On Creating a Patient-centric Database from Multiple Hospital Information Systems

J. Bettencourt-Silva1; B. De La Iglesia1; S. Donell2; V. Rayward-Smith1
1School of Computing Sciences, University of East Anglia, Norwich, United Kingdom; 2Faculty of Health, University of East Anglia, Norwich, United Kingdom

Keywords
Data collection, data retrieval, hospital information systems, methods

Summary
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.

1. Introduction

The use of electronic medical records (EMR) in clinical research has already been envisaged by the medical informatics community [1–3]. Indeed, different types of EMR systems can be used to select and, in some cases, retrieve data for epidemiological studies [4], clinical trials or other clinical study data management systems. In longitudinal or other observational studies, however, legacy patient information may be collected from regional or national data sources such as registries. The latter are 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 United Kingdom, within the National Health Service (NHS), 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) to manage EMRs. HIS were initially designed to support and monitor patient care, specific medical tasks and hospital management [5]. Only recently has attention been given to clinical research [6] and other secondary uses of such patient data [7] for which data extraction is key.

Historically, HIS developed from central to modular and distributed systems [8]. 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 [8, 9]. Similarly, the distribution of information processing and uniqueness of medical data [9] pose obstacles to data integration [8]. Commercially available HIS often focus on administrative tasks and lack knowledge-based functionality [10]. Hospitals may opt for implementing several commercial departmental systems, creating “islands” of information across various departments [11], 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 [12].
Extracting patient-centric data and EMRs from one or multiple HIS in a secondary care centre is often an ad-hoc process and 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 [13]. Furthermore, existing methodologies guiding knowledge discovery or data analysis projects fail to mention and examine data collection (retrieval or extraction) as a step in its own right.

The aim of this paper is to identify and expose the current obstacles and drawbacks of exploiting multiple HIS to retrieve and compile patient-centric data into a research database, and to facilitate future data collection and the use of electronic patient data in research.

A roadmap or process for data collection has been constructed based upon expert consultation and the experience drawn from the selected case study. Previous work on knowledge discovery in databases (KDD) methodologies [14, 15], KDD methodologies tailored to medical domains [16], and work on clinical data integration [17] were important in shaping the roadmap presented in this paper.

In order to capture a wide spectrum of patient-centric data from multiple systems, a case study on carcinoma of the prostate was carried out in at the Norfolk & Norwich University Hospital, a regional English NHS tertiary care centre.

This hospital provides secondary and tertiary care services to a population of over half a million people in the East of England, and the expected annual incidence of prostate cancer patients in this particular hospital is estimated to be about 300.

2. Background and Objectives

The approach to collecting patient-centric data from multiple HIS was drafted by initial consultation with domain experts as well as input from relevant segments of other methodologies such as the KDD roadmap [14] and the 6-Step DMKD [16]. The latter has been thoroughly applied to medical domains. Because collecting data from multiple sources ultimately results in data integration, input from previous work on this topic [17] was also crucial.

KDD methodologies have been developed to provide guidance to data mining and other data analysis and knowledge discovery projects using large databases. The common steps in such methodologies [14, 18] are (Fig. 1): domain understanding, data preparation, data mining, and evaluation of discovered knowledge. Data understanding is a step present in all methodologies reviewed and can be seen as a subset of domain understanding, data preparation, or a step on its own, depending on the methodology. The data preparation step may also be referred to as data cleansing or pre-processing. These common steps are illustrated in Figure 1 and provide the framework to embrace the data collection exercise presented in this paper.

A thorough 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 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 roadmap presented in this paper 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 paper focuses primarily on the method for data collection since this has received the least attention in the literature. Clinical outcomes of data analyses are omitted in this paper.

Data collection and preparation in KDD methodologies in medical environments may be achieved by following an information pipeline architecture described in [17]. The pipeline channels data from their sources to an operational data store (ODS) where the information is stored before it is further validated, cleaned and merged dur-
ing the Extract-Transfer-Load (ETL) process to create a core database. The core database implements a data model and is used to extract more specific data marts for reporting, visualisation, analyses and data mining [17].

The data collection process presented in this paper firstly expands 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 [11, 17, 19, 20]. Furthermore, global query systems have been developed [19, 20] 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 structure may pose obstacles to gaining access to systems and patient-centric data.

