PARTNER
Grant Agreement Number 215840

WP24 - D.1 Analysis Report of Existing Databases of Tumours
WP24 – D.2 Inventory of Existing Databases in Collaboration with Centres of the PARTNER Consortium

Vassiliki Kanellopoulos
CERN

Manjit Dosanjh, CERN
Karen Kirkby, University of Surrey
Norman Kirkby, University of Surrey
Rajesh Jena, Cambridge University Hospitals NHS Foundation Trust

May 30th 2010
Contents

1 Introduction 3

2 Tumour Documentation 3
   2.1 Tumour Documentation system 3
   2.2 Cancer Registries 4
   2.3 Clinical Registries Software 5
      2.3.1 ODSeasy, Germany 5
      2.3.2 Giessener Tumordokumentationsystem GTDS, Germany 6
      2.3.3 The SEER Data Management System, US 7

3 Example Tumour Databases 8
   3.1 Gene Databases 8
      3.1.1 IARC TP53 Mutation Database 8
      3.1.2 Human Tumour DataBase, France 8
   3.2 Single Projects - Treatment Outcome and Prognosis 9
      3.2.1 Database Projects by the Grampian University NHS Hospitals Trust, Scotland 9
      3.2.2 Non-functioning Pituitary Adenoma Database, Italy 13
      3.2.3 TumorAGENT, Germany 13
      3.2.4 The Lung Cancer Database, Japan 14
   3.3 National Cancer Registries 15
      3.3.1 ECRIC (Eastern Cancer Registry and Information Centre), UK 15
      3.3.2 RTDS (Radiotherapy Database Project), UK 16
      3.3.3 CBTRUS (Central Brain Tumour Registry of the United States), USA 17
      3.3.4 SEER Cancer Database (Surveillance, Epidemiology, and End Results), USA 17
      3.3.5 NCDB (National Cancer Database), USA 18
      3.3.6 NODB (National Oncology Database), USA 19
   3.4 European Databases 19
      3.4.1 GENEPI Entb1&2 (ESTRO, European Universities & Institutions) 19
      3.4.2 Conticabase and Conticagist 21
      3.4.3 EUROCARE (EUROpean CANcer REgistry-based study on survival and CARE of cancer patients) 22

4 Grid-related Cancer Infrastructures 23
4.1 CaBIG (Cancer Biomedical Informatics Grid), US .................................. 23
  4.1.1 EyecancerBig ................................................................. 24
  4.1.2 LEAD (Lymphoma Enterprise Architecture Database Platform) ........... 24
4.2 BioGrid Australia ................................................................. 25
  4.2.1 ACCORD (Australian Comprehensive Cancer Outcomes and Research Database) ................................................................. 26
  4.2.2 CART-WHEEL (Center for Analysis of Rare Tumours) ..................... 28

5 Conclusion ................................................................. 29

Appendix ........................................................................... 31

References ...................................................................... 34
1 Introduction

Knowledge discovery can be defined as "the non-trivial process of identifying valid, novel potential useful, and ultimately understandable patterns in data"[1]. Knowledge discovery is very important in the field of medicine and with the vast amount of data taken today, data-mining\(^1\) techniques become increasingly relevant. Previously, the examination of large datasets, e.g. biomedical data, was very hard if not impossible. The adoption of computational power introduces new possibilities to present and process data in a new way.

The data mining of medical data also introduces new challenges due to the "uniqueness of medical data"[2]. This is reducible partly to the heterogeneity of medical data and to ethical, legal and social issues that have to be addressed.

Medical data is very heterogeneous because the disease of a patient is described with entities like images, doctor's notes or lab work. The interpretation of the physician is difficult to standardise and doctors need to agree on the same terms to describe the same condition. Test results can be imprecise and therefore deduced interpretation could contain errors.

Ethical, legal and social issues contain among other things that personal data has to be kept private and secure. It is necessary to establish guidelines how data is handled and in particular transferred and who has access rights. This differs between different forms of patient identification: anonymous data has the least sophisticated requirements while identified data can only be collected under significant review by federal guidelines and with the patient giving consent.

Especially Oncology is a very complex and multidisciplinary field including not only clinical but also biological and genetic data that poses demands on the sophisticated exchange of experience. The basic requirement for this is the acquisition, processing and analysis of data. The introduction of databases in medicine led to a more sufficient exploitation of data.

The databases in 3 and 4 will exemplify the variety of tumour databases.

2 Tumour Documentation

2.1 Tumour Documentation system

A precise documentation of tumours and their course of disease is an essential key tool for clinical oncology. A Tumour documentation system can be a tool for involved doctors to retrieve information of previous diagnostics and treatment of their patient. All information is coded that the data can be

\(^1\)Data can be examined to identify patterns and establish relationships.
analysed. The comparison of results of different diagnostic methods can improve the quality of cancer diagnoses as well as the comparison of treatment results and their complications can be a powerful quality assurance tool. The correlation of initial diagnostic findings with aetiopathology and treatment parameters supports the finding of prognostic factors and the evaluation of treatment courses. Thus the data can be used to plan novel clinical trials. Finally a tumour documentation system provides epidemiological data and data for scientific research. All this will lead to an improvement of cancer treatment.

A Tumour documentation can be divided into a basic and a organ-specific documentation.

The basic documentation contains information about the patient, the adjuvant therapy (radiotherapy, chemotherapy, medication) and the course of the disease. The documentation is the same for all tumour entities.

The organ-specific documentation will record specific tumour entity related extra information in detail about pre-therapeutic data, treatment, pathology, stage and classification. Datasets for different tumour entities comprises characteristic handling for each disease and thus varies in content.

National cancer organisations and governmental health programs compile usually an own dataset which they recommend clinicians to use. The meaning of each datafield is explained in a data dictionary as well as the form how its content needs to be recorded. This is done to assure that data can be compared and used for analysis. Examples for this are the National Cancer dataset[3], the German tumour documentation[4] (for an overview see 5), clinical cancer registry dataset of the cancer registry New South Wales Australia[5], the SEER Program Coding and Staging Manual, Data Standards and Data Dictionary[6] and the CoC’s Facility Oncology Registry Data Standards[7].

2.2 Cancer Registries

Cancer registries are institutions which collect data about cancer diseases from multiple sources over long periods of time.

Clinical or hospital cancer registries are conducted at hospitals or comprehensive cancer centers with the intent to improve cancer therapy. For this reason data of cancer patients are collected that contain information about diagnosis, treatment and the course of the disease. Quality and outcome of cancer therapy can be judged which enables the comparison of different therapies and therapists.

Population-based or epidemiological cancer registries collect and proceed data for the analysis of cancer incidence rates, survival and mortality in dependence of time and region. The data which is collected through doctors and pathologists can be used to perceive risks factors and possible causes for cancer and their prevention. They require a subset of the dataset used in the hospital cancer
registries.

