Laterality is "9"; however, further checking of tumour size values suggests that 25 records should be "L" and one should be "R".

Specific Data Issues

As a different data pre-processing example, the success of conservative primary breast cancer surgery (wide local excision) can be compared with radical breast cancer surgery (simple mastectomy). To achieve this a suitable cohort of patients has to be extracted from the SBCDS dataset, with the first step to define a set of criteria to ensure that only appropriate/comparable records are extracted. The next step is to retrieve the required data using SQL statements. During this process, some previously unknown erroneous data was found, e.g. some data suggested that patients could have multiple surgeries on the same date – either multiple occurrences of the same surgery or one surgery type followed by the other. For example, within the same day, one patient has received the wide local excision surgery on the left side and simple mastectomy on the right side.

This issue highlighted episodic attribute data that was orphaned by the original system. It was caused by initial entry (e.g.) on the left side followed by a change to the laterality field and entry of data into the other (right) side, but without removal of the data from the side originally entered (i.e. left). When the record was saved, the

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

spurious data on the unselected side was stored along with the attribute data for the intended side. Roughly about 0.1% records have more than two cases of simple mastectomy or wide local excision on a specific date – this erroneous data can be updated in the database system by use of suitable SQL queries – for the purpose of data analysis in this study, it was decided to simply eliminate this group of data entries.

Normalisation

Breast cancer data represents a very tough challenge to analyse and present in a way that is not confusing and potentially misleading. The data needs a degree of normalisation (i.e. establishing subsets of data that are logical to compare and contrast) before much sense can be made of it. For example, it can be questionable trying to perform analysis on data for deceased and alive patients together and significant thought must be given to what certain event profiles actually mean. It is straightforward to break down the data into distinct groups for further analysis and interpretation. For example, those patients (1) who are still alive; (2) who have died and cancer has been assigned as the cause of death; and (3) who have died, but due to causes that are not considered to be related to cancer.

A first step in making sense of these groups is to factor in the length of time since initial diagnosis. More complex analysis would involve looking at whether longer survival correlates with certain patterns of treatment. Comparing patients based on age group and treatment packages is definitely of interest and one of the most difficult challenges is how to summarise the data to make it understandable without hiding potentially valuable information. The comparison and distribution of survival periods are analysed in the Results section based on the different groups.

3.3 Multi-dimensional Modelling

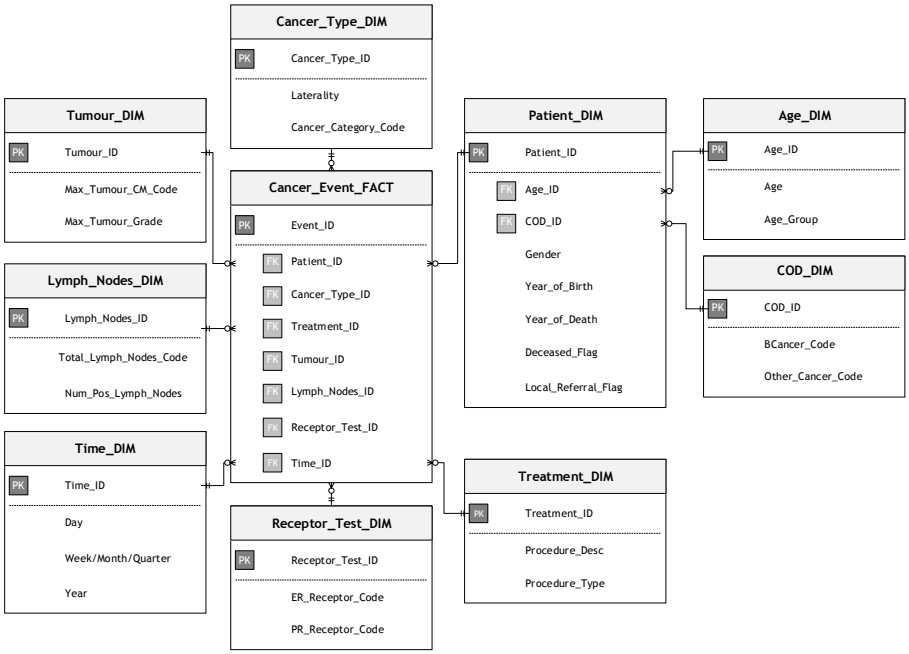
The data to be warehoused includes anonymised patient records and diagnosis records from SBCDS, which also has direct searches of further information from the UHS-CDE, e.g. eDocs, eQuest and the Patient Administration System. It has been extracted from the breast cancer data system and transformed into patient master and cancer detail tables, before loading to the university-hosted data warehouse.