Knowledge of the different types of heterogeneities in distributed database systems [20] 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; and query languages, where retrieval and manipulation of query results may differ. There is also semantic heterogeneity pertaining to the meaning, interpretation, or intended use of the same related data across systems [20]; and autonomy (design, communication and execution), also leading to heterogeneity in federated database systems because of the organisational entities and their independent control over different systems.

To address some of the above issues, the concept of metadatabases evolved from the traditional data dictionaries [19, 21]. A conceptual data model (CDM) may specify the lifecycle of a database system [21]. 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 [21]. This is particularly important since medical information is essentially bound to the context of its production [3]. Indeed one of the outputs of the data collection process presented in this paper is a metadatabase comprising metadata files for each data source.

One of the benefits of using a metadatabase in the context of the work in this paper 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 [4] 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 [4].

Section 3 defines the problem specification (3.1) and explains the initial process of selecting and identifying the data sources (3.2) and may be regarded generically as domain understanding. It is, however, in the Section 4, that the approach for data collection is thoroughly described. Data cleansing and collation are discussed in Section 5 and evaluation of the work in Section 6.

3. Preliminary Work
3.1 Problem Specification

The criterion used in this case study was included patients diagnosed with prostate cancer (ICD C61.X) during a period of five years from 2001 to 2005. The aim of this study was to use the inclusion criteria to attempt a methodological approach to data collection from multiple sources in a hospital. Once a research protocol comprising this information was agreed with the research team, credentials were sought and approved by the local research ethics committee and research governance committee. This process involved ethical and research governance approvals of a research protocol and was time consuming. It was necessary to exclude patient sensitive data such as names and addresses from this work. The protocol included an extensive literature survey on the problem domain (prostate cancer in this case). The application for credentials also required a list of information systems (data sources) involved in the study. In general, this might not always be known a priori and amendments to the original documents might then be necessary.

3.2 Identifying Data Sources

Depending on the study, different sources may be of different relevance but, in order to provide an accurate and holistic view of each patient, the aim is to collect data from as many sources possible where authoritative information on a patient exists. When so much data is collected and analysed, less biased results are expected as a better understanding of issues such as comorbidity factors and the pathway of a patient is possible. However, it has been argued that requiring data with no practical importance creates a source of misleading information [17]. Efforts need to be put towards the definition of practical importance or relevance of data required for the study.

The selected case study on prostate cancer involved 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 was needed. Data elements for the sole purpose of validating or linking data needed to be included as well as other fields or datasets that may have had an impact on the understanding of the problem domain. At this 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 and further specific data about the sources would be later identified as that is also one of the purposes of the process described in Section 4.

A conceptual data model (CDM) may also be of use in the identification of the variables for a study and those that are out-
side of it [17] but this requires previous knowledge of the sources and data elements. The approach presented in this paper 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.

In the case study, seven authoritative data sources were identified as of interest but because of authorisation constraints, only four were used. This was considered sufficient to illustrate the issues of retrieving data from multiple sources in hospitals. The aim in this paper is to work with local data and, for this reason, national or regional data sources are disregarded for the most part within our methodological approach. Nevertheless, this is later discussed as a way of validating the study data.

Gathering information on clinical coding was important at an initial stage so that cohorts of target patients could be retrieved first. This was important because not all sources provide accurate clinical coding or another way of identifying the target population.

As described in Table 1, out of the four 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.

### 3.3 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 to be retrieved. The hospital Patient Administration System (PAS) contains important appointment information for all patients as well as clinical coding (ICD-10 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 histopathological reports with respective tumour markers. The remaining three systems were identified as of importance but, for different reasons, their data was not included in this case study. The Theatre and Radiotherapy systems were difficult to access due to authorisation issues, and the cancer register system maintained by the information department was recently implemented and did not include enough retrospective data. Nevertheless, reports from the operating theatre as well as radiotherapy treatments and results are an important part of the pathway for patients with cancer. The cancer register system managed by information services appeared to have the potential to store a complete dataset for cancer; however, at the time of writing, this system had been used primarily for monitoring cancer targets.

### 4. Data Collection

Once the data sources are identified, the process described in Figure 2 below illustrates the steps taken to extract data from a single source. 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 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 4.1), data understanding (Section 4.2), extraction preparation (Section 4.3) and extraction & evaluation (Section 4.4).

### 4.1 System Understanding

#### 4.1.1 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 contains 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 (as described in step 4.1.2 below). 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 3 is revisited and the process in Figure 2 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 obtain 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 understanding 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 our study, we 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 may occur on a patient record after its retrieval.