National cancer registries ideally collect data from all cancer patients of the area they are responsible for. For example England conducts nine regional registries which collect information on cancer cases with the help of a standard dataset of the NHS[8]. The completeness is necessary for a meaningful, unbiased evaluation.

Many countries therefore introduce an act including the obligation to register every tumour case for epidemiological reasons. For the same reason efficiently working cancer registries do not ask their patient for consent. Informed consent has legal limitations and leads to loss of completeness. Sweden and Norway for example have collected data exhaustively without consent and backed by legislation.

Medical data is sensitive and so it is necessary for registries to propose confidentiality guidelines to protect the individual's rights. They include an explanation of security measures, security procedures and conditions for sharing data. Take as an example the “guidelines on confidentiality in population-based cancer registries in the European Union”[9] which are adopted by the European Network of Cancer Registries².

2.3 Clinical Registries Software

2.3.1 ODSeasy, Germany

ODSeasy[10] is a German-nationwide market-leading commercial software for the clinical documentation of gynaecological diseases (version 2009), especially breast cancer. It is installed in more than 300 hospitals in Germany. The manufacturer asthenis GmbH³ claims that it serves the purpose of using it in clinical routine and for quality assurance with a easy and clear structured GUI. It seems as if the latter is the main reason for its vast national distribution: German law requires a certain set of quality management data that can be extracted out of the software. Besides, there was an effort to improve breast cancer treatment in Germany during the late 90ties. To achieve this the German Association of Senology⁴ and the German Cancer Association⁵ formulated a set of requirements to ensure that women were treated by a team of experts and awarded hospitals meeting this standard the certificate “certified breast centre” (“zertifiziertes Brustzentrum”). The necessary component of sufficient tumour documentation is met by ODSeasy; it provides a comprehensive documentation from therapy decision (MDT) to follow-up care. Further functionalities provide support for compliance

²http://www.encr.com.fr/
³http://www.asthenis.de
⁴http://www.senologie.org/
⁵http://www.krebsgesellschaft.de/
of evidence-based guidelines and data exchange between hospitals and GPs.

- **Data:** Patient master data, anamnestic/diagnostic data, treatment data (systemic, surgical, radiation) with side-effects and complications. The data entry happens with the help of standardised dialogues. Uses always newest version of ICD-codes and ICPM-codes.

- **Technical:** The software needs a windows operating system and can be run on a server for several PCs. Adapted HL7 interfaces to various HIS, BAPI, ODBC. For statistical analysis there are interfaces to MS Office.

### 2.3.2 Giessener Tumordokumentationsystem GTDS, Germany

GTDS[11, 12, 13] was developed by the Department for Medical Information Technology of the University of Giessen for CCRs and follow-up. By November 2009 it is used in more than 40 cancer centres in Germany. Beyond the basis tumour documentation the system contains functions for clinical routine like medical reports, discharge letters and automatic generation of therapy protocols. There is support for follow-up management through guidelines and treatment support with the help of a tool to design and assign chemotherapy. With the help of an integrated Arden-Syntax trial eligibility of patients can be checked.

In Germany CCRs are usually run by Comprehensive Cancer Centres. Health care is divided between GPs, hospitals and ambulances. Hence information from different institutions have to be integrated. This fact is supported by GDTS and makes interdisciplinary patient care in oncology possible there are various interfaces implemented. This enables the import of administrative, laboratory or radiology data or pathology reports which already are available in the hospital. On the other hand, data can be exported to other systems like ECRs. This should reduce the cost of data gathering. In Germany more than 90% of the reports to the ECR (“Gemeinsames Krebsregister”) in the new states are transferred electronically from GTDS systems into CCRs. All data can be exported for statistical analysis.

From 1991-2000 the project was funded by the government and several cancer associations. Maintenance of the software is financed by service contracts with each end user.

- **Data:** [14] Patient's demographics and identification, information about the disease and condition at diagnosis/during the treatment/end of treatment/autopsy, planned/carried out treatment

---

6 Interface to ISH-med, which is a commercial HIS from SAP.
7 The Arden syntax is a language for encoding medical knowledge and can be used for decision making software
and documentation of pain, organisational data (e.g. appointments), socioeconomic status\textsuperscript{8}, QOL, palliative oncological treatment.

- **Technical**: Interfaces include exchange of HL7, BDT and ASCII messages. The architecture is a client/server application. It requires an initial installation of a run time environment on each PC. The access can also be done via web browser (JAVA language, Tomcat Servlet Engine and XML libraries). The system is based on the ORACLE database and development tools (Developer/2000 Forms and Reports).

### 2.3.3 The SEER Data Management System , US

The software\textsuperscript{15} is still being developed for and with SEER registries under the aegis of the NCI in order to have a common system which decreases system maintenance, standardise datasets and therefore improve data quality, data consistency and response time. The software provides all necessary registry functions like the import of data, editing, linkage, consolidation and reporting.

For developing the system a methodology was selected which was based on Flavin\textsuperscript{9} A more object oriented methodology like UML was deliberately not considered. The software will use a relational database model.

- **Data**: See 3.3.4

- **Technical (proposed)**: Operating system will be a common variant of Unix, the DBMS will most likely be Oracle, Sybase or PostgreSQL (open source). The system will include the following components: Apache Web Server, Application Server, client components will only run on Window PCs and be web-based or standard executables, disk storage subsystem, tape library and backup software, firewall.

\textsuperscript{8}Information on income, education and occupation.

\textsuperscript{9}Fundamental Concepts of Information Modelling.
3 Example Tumour Databases

3.1 Gene Databases

3.1.1 IARC TP53 Mutation Database

The database [16] gathers all TP53\textsuperscript{10} gene variations which are identified in humans and tumour samples. Scientist and clinicians can use the database as source for their research to understand molecular pathology of cancer, to find tumour-specific mutation patterns and to analyse genotype/phenotype relationships.

The database is developed and maintained by the IARC. The Data is manually compiled from peer-reviewed papers and other databases. Data can also be submitted directly via email in a prescribed way. There is a tool to check if the format is correct.

Part of the database can be searched through an open access web-portal for several criteria like mutation spectra, tumour spectra cell lines etc. Data can either be downloaded or shown online in table and graphs. The datasets cover somatic mutation\textsuperscript{11}, germ-line mutation\textsuperscript{12}, polymorphism\textsuperscript{13}, biological properties of p53 mutant proteins and TP53 gene mutation status in human cell-lines.

3.1.2 Human Tumour DataBase, France

The Hospital and Bioinformatics Unit of the Institute Curie established the database HuTuDB\textsuperscript{17} for basic research in oncogenesis and cancer treatment. The data is used to group tumours on a molecular level, find bio-markers and gene predictors for prognosis and diagnosis and to discover new therapeutical targets. Patient information is anonymously stored centrally for genome and transcriptome\textsuperscript{14} analysis. The database access is restricted. Authorised persons only can consult, query and update the database through a user-friendly interface.