Design and implementation of the data structure is the first step in data warehousing and one of the key aspects for healthcare warehouse design is to find the right scope for different levels of analysis. Questions to ask include which tables are needed and how are they connected to each other. A snowflake schema has been designed initially based on SBCDS and is shown in Figure 4, where the granularity comes from patient events during hospital visits. The schema employs the original structure of the SBCDS data, which contains information about the patient, treatment, time and age etc. These groups are turned into dimensions, as seen in Figure 4, in order to allow an easy way to analyse the data.

In this schema the available data is divided into measures – fixed facts which can be aggregated – and dimensions, which provide ways to filter the available data. It

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

Figure 4 Snowflake Schema example



results in one fact table deriving from cancer events and multiple dimension tables and, when visualised, the fact table is usually in the middle – the schema can be viewed as a snowflake (Figure 4), hence the name [9].

This data warehouse model has facilitated the exchange of health-related information among the Solent Health Informatics Partnership for research purposes. The data in the warehouse is held in a structured form and available for analysis through OLAP and Tableau. After implementing the design, a working data warehouse has been created. Users are then able to look at the summarised or aggregated data to various levels – through joining the fact table to the selected dimension tables (e.g. patient, treatment, time, age, tumour, cancer_type etc.).

4 Results

This section highlights what can be achieved through visual analytics and data mining on the SBCDS dataset, in particular by using the Tableau and Weka software. These results are indicative only, serving to illustrate the general approach and techniques described. The focus for data mining and modelling will be on sequential patterns mining and classification.

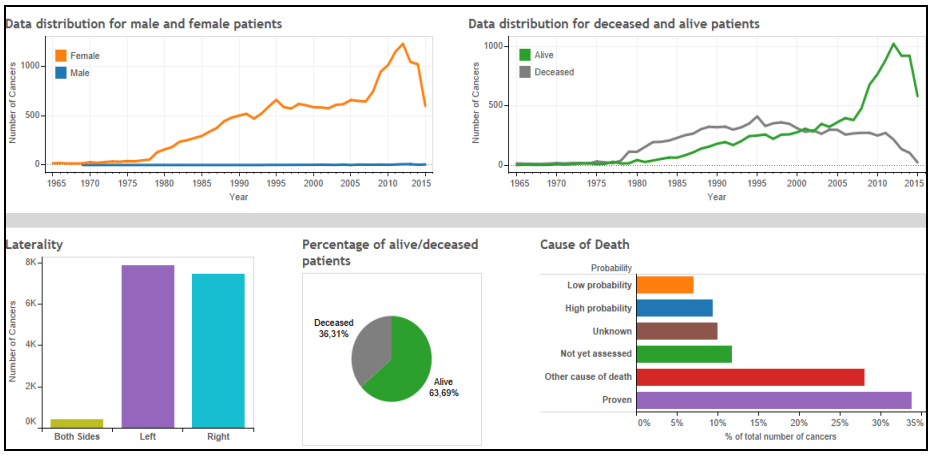
The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

4.1 Visual Analytics

Dashboard

Dashboards are typically used as a means of displaying live data and each dashboard is a collection of individual indicators, designed in such a way that their significance can be grasped quickly. Figure 5 gives an example of the diagnosis dashboard from Tableau and contains five charts each demonstrating a different situation.