### 4.1.2 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 hospital’s 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 (step 4.1.1), understanding the data (step 4.2.1 and 4.2.2) and obtaining system access (step 4.1.3). 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 our 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.

### 4.1.3 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 4.1.2 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 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.

### 4.2 Data Understanding

#### 4.2.1 Data Familiarisation and Understanding

The identification and ranking of the relevant data fields from the source is essential [18] and can first be achieved by familiarisation using the system’s GUI together with any data extraction tools when available.

A set of examples to achieve database familiarisation has been described in [14] and, from these, the following are considered particularly 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 [14], 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. One of the issues arising from the case study was that decimal places are handled differently across data sources. The extracted datasets were formatted at a later stage so that the retrieved field type represents the true data type.

- **Determination of database field semantics**
  Two or more fields, although different, may be based on the same or similar measurements [14]. An understanding of the data fields is therefore important and detailed explanations should be in-

---

**Table 2** Summary of items gathered when understanding the system

<table>
<thead>
<tr>
<th>Metadata – Understanding the System</th>
<th>Examples from the case study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriateness and authoritativeness of system</td>
<td>All identified systems are authoritative and contain crucial information on prostate cancers.</td>
</tr>
<tr>
<td>Data flow chart</td>
<td>Overall understanding of hospital information systems’ data flow.</td>
</tr>
<tr>
<td>Data extraction method (if known)</td>
<td>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</td>
</tr>
<tr>
<td>Preview and system training</td>
<td>Previewed all systems, training on PAS, PACS, LAB.</td>
</tr>
<tr>
<td>System credentials</td>
<td>Read-only access to the systems was issued.</td>
</tr>
<tr>
<td>Key data elements present (if known)</td>
<td>Hospital number, NHS number, Date of Birth present on all systems.</td>
</tr>
<tr>
<td>System Updates (impact)</td>
<td>All systems were live and updates occurred daily, but no impact on study data.</td>
</tr>
<tr>
<td>System Limitations</td>
<td>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).</td>
</tr>
<tr>
<td>System Live Date</td>
<td>Date when system went live.</td>
</tr>
<tr>
<td>Last System Update Date</td>
<td>Date when system was last updated (impact on data fields and values).</td>
</tr>
</tbody>
</table>
included in the data dictionary. It is also important to expand any abbreviated field names. During the case study, the most common issues regarding field semantics were fields that concatenate clinical or procedure coding together, and fields that represent the same value but in different formats (e.g. dates).

- **Determination of field value semantics and their plausibility** [16, 18]
  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 [4] and, if so, whether it is possible for it to be recovered or estimated at a later stage.

- **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 [17], these need to be sought and ranked 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. When the data collection process (Fig. 2) runs at least two complete iterations, hence evaluating more than one system, it is possible to compare the focal data elements across sources’ metadata.

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.

An example from the case study is given in Table 3 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).

### 4.2.2 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 metadata database is created (Section 5). 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 4.

### 4.3 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 ven-
4.4 Extraction and Evaluation

4.4.1 Cross-Validation

Using the method selected in step 4.3, 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 [4]. This can be investigated further in the data quality step (step 4.4.2) but, in the current step, the major concern is to ensure the data retrieval method is extracting the desired data.

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 our case study).

4.4.2 Extract Finalised Dataset and Data Quality

Upon a satisfactory cross-validation, the final dataset is extracted using the method selected in 4.3. 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. Indeed, such statistics may suggest that data cleansing is necessary and this is often covered in the next steps of DMKD methodologies (data cleansing and pre-processing). Nevertheless, it is important to include a brief step for checking quality assurance at the end of the process 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 [4]. The most relevant data cleansing operations in the course of the case study were:

- **Outlier handling**
  Many outliers may be classified as either errors or groups of interest [14] and, at this stage, only erroneous outliers are dealt with as the latter 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. 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**
  Missing values may be removed, estimated, or simply marked as missing. Estimation, however, may need to be performed in a separate (training) dataset so

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Extraction method used in the case study</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAS</td>
<td>OALP 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.</td>
</tr>
<tr>
<td>PACS</td>
<td>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.</td>
</tr>
<tr>
<td>ONC</td>
<td>Due to access restrictions, the domain expert agreed to run a back-end SQL query to search for out cohort.</td>
</tr>
<tr>
<td>LAB</td>
<td>The same OALP software as above was used but using a different schema to access a different data source and this required different access credentials.</td>
</tr>
</tbody>
</table>
that the impact can be assessed before the data is changed. A simple but risky approach [14] 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 [24], and more complex, may involve training a neural network to predict missing values using the remaining fields [25].