The tumour entities which are examined cover breast carcinomas, uveal melanoma, paediatric tumours and sarcomas.

- Data: clinical patient information and transcriptomic data. The samples used have to be eligible for genome and transcriptome analysis. Own transcriptome data can be uploaded for further

\textsuperscript{10}TP53 is a growth-inhibiting gene which encodes the tumour suppressor protein p53. The protein has several mechanisms to prevent cancer in being able to activate DNA-repair, stop the cell cycle and initiate cell death if the DNA damage is irreparable.

\textsuperscript{11}Mutation in a body cell which was produced after conception and therefore is not passed on.

\textsuperscript{12}Mutation in germ cells which are transmitted to offspring.

\textsuperscript{13}Variation of a gene which occurs in a population.

\textsuperscript{14}The transcriptome is the set of all RNA molecules. The transcriptomes of cancer cells are used to understand the processes of carcinogenesis.
3.2 Single Projects - Treatment Outcome and Prognosis

3.2.1 Database Projects by the Grampian University NHS Hospitals Trust, Scotland

The college of Life Science and Medicine of the University of Aberdeen provides a database design service. They give rudimentary guidelines on database design[18] and data confidentiality[19]. In the following is a list of applications:

**Colorectal Initiative Tumour Database** The database[20] was developed to determine potential predictive and prognostic markers and novel therapeutic targets. For this reason it includes clinical details for colorectal tumours and adjacent uninvolved tissue. It also contains molecular analysis of samples.

- **Data:** Patient Identification, Sex, Date of Birth; date of death; Date and time of surgery; hospital, Surgeon; Biopsy number, tumour site, tumour size; location normal tissue with respect to tumour; diagnosis, duke’s stage\(^{15}\), date of diagnosis; operation type; node and TNM staging data; date of cancer recurrence; date of follow-up, treatment; metastasis sites, other tumours; chemotherapy details; transfusions prior to treatment; carcinoembryonic antigen levels at diagnosis and start of treatment; liver function tests.

- **Technical:** Data is held in Excel on a PC within the Institute of Medical science. For statistical analysis the data is exported into SPSS.

**The Scottish Urology Cancer Database** The database[21] analyses the data of patients that were treated for urological cancer. Recorded are cases who have been diagnosed a new primary urological cancer since January 2001 and after. Various hospitals (Aberdeen Royal Infirmary, Gartnavel General Hospital from Glasgow and Western General Hospital from Edinburgh) agreed on collecting data over a period of two years.

- **Data:** The database holds information on basic patient information, clinical information on organs affected by cancer and information about chemotherapy and radiotherapy. The data listed

\(^{15}\)Colorectal cancer can be staged according to the Dukes system. In the UK though now mainly the TNM classification is being used which is more precise.
below is managed in nine relational database tables which are related through a study identifier.

Patient registration: consultant/centre, patient postcode/date of birth/sex, hospital Identifier, organ.

Kidney: date/source/priority of referral, symptoms and duration, referral day, date first seen by hospital clinician, date of first consultation with surgeon, date and basis of definite diagnosis, site of tumour/area affected, number/size/type of tumours, Fuhrman Nuclear Grade\textsuperscript{16}, clinical staging, initial treatment intent, radical/partial nephromectomy, laparotomy inoperable, laparoscopy, embolism, other surgery, immunotherapy, radiotherapy, chemotherapy, oncology clinician/hospital, date of first definitive treatment, reason for delay, clinical trial information.

Renal pelvis & ureter: analog to dataset for kidney.


Prostate: analog to dataset for kidney, reference to specific treatment and tumour linked data like Gleason Score\textsuperscript{17}, TURP and PSA. Distinction between radical and palliative radiotherapy.

Testicular: analog to dataset for kidney, reference to specific treatment. Documentation of lymphatic/vascular invasion.

Penile: analog to dataset to kidney, reference to specific treatment.

Chemotherapy: consultant/hospital, oncology identifier, treatment intent, prescribed treatment with to and from dates, chemotherapy completed as prescribed, reason for non-completion, response, clinical trial information.

- Technical: The audit data is held on an Access\textsuperscript{18} relational database on personal computers. The data input is managed by using a form layout. Through selection buttons simple queries can be performed and reports produced. There are no external links to other systems. Single users can log onto the database via password.

**Lung Cancer Database** The Database\textsuperscript{[22]} was created in 2002 by the Medicine Assessment Research Unit to record clinical information on ca 1400 patients with lung cancer. The data were recorded from April 2000 to March 2002.

\textsuperscript{16}Most commonly used Kidney Cancer Grading. The grade affects the prognosis, but doesn’t currently affect treatment (Same treatment for same Stage regardless the Grade). Nuclear grade means that the system is based on just the appearance of the nuclei of the cancer cells, rather than the appearance or structure of the cells as a whole. Nuclear characteristics used in the Fuhrman Grade particularly indicate how actively the cells are making protein.

\textsuperscript{17}Staging system to help to evaluate the prognosis of prostate cancer. Together with other parameters, it is incorporated into a strategy of prostate cancer staging which predicts prognosis and helps guide therapy. A Gleason score is given to prostate cancer based upon its microscopic appearance. Cancers with a higher Gleason score are more aggressive and have a worse prognosis.

○ **Data:** The data is divided into tables related through the PAS number and are listed below.

- Patient history: identifier to other subsystems, name/date of birth/sex, date and cause of death, smoking/alcohol habits, treatment with chemotherapy/radiotherapy, respiratory history, history of malignancy, diagnosis of lung cancer, tumour-class/histological subtype, information about resection, TNM stage
- Substance history: type of substance abuse
- Occupation history: occupation
- Drug history: name of drug, route of administration/dose/frequency/duration
- Chemotherapy history: name of chemotherapeutic drug, start/end, dose/route\(^{19}\)/number of cycles
- Radiotherapy history: radiotherapy site/total dose, start/end
- Malignancy history: location of previous malignancy
- Family history of malignancy: relative/site of malignancy, age at onset/age at death
- Symptom history: symptom
- Investigation history: investigation like biopsy, bronchoscopy etc.
- Tissue diagnosis: biopsy/cytology/full specimen number, request date, results of analysis

○ **Technical:** The data is held in an Access database located on a confidential server on the university network. There is electronic data transfer from the lab system while clinical and epidemiological information from hospital notes are entered into the database via special forms.

**The Aberdeen Breast Cancer Database.** The database\(^{23}\) was started in 1999 and was retrospectively populated with data of patients with breast cancer taken out of the Aberdeen Royal Infirmary’s Patient Records System between 1991 and 1995. The aim is to collect information of a group of patients over a 10 year follow-up period and evaluate various treatment modalities. The database includes approximately 2500 patients which undergo a yearly follow-up.