Figure 5 SBCDS Dashboard example



The first chart at the top left of Figure 5 shows the data distribution for both male and female patients. Indeed men can develop breast cancer as well, in the small amount of breast tissue behind their nipples [1]. There are fewer than 70 male patients with about 100 records in the dataset. The study here focuses on female patients, of mean age 61 years old at the diagnosis.

The second chart at the top right of Figure 5 shows the overall data distribution for deceased and alive patients, where the earliest patient was diagnosed in 1960s and the most recent one was in November 2015. The rest of the dashboard shows other general information about the patients recorded in the data warehouse, i.e. laterality of the breast cancer, percentage of alive/deceased patients and probability that cancer is the cause of death.

Survival Analysis – Time and Age Groups

Analysing the survival time of patients is more complicated than creating the basic dashboard. First, before beginning the process, groups are needed to compare the survival time – e.g. using age groups and grouping the diagnosis time into decades. As a basic filter, only deceased patients and those who are diagnosed with primary breast cancer are counted for these diagrams, which results in 5,836 patient records. Depending on their age some of these patients are also filtered out. Using Tableau,

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

Figure 6 suggests that the trend of survival time for this cohort of patients has improved in the past four decades, both in the short term and (cumulatively) in the longer term.

Figure 6 Survival time comparision by decade and age group

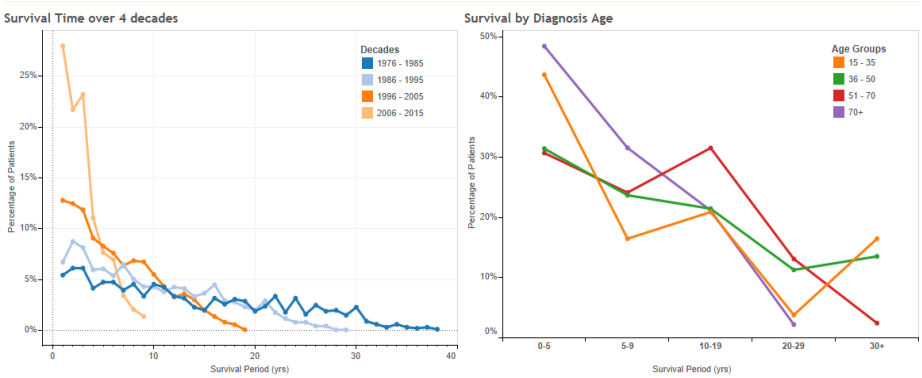


Figure 6 [Left] illustrates how long patients have survived after breast cancer is initially diagnosed. One inference is that patients diagnosed later can survive longer. This is best seen from 0 to 10 years – after that many of the patients with a later diagnosed cancer are still alive. This is also shown in the relatively steep fall for the 2006-2015 decade, which corresponds to the newest diagnosed patients. These diagnoses are too recent to have a longer survival time.

Age_at_Diagnosis is one of the key sub-divisions of the data. Much is already known (whether proven or not) about how the age at diagnosis influences survival prospects. It would be expected that patients who are diagnosed with cancer at an early age often die more quickly than patients who are diagnosed later in life – Figure 6 [Right] illustrates this hypothesis. The graph shows the percentage of patients overall who lived for a certain period of time divided by age group, giving a similar trend to the previous diagram.