When possible, a quantification of the amount of missing and erroneous data in a dataset is important metadata and should be added to the metadata.

5. From Data to Operational Data Store

One of the two outputs of the process described in Figure 2 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 3, below.

![Fig. 3 Simplified schema for a metadatabase containing multiple sources and their respective data fields. Table sources (left) contains the list of sources retrieved together with their metadata, and table data fields (right) contains the attributes collected for each data source, with respective metadata.](image-url)

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 particular data transformations were carried out to build the ODS. The purpose of this was to facilitate further data to be added to the ODS in a longitudinal fashion. 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 domains from. It is also expected that most of data understanding already occurred as part of the process presented in this paper. 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 [14, 16, 17].

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 our 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 de-identified 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 to heterogeneous data integration [26]. Schema matching is possible using the metadata collected for each system available in the metadatabase but was not used during the case study.

6. Discussion and Evaluation

The collection of patient-centric data from secondary centres, such as the one described in this paper, 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, together with a case study on prostate cancer, provided the framework for the development of the patient-centric data collection process presented in this paper (Fig. 2).

The process exposes the lengthy journey of retrieving patient-centric multi-source data for research or other secondary uses, and it was tested using several data sources from a regional English hospital. A metadatabase (Fig. 3) as well as study datasets were successfully created as outputs of this process. Indeed it has been identified that the collection of metadata and creation of a metadatabase in parallel with the study database are crucial to its understanding, linkage, and validation.

Another important outcome was the contextualisation of the retrieved data, which 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).

Overall the case study showed the collected data was of acceptable quality when it came 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 320,000. It is common, however, to find greater amounts of erroneous and missing data in legacy systems collected by clinicians [16]. It is also common to find non-focal data elements missing, which will not necessarily impact linkage, but may influence the study results.

Because each data source is invariably linked to a HIS, the work carried out in this paper has also confirmed that the current HIS implemented in a 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 where an additional software tool had to be developed to extract the data.

On average for each data source that was put through the process, the most time consuming step was system understanding where a considerable amount of time was spent liaising with hospital IT managers and other staff in order to obtain access credentials. The second most time consuming step was extraction preparation where a significant amount of time was spent on finding appropriate ways to extract the data. Understanding the data and extract and evaluation were, respectively, the least time consuming steps. Nevertheless it is expected that in other studies, understanding the data may be one of the most time consuming steps.

The process and its steps cover all actions taken whilst carrying out the case study. The sequence in which the steps are followed varies significantly from data source and for this reason several recursive links (feedback mechanisms) were added. Given the similarities across secondary care centres and the way in which the process was designed, accounting for a thorough review of the relevant literature, it is expected that the process can be applied to most data retrieval exercises in a similar setting. Nevertheless, the fact that it was tested in a single secondary centre may introduce limitations to its generalisation. Further work is needed to validate the process in other settings and with other data mining or data analysis methodologies.

7. Conclusions

Collecting routine data from secondary care centres is a non-trivial task, poorly described in the literature. The work presented in this paper focuses on identifying and understanding the common steps to retrieve such data to create a patient-centric database for research; data analyses or knowledge discovery. A process to facilitate such work is introduced. It is hoped that, by exposing the complexity of data retrieval in hospitals through the steps of the process, this experience contributes to research and future data analysis studies, and that analysts are able to ensure the context of their data is fully understood.

Data ownership issues and related bureaucracy still pose one of the major obstacles to carrying out work such as this. It is our belief that research studies should be encouraged to use routine 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 reassure 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.

Furthermore, the practical obstacles of working in a heterogeneous multi-source environment where data access is limited have been addressed and the key informational requirements for a successful data collection task in hospitals have been introduced. It has also been identified that it is possible to obtain reliable, routinely collected, data from a secondary care centre for research. Further work on the issues of extracting patient-centric data from other sources and settings is needed to encourage the secondary uses of such data.
Acknowledgments
This work was supported by grant FSE-POPH-BD-43770-2008.

References
23. Bettencourt J. Extracting Patient-Centric Data from the NHS: A Case Study in Prostate Cancer at the Norfolk & Norwich University Hospital. School of Computing Sciences, University of East Anglia, Norwich; 2009.