○ **Data:** The database holds information about patient identification, referral details, epidemiology, family history, clinical findings, initial investigation and surgery, pathology, radiotherapy, hormone and chemotherapy. The data are managed in 53 relational database tables which are related through a unique patient identification number.

- Basic patient’s data: forename/surname, patient’s address/postcode/phone, sex, marital status, date of birth
- Current patient status: vital status/recurrence of disease, status date

\(^{19}\)Chemotherapy can be given in different ways
biochemistry results: date of result, alkaline phosphate/asparate aminotransferase/gamma glutamyl-transferase/calcium/haemoglobin level, white blood cells/platelet count
clinical findings from first contact: affected side/situation, vertical/horizontal tumour size, tumour/nodal/metastasis stage details, metastasis details
chemotherapy follow-up: date of follow-up, chemotherapy treatment, drug details
metastases information: date of record creation, assessment of metastasis stage, metastases details
epidemiology from first contact: age at menarche/menopausal stage/date of last period, family history of breast cancer and relations affected, family history of non-breast cancer and relations affected, type of hysterectomy/age at operation
surgery details from follow-up: date of follow-up information, breast/axillary surgery, treatment code for surgical procedure
Details of hormone therapy courses: date of treatment, treatment code for hormone therapy
Initial Chemotherapy hormone therapy: date of hormone treatment, adjuvant/therapeutic treatment, treatment code for hormone therapy, type of chemotherapy, date of initial chemotherapy/date of other chemotherapy, date contralateral tumour diagnosed, loco-regional control at death, breast cancer contribution to death
Details of initial contact pathology data: unique pathology identification number, histology details, axillary nodes and numbers of nodes involved/removed, maximum Tumour size and grade, oestrogen and progesterone receptor tumour status and value, human epidermal growth factor receptor-2
Details of initial radiotherapy treatment: code for radiotherapy treatment type, treatment start date, site of treatment and dosage, boost date/duration/dose
Details of investigation: date of investigation, result of chest and skeletal x-ray examination
Initial patient contact information: date of birth/age at referral, postcode at referral, date of referral, consultant/hospital at referral, radiotherapy department number and radiotherapy consultant, date of anniversary of first treatment, screen detected cancer, smoking status
Review data: date of review, tumour present or absent, primary site, nodes present or absent, skin/breast, lymphoedema, brachial plexopathy, metastasis, weight, treatment
Details of follow on scans x-rays: date of follow-up scans/x-rays, bone scan/mammogram/breast ultrasound/magnetic resonance imaging results/fine needle aspiration
Treatment summary: date of summary, drug name/prescribed dose, course/pulse dates
Surgery follow-up details: nature and date of follow-up for first breast procedure, axilla and date
of follow-up for second breast procedure, axilla details for second date

- **Technical**: The data are held on a PC in an access database. Input is managed by using a form and simple queries and reports can be made and produced through a selection list. There are no external links to any other systems and access is only possible for single users via a security password.

### 3.2.2 Non-functioning Pituitary Adenoma Database, Italy

Non functioning pituitary adenoma are rare and have a prolonged natural progression. Therefore the long-term outcome of this disease is not well established. To improve treatment of these tumours a database\[24\] was developed to evaluate clinical data and post surgery and radiotherapy outcome.

For this reason through the year of 2004 seven endocrinological centres in northern-western Italy collected retrospectively information on demographic, clinical and biochemical presentation, therapeutic approaches and long-term outcomes of 295 patients were they had been referred for diagnosis, cure and follow-up. The protocol was approved by local ethical committee of each hospital.

- **Data**: The database was divided into two parts: principal record including information about the patient at the time of recruitment and several follow-up records. Each patient received an identification number and personal information was available only to principle investigators or co-investigators.
  - The database contains information about risk factors, associated diseases, signs and symptoms and hormonal, clinical, biochemical, radiological, ophthalmological and outcome data.

- **Technical**: Access 2000 database. Data analysis is done with STATA8 software\[^{20}\].

### 3.2.3 TumorAGENT, Germany

TumorAGENT\[^{25, 26}\] manages the data of breast cancer patients and their tumour sample data to enable translational research\[^{21}\] and thus provide a tool for genome and proteome analysis of tumours. The aim is to understand the pathogenesis of breast cancer and help to develop a framework for cancer treatment. The interdisciplinary database was developed within four years by the department of gynaecology and obstetrics of the university hospital of Tübingen and TWT GmbH, a software

\[^{20}\] Data analysis and statistical software by StataCorp LP, http://www.stata.com/stata8/

\[^{21}\] Definition by the The Translational Research Working Group of the NCI: “Translational research transforms scientific discoveries arising from laboratory, clinical, or population studies into clinical applications to reduce cancer incidence, morbidity, and mortality.”
company, and is fully operational since July 2005. Implementations can be found in the University hospital of Tuebingen and at the Indiana University, US.

The database consists of medical patient information and tissue management. Before surgery an appropriate patient consent has to be obtained. For the data German data protection guidelines were applied: the medical professional code\textsuperscript{22}, the hospital act and the data protection act\textsuperscript{27} of the Federal state of Baden-Württemberg.

The database offers different role permits and interdisciplinary support like integration of workflow support for the laboratory.

- **Data**: The patient management consists of patient treatment data, including analysis about courses of disease, therapy, all relevant tumour related data to a patient and time-related information like therapy history, patient’s history and follow-up data. The blood and tissue management consists of information about sample storage for tissue and blood samples, history of handling and use of each sample. The availability of paraffin samples is documented. Altogether a documentation of approximately 800 attributes altogether, is possible with the help of 100 forms and 30 entities.

- **Technical**: Dynamic web application on IEEE\textsuperscript{23} Standard on “thin client\textsuperscript{24}” concept which is based on a Java web-application and is platform independent. The database is Oracle-based. The User-interface, business logic and database abstraction layer are defined and described in XML. SQL Queries based on ODBC and JDBC to other databases and software systems are possible. The users do not have to know any programming languages. Tumour Agent has an integrated role and authorisation model and a security model throughout the database to server and client browser which is realised by encryption and access restriction. This enables database management, data mining and documentation. Includes all medically relevant data protection measures and contains the output languages German and English.

### 3.2.4 The Lung Cancer Database, Japan

The National Cancer Center Hospital East in Japan established a large-scale lung cancer registry\textsuperscript{28} for basic and clinical research in 1999. Enrolled are 2506 patients that were newly diagnosed with primary lung cancer. The data taking was two-fold: The Baseline survey was done through two

\textsuperscript{22}Principles that guide the professional conduct of doctors in dealing with their patients, society, and with another.

\textsuperscript{23}The Institute of Electrical and Electronics Engineers is an international organisation for the progress of technology related to electricity and one of the leading standards-making organisations in the world.