Survival Analysis – Sample Treatments

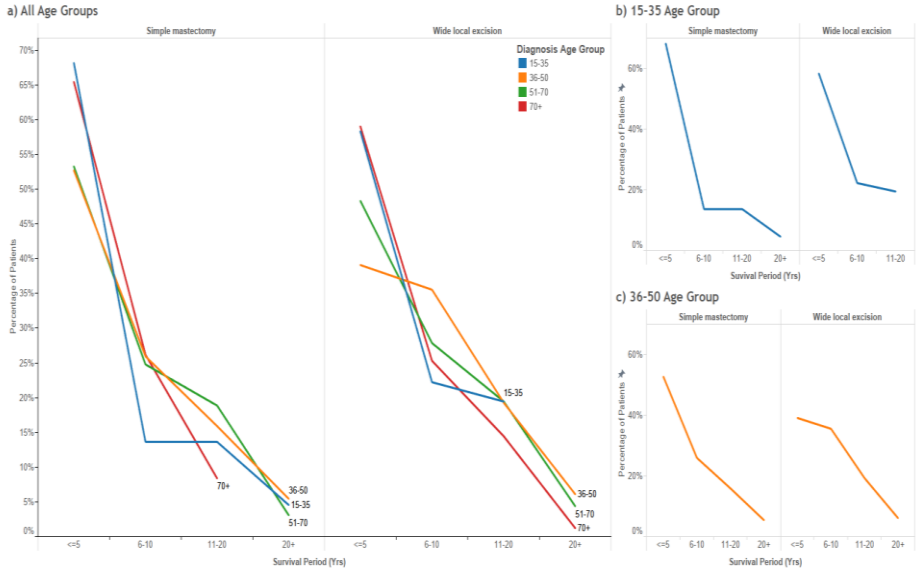
The example here compares two different types of breast cancer surgery. The patient age-bands and survival period groups are defined using Tableau's grouping functionality. They are then used to produce visuals comparing survival time against the two different surgery types, with a much greater success than the raw values.

The graph produced in Figure 7(a) shows the percentage of patients for each sub-cohort (age-band), rather than a percentage of the total. This illustrates that survival statistics for each surgery type can be compared appropriately, by looking at the percentage of patients in each age-band and survival period groups.

Figure 7(b) shows the survival statistics for age-band 15-35. Both surgeries have a relatively high percentage of patients who survive less than five years. Figure 7(c)

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

Figure 7 Wide local excision vs simple mastectomy



displays the same data but for age-band 36-50. The key difference here concerns the 6-10 years survival group, which has a greater percentage of patients following a wide local excision than a simple mastectomy.

Typically the graphs follow the common hypothesis that patients with cancer diagnosed at a young age have a reduced survival period. On the whole a greater percentage of patients are present in the 6-10, 11-20 and 20+ years groups for wide local excision, which could suggest that for this dataset the surgery is performing better than simple mastectomy. It is hard to derive any real conclusion from this analysis of course, but it does provide insight into which areas to concentrate further research.

4.2 Data Mining and Analytics

Sequential Patterns Mining

Querying the patient profiles is the starting point before pre-processing for sequential patterns mining. Based on the data model from Figure 3, the relevant attributes selected for this purpose are: BCI_ID, Cancer_Category_Desc, Year_of_Death and Diagnosis_Month. Table 2 shows some raw output from an SQL query against the data extracted for the case study – the data has been limited to local referrals. Due to the limit of the table length, the query interleaves the date of each disease presentation event to give up to five event type/date pairs after the initial diagnosis.

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

Table 2. Sample results of patient event profiles

INITIAL	PRES1	PRES2	PRES3	PRES4	PRES5	STATUS
Primary	Loco-RR	Loco-RR	Risk-Reduce	-	-	Alive
Primary	Loco-RR	Other	Loco-RR	Metastatic	-	Dead
Primary	Loco-RR	Primary	Loco-RR	Loco-RR	Loco-RR	Dead
Primary	Metastatic	Metastatic	Metastatic	Metastatic	Metastatic	Dead
Primary	Other	Loco-RR	Loco-RR	Loco-RR	-	Alive
Primary	Primary	Metastatic	Other	Metastatic	Metastatic	Dead

Key: Primary Breast Cancer (Primary), Loco-Regional Recurrence (Loco-RR), Metastatic Disease (Metastatic), Risk Reducing Mastectomy (Risk-Reduce), Other Cancer Diagnoses (Other)

For the November 2015 dataset, there are 178 distinct disease event sequence profiles which correspond to 12,139 instances. The following pre-processing approach has been pursued to ensure the data is represented as accurately as possible: removal of instances where (1) there is no presentation at all; (2) initial presentation is anything other than primary breast cancer; (3) two or more presentations of primary cancer exist (when cancer is unilateral); and (4) the total number of events is less than three.