\textsuperscript{24}Computer or software program which depends on its server to fulfil its job.
patient self-administered questionnaires concerning demographic data, health habits, psychological factors, follow-up for vital status and food habits. In the hospital medical and biological data were taken. Patients who did not completely filled out the questionnaires were reinvestigated by research assistants. A sampling bias and thus the limitation of the statistical analysis are due to the fact that patients of only one institution were used.

The protection of personal information was granted through personalisation. Clinical data was managed by researchers and research secretariat.

- **Data:**
  Demographic data: age at time of cancer diagnosis, sex, education level, marital status, smoking history, past history of cancer, family history of cancer, health habits, psychological factors (including the mental adjustment to cancer scale, the Eysenck Personality questionnaire\textsuperscript{25}-revised for Japan, the hospital anxiety and depression scale\textsuperscript{26} Fagerström Test for Nicotine Dependence\textsuperscript{27}, mental adjustment to Cancer (MAC) Scale).
  Food: food frequency questionnaire for dietary habits which contains 138 foods.
  Medical information from medical charts: histology, clinical stage, pathological stage, cancer treatment of first line, performance status, symptoms.
  Follow-up data: Assessment vital status, follow up survey (between 1999-2004)
  Biological data: urine specimens and blood samples (both before treatment), DNA from blood lymphocytes.

### 3.3 National Cancer Registries

#### 3.3.1 ECRIC (Eastern Cancer Registry and Information Centre), UK

ECRIC\textsuperscript{[29, 30]}, member of UKACR\textsuperscript{28}, is a cancer registry for all malignant tumours, benign brain and CNS tumours and some precancerous lesions covering about 5.5 million people in the east of England. The aim is to collect and keep data and make them available for the national statistics service, health professionals who want see how they perform, researchers and patients.

If incoming information show evidence for cancer (usually pathological information) the case is registered. About 40 tumour types are recorded, the cancers are staged at ECRIC by experts.

\textsuperscript{25}Questionnaire to assess the personality traits of a person.
\textsuperscript{26}The HADS consists of a 7-item anxiety subscale and a 7-item depression subscale to assess symptoms of anxiety and depression during the preceding week in medically ill patients.
\textsuperscript{27}An instrument for assessing tobacco dependence, which evaluates, among other factors, depth of inhalation, time from awakening to day’s first cigarette, smoking when bedridden with illness, and difficulty in smoking cessation.
\textsuperscript{28}\url{http://82.110.76.19/}
Data sources include data from PATH, PAS, medical records department, radiotherapy databases, MDT, CWT, Death Cards and other UK cancer registries which can be retrieved electronically as well as manual input from the registrars. The data feed is almost real-time (there are registries which update their data only annually). Because ECRIC does not impose a certain format they use different software to convert the data to a standard dataset.

ECRIC also holds the NHS National Brain Tumour Registry[31] for England. It holds approximately 1 million records from 1971 until now.

- **Data:** A minimum common dataset is required (for all English registries): Administration details, patient demographics (Birth date, ethnicity), diagnosis (date of diagnosis, who diagnosed), tumour details (site, behaviour, morphology and stage of tumour), treatment modality, death details, additional optional data. cancer registration dataset NHS[8]. One dataset comprises one cancer case a patient.

- **Technical:** To retrieve data for incidence, mortality and survival, a web-based portal, NCSI [32] was designed.

### 3.3.2 RTDS (Radiotherapy Database Project), UK

RTDS[33, 34] is a project by the NATCANSAT. Clinical and management data from seven radiotherapy facilities within the UK are collected routinely with the primary aim to commission and monitor services. Besides the data will be used by cancer registries, national audit and research programs.

Since April 2009 it is mandatory for electronically linked facilities to submit monthly data, from April 2011 this will be mandatory for all facilities.

Radiotherapy is a machine and computer intensive field and therefore many data are present in electronic form. These data can be extracted directly from the radiotherapy equipment software. Before uploaded the data needs to be quality assured.

- **Data:** [35] Patient pathway (timeline); patient identity, patient characteristics including age, gender and ethnicity; care episode including consultant, diagnosis scheme, primary diagnosis, secondary diagnosis; attendance occurrence including attendance record with outcome, service agreement details, clinical procedures and treatment site; referral including priority, requested service, data of referral request, and referrer; radiotherapy including treating site, treatment decision date, priority, treatment starting/ending date, treatment prescription with treatment region, number of fields, prescribed and actual dose, fractions, treatment modality, beam type, beam energy and exposure time.
- **Technical:** Monthly transfer of data as an AES256\textsuperscript{29} encrypted file, which has to have a NAT-CANSAT approved format. The following tool-kits for data transfer from the OIS are available: Lantis (Siemens), Impac mosaïq/MultiAccess (Elekta), Aria (Varian), record and verify system: Oncentra Visir (Nucletron).

### 3.3.3 CBTRUS (Central Brain Tumour Registry of the United States), USA

CBTRUS\textsuperscript{[36, 37]} is a national database which was started in 1992 as a not-for-profit corporation. It includes all primary malignant and non-malignant tumours of the brain, central nervous system, pituitary and pineal glands, and olfactory tumours of the nasal cavity. The database describes the incidence and survival rates and can evaluate diagnosis and treatment. The collected data is the largest aggregation of population-based data of this tumour entities in the US and provides the base for etiologic studies and tumour prevention.

The database has been developed by compiling data from state cancer registries that include information on both malignant and non–malignant primary brain tumours (11 collaborating sites).

- **Data:** Epidemiological data. Follow-up data is not available.

- **Technical:** CBTRUS can assist in accessing mortality, survival and treatment information of other databases.

### 3.3.4 SEER Cancer Database (Surveillance, Epidemiology, and End Results), USA

The SEER database is a program by the NCI which studies cancer epidemiology and outcomes \textsuperscript{[38, 39]}. For this reason it collects cancer survival and incidence information from 18 population-based cancer registries throughout the US (covering 26\% of the US). It was created in 1973 and represents a very important source for cancer statistics in the US (“gold standard”) holding more than 6 million cancer cases.

There are various applications for using the data: The database has been linked to Medicare\textsuperscript{30} billing data where much more detailed patient information is available. It is claimed this allows a better investigation of cancer outcomes.

SEER data is very useful on demographics and outcomes of rare malignant diseases. It is possible to see trends in diagnosis, stage migration\textsuperscript{31} and survival.

---

\textsuperscript{29}Advanced Encryption Standard is a symmetric-key encryption standard adopted by the U.S. government with a key size of 256 bit.

\textsuperscript{30}Health care insurance administered by the US government for elderly and/or disabled people.

\textsuperscript{31}Change of the distribution of a cancer stage induced by either a change in the staging system or an improvement in technology for detection of tumour spread.
Second malignancies, national and regional trends for treatment and outcome can be investigated. Finally, because there is a sufficient number of patients across different spectrums of presentation available, the data can be used to prove prognostic models of patient survival.

- **Data:** Registry, marital status, race/ethnicity, gender, age, birth date/place, date of diagnosis, primary site/grade/histology, whether first malignancy/sequence of reported malignancy, extent of disease at the time of diagnosis, extent of lymph node dissection and pathology, tumour markers and additional pathological and clinical information, staging, extent of surgery, radiation, date and cause of death. To ensure the quality and comparability of data, there is a SEER Program Coding and Staging Manual, Data Standards and Data Dictionary[6]. The data are updated annually.

- **Technical:** SEER Data Management System (2.3.3). Data can be accessed through the web, analysed and linked to other national data sources.

### 3.3.5 NCDB (National Cancer Database), USA

NCDB [40] is a nationwide oncology outcome database for all kinds of tumours. The aim is to explore trends in cancer care, create regional and state benchmarks for participating hospitals, and serve as the basis for quality improvement [38].

The NCDB is a joint effort of the CoC and the American Cancer Society. It collects a vast amount of data from more than 1,400 CoC-approved cancer programs.

For brain tumours it accounts for the largest database of its kind in the US: newly diagnosed cases are identified and follow-up is conducted on all primary brain tumours from hospitals accredited by the American College of Surgeons. Compared to SEERS3.3.4 or CBTRUS3.3.3 it contains more complete information regarding the treatment of these tumours [36].

- **Data:** patient characteristics, tumour staging and histology, type of first line, treatment, disease recurrence, survival, treatment outcomes. There are specialised datasets for the following tumour sites: breast, colorectal, gynaecological, head and neck, intracranial and central nervous system, liver, melanoma, pancreas, sarcoma, thoracic oncology, upper gastro-intestinal and urology. For some tumour sites radiotherapy data are under-reported due to outpatient treatment.

- **Technical:** Data is collected by approved cancer programs (cancer registries) and submitted to the NCDB using nationally standardised data item and coding definitions [7] as well as nationally
standardised data transmission format specifications. The data is submitted via file-upload after logging on the website http://www.facs.org/cancer.

Data confidentiality is very important and compliance is ensured with the Health Information privacy by the US government\(^{32}\). A Quality Integration Committee evaluates amongst other things the quality of the cancer registry data.

### 3.3.6 NODB (National Oncology Database), USA

The NODB is a commercial product by IMPAC which claims to be one of the largest longitudinal oncology databases worldwide\(^{33}\). It allocates data on cancer incidence, treatment and outcome. In 2009 the database held more than 2 million cases since 1985 from hundreds of oncology treatment facilities nationwide.

- **Data:** Demographics (age, sex, race, state, county, patient’s history of cancer, family history on cancer, vital status), tumour indicators (histology, positive lymph nodes, Staging, tumour size, differentiation, sites of distant metastases, prognostic test results) treatment modalities segmented by first and subsequent course (surgery/radiation type, radiation/surgery sequence, chemotherapy data, chemotherapy regimens, hormonal therapy/immunotherapy data), information on survival, recurrence and disease progression.

### 3.4 European Databases

#### 3.4.1 GENEPI Entb1&2 (ESTRO, European Universities & Institutions)

GENEPI Entb1\(^{42}\) & Entb2\(^{43}\) is a project by ESTRO, European universities and institutions with the aim to predict the effects of irradiation. With the help of a central database, which holds detailed information on the outcome of a large number of patients who received radiotherapy, radiation effects and genetic determinants of the variation in individual radio-sensitivity are being evaluated. The database is also linked to distributed European tissue banks containing samples of normal and tumour tissue. It is the largest infrastructure of its kind worldwide.

European guidelines for research were developed for the implementation of tissue and databanks. The GENEPI database policy \(^{44}\) leads to a frame of reference for researchers to exchange tissue with compliance to existing legislative and appropriate ethical standards. Participating centres are required to obtain consent from the local ethics committee which is written down in an informed

\(^{32}\)http://www.hhs.gov/ocr/privacy/

\(^{33}\)Data of patients are collected throughout their lifetime.
consent procedure\[45\]. To ensure data quality a detailed protocol for processing and storage of tissue samples\[46\] is given.

After the start of recruitment of patients in February 2004 the database was opened May 2004. In December 2005 6790 patients were recorded with 12017 tissue samples. Approximately 1600 prospective recruited patients, tissue, treatment parameters and outcome data were available in the central database. By September 2009 the number of patients included in the database was 10,000.

GENEPI Entb2 is the further development of the GENEPI Entb1 database: the database should be more user friendly and allow easy search at a distance. The functionality was extended to the storage of treatment plans with images and dose volume histograms. There was a bigger need for capacity to store genetic data. Also, the network of tissue banks was expanded to centres which do not have a laboratory or tissue bank of their own to contribute data and tissue from patients with a well documented overreaction to radiotherapy. This might make it possible to identify the genes that are responsible for over sensitive reaction to radiation. For this reason clinically radiation hypersensitive patients are compared with patients without abnormal reactions to radiotherapy from the GENEPI Entb1 database.

Interestingly the change from GENEPI Entb1 to GENEPI Entb2 lead mainly to less data fields because the total completion (done by documentation secretaries) was not feasible\[45\].

Finally, the main objective for the last 6 months of the project were to find the genetic pathways involved in patients who are overreacting to radiotherapy. This group then could be treated with lower doses and different therapy modalities to prevent severe side effects like late irreversible tissue damage. For this reason the database is used to identify the gene profiles of the patients that exhibit extreme radio-sensitivity without known hereditary hyper-radio-sensitivity syndromes. To ensure the evidence is linked to the genetic pathways, patients have to conform to certain eligibility criteria to be admitted to the database.

GENEPI Entb1 and Entb2 were funded under FP5 for originally 36 months each (Nov. 2002-Oct. 2005 and Sept. 2006 - Aug. 2009, was prolonged until Feb. 2010). Within the time period of funding the data input was supported by this funding. Afterwards it should run autonomously through a commercial model. External users of the data will have to pay and centres that feed in data or hold tissue banks will be reimbursed. The use of database is foreseen for at least 10-20 years.

- Data: Normal and Tumour tissue bank for patients with head and neck, breast, lung, rectal or prostate cancer with linked outcome database. The tissue bank contains lymphocytes, epithelial and Tumour tissue of cancer patients as well as lymphocytes of healthy individuals which will be stored for at least 20 years. The tissue bank from Entb2 contains skin fibroblasts, whole blood,
lymphocytes, plasma and lymphoblastoid cell lines from clinically radiation hypersensitive patients. The GENEPI database contains an essential dataset for each de-centrally stored tissue and describes which additional data are available in distributed databases. The already stored tissues from trials could be included which speeds up investigating late radiation effects. Most information on treatment parameters and outcome data is stored in local clinical databases but are also stored in the GENEPI database: clinical information about the patient, epidemiological information, diagnosis, treatment data (including radiotherapy), outcome and follow-up data. Medical issues are described in the most common standard and explicit described.