This dataset is then divided into two sub-groups: alive (188) and deceased (1,957). The GSP (Generalized Sequential Patterns) algorithm has been used through Weka for sequential patterns mining [3]. Five sequential patterns are shown below for alive patients under a minimum support threshold of *minsup*=5%, where the numbers of patients are in brackets.

[1] <{Primary}{Loco-RR}{Loco-RR}> (39)
[2] <{Primary}{Loco-RR}{Metastatic}> (20)
[3] <{Primary}{Loco-RR}{Other}> (14)
[4] <{Primary}{Metastatic}{Metastatic}> (12)
[5] <{Primary}{Other}{Loco-RR}> (10)

These results are maximal patterns, i.e. they are not contained by other sequential patterns. When the same mining approach is applied to deceased patients and under the same setting, i.e. *minsup*=5%, there are three sequential patterns:

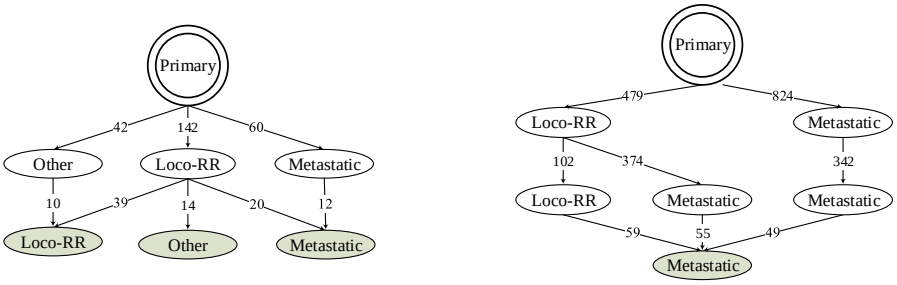
<{Primary}{Loco-RR}{Loco-RR}{Metastatic}> (59)
<{Primary}{Loco-RR}{Metastatic}{Metastatic}> (55)
<{Primary}{Metastatic}{Metastatic}{Metastatic}> (49)

A directed acyclic Sequential Patterns Graph (SPG) has been used to represent the maximal sequence patterns [11]. Figure 8 shows the SPGs for both alive and deceased patients when *minsup*=5%. It can be seen that nodes of SPG correspond to elements (or disease events) in a sequential pattern and directed edges are used to denote the sequence relation between two elements. Any path from a start node to a final node corresponds to one maximal sequence and can be considered optimal.

Taking the left-side path from Figure 8(a) as a sample for illustration, 42 patients are found with the diagnosis pattern of <{Primary}{Other}> with a support of at least

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

Figure 8 SPG for maximal sequential patterns when *minsup*=5%



(a) Alive (b) Deceased

5% – out of this group there are another 10 presentations with Loco-Regional Recurrence. Down the left-side of Figure 8(b), there are 479 patients with the diagnosis pattern of <{Primary}{Loco-RR}> – while 102 of these cases further present with Loco-Regional Recurrence, there are 374 instances of Metastatic Disease at the same level. These proceed respectively to another 59/55 presentations of Metastatic Disease.