Random patient number, age at first radiotherapy, gender, ethnic background, cancer diagnosis, UICC stage, genetic known risks (there are some known genetic diseases which are linked to radio-sensitivity), past medical history, family history of radio-sensitivity, first and last day of RT, total dose planned and received, number of fractions planned and received, number of fractions per day and time interval between them, dose specification point (e.g. isocenter, at 5 cm depth), beam set-up, beam energy, number of fields, number of fields treated per fraction, systemic treatment and their dates, type or systemic treatment, description of acute side-effects of RT (CTCAE 3.0 grade, history, dates, treatment).

- **Technical**: The database is central. In the first version, data could be uploaded via a secure online data entry. Now data is submitted via email in text format. External data in csv-file format can be included in the database through a special software program. This makes bulk import of old data possible. Consistency checks before import and in the database (min-max values, no double listing of patients, check for complete datasets).

### 3.4.2 Conticabase and Conticagist

Conticabase[47] is an European sarcoma database and tumour bank which provides data for bone and mesenchymal tumours (except GIST). Conticagist[48] is the corresponding database for GIST. Both projects are under the auspices of Conticanet, the leading connective Tissue Cancer Network[49]. It is a Network of Excellence funded by the 6th Framework Program of the European commission (start: 2006). The aim of the projects is to produce a joint effort to understand these rare tumours.

Conticabase has more than 40 registered centres who provide almost 8000 patients for its database.

---

34 Comma-separated values file.
35 Loose connective tissue of the embryo
36 Gastrointestinal stromal tumour is a most common mesenchymal tumours of the gastrointestinal tract
Conticagist contains data from almost 900 patients.

Both databases contain information on the description of the tumour, treatment and follow-up, tumour sample availability, molecular biology analyses and a virtual slide system for pathological review of the tumour samples. To assure data quality a mandatory set of data is needed to enter a patient. A data dictionary explains each data field.

The databases are accessible only for members. Depending on their user profile they have the right to create new cases, add information and extract data from the database.

- **Data Conticabase**: Patient’s information like treating centre, age, sex; Information on the tumour like first consultation, site/size/depth of primary tumour, loco-regional extension, lymph nodes, metastasis, date and age at diagnosis, histotype, grading, morphological classification, differentiation, mitotic count, necrosis, vascular embol, date of first treatment, surgery of primary tumour, margin resection, tumour spillage, growth pattern, radiotherapy of primary tumour, dose, chemotherapy, complete remission after treatment, local recurrence, localisation of metastasis, patient status at latest follow-up, image of the tumour; information on the tumour sample like how the sample was performed/what type of tumour, treatment of tumour before sampling, sample frozen/paraffin; Information on the paraffin/frozen sample including number of samples, various tumour markers specific for sarcomas with the performed analysis technique, cytogenetic analysis, mutation analysis. Image of the tumour sample.

The dataset for Conticagist is slightly different taking certain attributes for GIST into account.

- **Technical**: Java software running on apache tomcat\(^{37}\) with a MySQL database Server. All tools used are open source.

The security model contains full https encryption with a certificate from an official authority, business level security model based on roles and centres and daily backups.

History of all user logins, journal of patients data modification (including creation and deletion). Identified partners with an additional certificate can exchange data through a secured web service layer.

### 3.4.3 EUROCARE (EUROpean CANcer REgistry-based study on survival and CARE of cancer patients)

EUROCARE examines the survival of European cancer patients with solid tumours and for this reason collects anonymised data on more than 13 million cancer cases from epidemiological can-

\(^{37}\)Open source servlet container which provides a web server environment for Java code to run.
cancer registries in 23 European countries\[50\]. On a large random sample of data it also evaluates the outcome treatment of cancer patients. The recent project version EUROCARE-5 covers survival data on patients diagnosed with cancer until 2007.

The data is collected in a database located at the Cancer Epidemiology Unit and Data Analysis Centre, Istituto Superiore di Sanità, Rome.

The confidentiality guidelines by the IARC are abided. All data remain property of the contributor which implies that he can oppose the usage of his data for any analysis. If data is used for another purpose than in the EUROCARE protocols foreseen the consent of the owner is required.

Parts of the EUROCARE-4 results can be accessed on-line through a web portal.

- Data: For each patient the following variables are compulsory: Patient identification code, sex, date of birth, date of diagnosis, date of tumour registration, date of death or last known vital status, vital status, primary tumour site, microscopic confirmation of diagnosis, morphology and behaviour\[^{38}\], extent of disease at diagnosis. Optional (among others): Stage at diagnosis, TNM classification, tumour size, number of metastatic lymph nodes, treatment, cause of death.

- Technical: A dedicated server is used for data storage and analysis. This server is not connected to the web. In the newest project version EUROCARE-5 data will be uploaded through a secure, web-based procedure. The data has to conform to a certain format.

### 4 Grid-related Cancer Infrastructures

#### 4.1 CaBIG (Cancer Biomedical Informatics Grid), US

CaBIG\[51\] is a nationwide network of tools and IT grid infrastructure in the US which was launched by the NCI to connect doctors, scientists and patients to share their knowledge, resources and data. The initiative started in 2004 with a three year pilot phase. By now the technology has been widely adopted, more than 2000 researchers from 700 organisations and most of the 65 NCI-designated cancer centres are working with CaBIG to collect and retrieve data.

The goal is to improve treatment outcome of cancer through prevention, better diagnosis and detection. For this data from heterogeneous sources need to be searchable and integrable by its users. This approach might be the foundation for personalised medicine.

The underlying infrastructure caGRID\[52, 53\], a grid enterprise architecture, is model-driven and service-oriented with some core tools. The Cancer Common Ontologic Representation Environment

\[^{38}\text{Benign, malignant or carcinoma in situ}\]
(CaCORE)[54, 55] gives the background to develop new tools and assure syntactic and semantic interoperability. Semantic interoperability means that certain controlled standard vocabularies are used that underlies the description of data that different systems can interpret and work with. Syntactic interoperability relies on common interfaces and protocols in a way that data services and analytic tools can be promoted and used by all clients.

The available tools in CaBIG provide support for clinical trials, pathology, the collection, analysis and management of basic research data, collection, annotation, sharing and storage of medical images and the management of biobanks. All software is open-source and open-access. Information and analytic resources are locally managed and can be securely accessed. The caGRID uses its own developed querying language CQL which is object oriented.

CaBIG also contains a security framework which fulfils privacy requirements and intellectual properties by means of policies and procedures while sharing data.

Below are two examples for databases where built within the CaGRID infrastructure.

### 4.1.1 EyecancerBig

EyecancerBig[56] is a project under way by the AJCC and the UICC to develop a database with associate tissue bank which contains medical and biological data about eye cancers. The goal is to create a common terminology and set of standards to pool and share treatment data from eye cancer specialists around the world. This will enable the evaluation of competing treatment regimes and communication between researchers and doctors. The database will be used as a base to develop an ophthalmologic oncology specific electronic medical record system in the future.