Classification

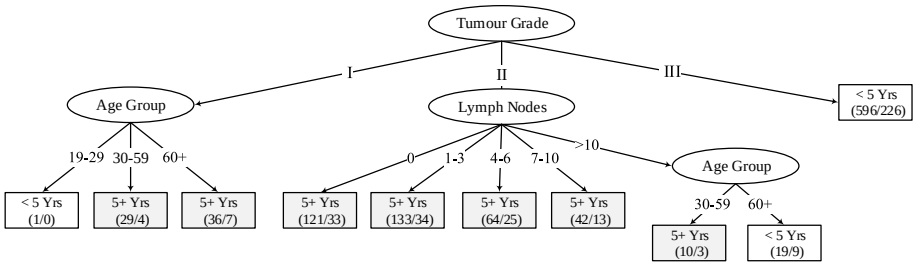
All of the classification algorithms have been evaluated using Weka and compared according to a number of measures: (1) accuracy, sensitivity and specificity, (2) n-fold cross validation and (3) receiver operating characteristic. In this case study, a predictive model was created by applying the Weka Decision Tree J48 algorithm to the prepared dataset for the deceased patients whose cause of death was breast cancer on either the left or right side (i.e. not both). The 10-fold cross validation technique was used where the data was divided into 10 sub-sets of the same size.

The following variables have been selected for the predictor set: Max Tumour Grade, Positive Lymph Nodes and Age Group. The outcome status is either "Survive more than 5 years" or "Survive less than 5 years". There were 1,053 records containing full values for all the above variables. The complete decision tree is shown in Figure 9 with 10 rectangular boxes representing the leaf nodes (i.e. classes). Each internal node represents a test on the variable and each branch represents an outcome for the test. There are two numbers in parentheses in the leaf nodes: the first shows how many patients reached this outcome and the second shows the number of patients for whom the outcome was not predicted to happen.

Based on Max Tumour Grade – an indicator of how quickly a tumour is likely to grow and spread – some rules can be extracted from the decision tree in Figure 9. For example (1) if the tumour is Grade III then the patient is likely to survive less than 5 years; (2) if Grade II and the number of positive lymph nodes is no more than 10, then the patient is likely to survive more than 5 years; and (3) if Grade I and the patient age group is '19-29', then they are likely to survive less than 5 years.

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

Figure 9 Decision tree for survival analysis



4.3 Discussion

It should be recognised that the modelling of cancer patient records at University Hospital Southampton is to a large extent feasible because of the adherence to a strict IT strategy for over 20 years. A fundamental challenge to analysing clinical data arises from the need to correctly and consistently identify the patient within the variety of systems which hold the data of interest. For a large organisation like UHS, which has a master patient index (MPI) containing well over 2 million individuals, correct identification against existing records is a constant challenge. UHS has pursued a range of approaches to achieving a high-quality MPI with low levels of duplicates and accurate patient details.

Another UHS/IT strategy that is pivotal to moving ahead with clinical research and data analysis is the relentless drive to concentrate data in relational databases, mainly Oracle. An integration engine (e.g. Ensemble) and HL7 messaging have also been used to facilitate high quality data exchange between the systems employed to deliver healthcare. Good quality data is acquired and maintained by good planning, prescribed and proven operational procedures, and professional IT development and systems management – good quality analytical work and sound conclusions can only come from such good quality data.

Temporally-structured clinical datasets are a relatively new tool in mainstream clinical practice. They pose significant challenges in data representation and analysis. The Southampton Breast Cancer Data System is a wholly new, very large and complex data presentation and analysis system which is in continual evolution and validation. It provides an opportunity for experimentation with a range of data mining software tools and concepts, including the use of technology to identify outlier and "illogical" cohorts of patients with erroneous data.

Much thought must be given when deciding which data items to select for the proposed analyses and illustrations. Diagnostic tests are performed on patients for many reasons, and many patients suffering from a life threatening disease such as cancer may have other chronic and/or acute problems which almost certainly have little in common with the cancer. One must also consider that clinical data which could be of great value to the prospective analyses may not be available, either

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

because it is not stored digitally or cannot be acquired. In the case of the UHS Lifelines concept, chemotherapy and radiotherapy data would be a valuable addition to the otherwise fairly complete dataset and it is hoped that this data will be forthcoming in the future.

Having undertaken appropriate pre-processing above, it was possible to perform sophisticated analyses on temporally-structured data using the software tools described. This work is iterative and with several objectives – one goal is to identify data mining tools which can be integrated into the bespoke data systems within the UHS-CDE. In time, this will allow clinicians more readily to analyse and understand the consequences of treatment decisions and their clinical outcomes across a spectrum of complex and chronic diseases.

Visual analytics is also becoming one of the essential techniques for health informatics. It allows users (clinicians, researchers, administrators and patients) to derive actionable and meaningful insights from vast and complex healthcare data. The use of software such as Tableau has the potential to improve graphical output from SBCDS, e.g. for histograms of year-on-year survival for any defined cohort. In addition, enhancement of timeline visualisation using the MVC architecture could provide an elegant solution which would allow UHS to move forward with Lifelines as a universal tool and testbed.

5 Conclusion

Within the healthcare domain there has been a vast amount of complex data generated and developed over the years through electronic patient records, disease diagnoses, hospital resources and medical devices. This data is itself a key resource and could play a vital role enabling support for decision making through processing information for knowledge extraction. The growing amount of data exceeds the ability of traditional methods for data analysis and this has led to the increasing use of emerging technologies. An overview process-driven framework has been proposed which integrates the pre-processing, data warehousing, data mining and visualisation aspects.

As part of the University Hospital Southampton clinical data environment, the Southampton Breast Cancer Data System has been developed as a "proof of concept" system. SBCDS is a unique timeline and episode-structured system for the analysis of the entire breast cancer pathway to final outcome of thousands of locally treated patients. There are already some valuable and complex analyses that have been developed within SBCDS, and there is potential for further growth in functionality and capability of the system.

UHS Lifelines provides a conceptual model for the time-structured presentation of all key data for any patient or any chronic condition on a single computer screen. In particular, a Cancer-Lifetrak timeline has been developed to highlight the month of onset of key episodes of breast cancer progression, diagnosis, local recurrence, metastasis etc. It also permits measurement of time intervals between episodes and the correlation of these intervals with pathology and treatments.

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

One challenge of the Lifetrak representation is the overloading of the graphical interface by a concentration of many events over a relatively short period. A practical graphical user interface approach is thus needed which will handle this situation, so that the overriding story told by the data is not lost or corrupted. A sample dataset could be extracted for this purpose which includes the range of clinical data often associated with cancer patients, e.g. a mixture of pathology, radiology and clinical documentation events. The full set of records for all breast cancer patients comes to in excess of a million rows of event data and presents a bigger challenge for data management and predictive modelling.

Big data in healthcare is overwhelming not only because of its volume but also the diversity of data types and the speed at which it must be managed. The emerging NoSQL databases have significant advantages such as easy and automatic scaling, better performance and high availability. Using big data technologies has the potential to lead to more efficient and flexible healthcare applications. There are several challenges that need to be addressed to maximise the benefits of big data analytics in this area [24]: data quality; developing reliable inference methods to inform decisions; implementation of trusted research platforms; data analyst capacity; and clear communication of analytical results.

This paper has sought to give insight into how retrospective analysis of cancer treatment has the potential to identify some treatment pathways as more successful in terms of statistical outcome than others. The application of data mining and visual analytics could help prevent patients from being given sub-optimal treatment and help focus resources on treatments that are more successful. The cumulative benefits, both human and financial, might be enormous and it is hoped that others will take up the challenge to develop analyses where data is available to support that goal.

Acknowledgements. This research project has been supported in part by a Southampton Solent Research Innovation and Knowledge Exchange (RIKE) award for "Solent Health Informatics Partnership" (Project ID: 1326). The authors would like to thank Solent students who made some contribution to the work: in particular Chantel Biddle, Adam Kershaw and Alex Potter. We are also pleased to acknowledge the generous support of colleagues in the University Hospital Southampton Informatics Team, in particular Adrian Byrne and David Cable.

References

1. Bonadonna G, Hortobagyi, GN, Valagussa P (2006) Textbook of breast cancer: A clinical guide to therapy. CRC Press
2. Devi RDH, Deepika P (2015) Performance comparison of various clustering techniques for diagnosis of breast cancer. IEEE International Conference on Computational Intelligence and Computing Research, pp 1-5
3. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: An update. SIGKDD Explorations 11(1): 10-18
4. Han JW, Kamber M, Pei J (2011) Data mining: Concepts and techniques. Elsevier
5. Hand DJ, Smyth P, Mannila H (2001) Principles of data mining. MIT Press Cambridge, USA

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4

6. Holzinger A (2014) Trends in interactive knowledge discovery for personalized medicine: Cognitive science meets machine learning. IEEE Intelligent Informatics Bulletin 15(1): 6-14
7. Hu H, Correll M, Kvecher L, Osmond M, Clark J, et al (2011) DW4TR: A data warehouse for translational research. Journal of Biomedical Informatics 44(6): 1004-1019
8. Jerez-Aragones JM, Gomez-Ruiz JA, Ramos-Jimenez G, et al (2003) A combined neural network and decision trees model for prognosis of breast cancer relapse. Artificial Intelligence in Medicine 27(1): 45-63
9. Kimball R, Ross M (2013) The data warehouse toolkit – the definitive guide to dimensional modeling. John Wiley & Sons, New York
10. Lee YJ, Mangasarian OL, Wolberg WH (2003) Survival-time classification of breast cancer patients. Computational Optimization and Applications 25(1-3): 151-166
11. Lu J, Chen WR, Adjei O, Keech M (2008) Sequential patterns post-processing for structural relation patterns mining. International Journal of Data Warehousing and Mining 4(3): 71-89. IGI Global, Hershey, Pennsylvania
12. Lu J, Hales A, Rew D, Keech M, Fröhlingsdorf C, Mills-Mullett A, Wette C (2015) Data mining techniques in health informatics: A case study from breast cancer research. 6th International Conference on IT in Bio- and Medical Informatics, LNCS 9267, Springer International Publishing Switzerland, pp 56-70
13. Lu J, Hales A, Rew D, Keech M (2016) Timeline and episode-structured clinical data: Pre-processing for data mining and analytics. 32nd IEEE International Conference on Data Engineering (ICDE) – Workshop on Health Data Management and Mining, pp 64-67
14. Mahajan R, Shneiderman B (1997) Visual and textual consistency checking tools for graphical user interfaces. IEEE Transactions on Software Engineering 23(11): 722-735
15. Marr B (2015) Big Data: Using smart big data analytics and metrics to make better decisions and improve performance. Chichester, Wiley
16. Martin MA, Meyricke R, O'Neill T, Roberts S (2006) Mastectomy or breast conserving surgery? Factors affecting type of surgical treatment for breast cancer: A classification tree approach. BMC Cancer 6: 98
17. National Information Board (2014) Personalised Health and Care 2020. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/384650/NIB_Report.pdf
18. NHS (2014) Five year forward view. Available from: http://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf
19. Razavi AR, Gill H, Ahlfeldt H, Shahsavar N (2007) Predicting metastasis in breast cancer: Comparing a decision tree with domain experts. J. Med Syst 31: 263-273
20. Reenskaug T, Coplien J (2009) The DCI architecture: A new vision of object-oriented programming. Available from: http://www.artima.com/articles/dci_vision.html
21. Reps J, Garibaldi JM, Aickelin U, Soria D, Gibson JE, Hubbard RB (2012) Discovering sequential patterns in a UK general practice database. IEEE-EMBS International Conference on Biomedical and Health Informatics, pp 960-963
22. Rew D (2015) Issues in professional practice: The clinical informatics revolution. Published by Association of Surgeons of Great Britain and Ireland
23. Stolba N, Tjoa A (2006) The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making. International Journal of Computer Systems Science and Engineering 3(3): 143-148
24. Wyatt J (2016) Plenary Talk: Five big challenges for big health data. 8th IMA Conference on Quantitative Modelling in the Management of Health and Social Care

The final publication is available at link.springer.com via http://dx.doi.org/10.1007/978-3-319-64265-9_4