- **Data:** The dataset will be derived from the defined parameters for each tumour: epidemiologic, clinical diagnostic, photographic, ultrasound angiographic, ultrasonographic, radiographic, pathology, genetic, treatment, outcomes.

### 4.1.2 LEAD (Lymphoma Enterprise Architecture Database Platform)

LEAD[57] is a platform which integrates clinical and biomedical informatics research in order to facilitate the development of new clinical and therapeutic strategies for lymphoma. It was built using the CaCORE SDK. Different data with different standards can be integrated into one comprehensive database with the help of semantic tools.

It uses various sources to populate the database namely clinical data from emory university, SEER
registry data, cancer registry (IMPAC), laboratory, pharmacy and phase 1 lymphoma clinical trials data.

Different querying strategies were used to retrieve lymphoma datasets e.g. certain ICD-O histology and behaviour codes, free-text search of pathology reports or all medical documents. They had varying specificity for different sub-types of lymphoma.

The user can retrieve information from LEAD through a web browser.

- **Data**: Key elements feature registration data, chemotherapy, adverse events, histology, surgery, radiation therapy, lab data, physical exam, of treatment, baseline stage, concomitant procedure, response assessment, overall disease evaluation, baseline data, baseline history type concomitant medication, tumour specimen biopsy, haematology chemistry, toxicity, pharmacokinetics.

- **Technical**: Oracle 10 g relational database

### 4.2 BioGrid Australia

Bio Grid Australia is a non profit, national E-health initiative which provides a platform for clinical and scientific researchers to share and access clinical, laboratory and genetic data of diseases in existing databases across multiple sites nationwide. The databases holding patient data can be linked with research and genetic profiling data.

The data repositories are linked through the world wide web in an ethically approved and secured way. Only authorised researchers can extract and transform sub-sets of data and analyse them with their own tools.

Privacy and intellectual property is protected. Altogether more than 3 million dataset of patients are available.

- **Technical**: Service Oriented Architectures which uses distributed relational database technology (DB2). A virtual database is created spanning data sources at multiple institutions. Researchers can make queries using SQL. There is direct access for statistical analysis applications such as SAS. There is a local relational database at each institution. Views are used to the subset the data tables and fields. [59]

---

39Phase I trials are the first stage of human testing for the safety of a new drug in a small group of healthy persons.

40Side-effect that occurs in a person who participates in a clinical trial during the treatment or within a pre-specified period of time after the completion of treatment.

41All processes in the body the medication subjects.

42Business model which supports the communication between services from electronic data processing like software programs, databases and server.
The Australian Cancer Grid is the leading project of the BioGrid Australian platform. The aim is to enable data sharing on clinical and surgical outcomes of cancer treatment between researchers. The data is located at its hospital or research facility and can be combined for viewing for authorised persons.

4.2.1 ACCORD (Australian Comprehensive Cancer Outcomes and Research Database)

ACCORD is a data collection software for hospitals that can be used as cancer registry and outcomes database[60]. It was developed by BioGrid, the Victorian Partnership for Advanced Computing43 and launched in 2004 by Cancer Trials Australia44 in partnership with the Ludwig Institute of Cancer Research45. Data is collected of all cancers patients treated throughout Australia on different sites including the Royal Melbourne Hospital, Western Hospital, Box Hill Hospital, Peter MacCallum Cancer Center, Austin Hospital and the Royal College of Surgeons in Adelaide.

The data is taken within the hospital. To identify new or returning patients ACCORD is interacting with the HIS. The database can be queried by clinical researchers and clinicians and is free of charge for BioGrid members.

There are comprehensive datasets with data dictionaries for several tumour streams[61]:

- **Tumour stream Brain**: 2007/08 finalisation of CNS dataset and database. Future plans: increase functionality, eg. automatic electronic letters including treatment plans and outpatient discharge summaries. Linkage to other in-hospital databases, including pathology will allow research into bio markers predicting response to treatment.

- **Tumour stream Breast**: Has clinical functional features. Multidisciplinary breast minimum dataset. Future Plans: import of pre-existing databases at Royal Woman's, Royal Melbourne and Western Hospitals will be imported.

- **Tumour stream Colorectal**: Most mature database along the streams. Under the control of the colorectal surgical society of Australia46, data is being collected at interstate sites and internationally (Sao Paolo, Brazil). Electronic chemotherapy prescribing. Future plans: Links to other cancer databases in BioGrid including linkage with Radiation Oncology Victoria47, linkage to hospital pharmacy, pathology databases and to radiology databases.

---

43VPAC is a leading, independent and not for profit advanced computing R&D service provider which was established in 2000 by a consortium of Victorian Universities [http://www.vpac.org/].

44Not-for-profit service organisation that conducts cancer clinical trials [http://www.cancertrialsaustralia.com/].

45Global non profit institute which supports research to understand and control cancer ([http://www.llicr.org/]).

46[http://www.cssa.org.au/]

473rd largest service provider in Australia
○ **Tumour stream Head and Neck (incl. Thyroid):** The management of head and neck cancers involves specialised multidisciplinary medical care and input from sub-specialised health members. Start data collection at Royal Medical Hospital in 2007. The challenges is to include coordinating clinical services and all necessary disciplines for optimal treatment. The ongoing issue is to capture chemotherapy and radiotherapy data and there is a need for multi-centre collaboration to record all aspects of the patient’s treatment regimen. Future plans: Coordination of MDT and the patient’s journey including appointments, admission and discharge process.

○ **Tumour stream Lung:** Tumour type which requires most frequently multidisciplinary care and therefore need for thorough data collection. 2007-2008 the Victorian Cooperative Oncology Group\(^{48}\) lung group developed a minimum dataset, data dictionary is in progress which includes newly ratified lung staging system. dataset involves surgeons, radiation oncologists, respiratory physicians, medical oncologists and palliative care physicians. Coordination of data from all disciplines at the point of contact and linking of relevant external databases relevant to lung cancer, especially radiotherapy. Future: Analyse the impact of the new staging system fields of survival.

○ **Tumour stream Melanoma:** 2007-2008 development of a minimum dataset over three sites. Set consists of combined data for research and clinical activities for comprehensive clinical information to complement the Melanoma Tissue Bank (in collaboration with the Victorian Cancer Biobank). Future: Data Managers will be employed at each site to ensure high quality data for the project. Several other sites will be included.

○ **Tumour stream Sarcoma:** 2007-2008 definition of each Tumour subtype, definition and pilot of a minimum dataset at sites where the majority of sarcoma patients are treated. Building of an electronical database to share data between sites, expansion to other hospitals. Future: Development of an integrated Australian wide sarcoma database with comprehensive clinical data matched to corresponding tumour specimens in collaboration with the Australian Sarcoma Study Group\(^{49}\).

○ **Tumour stream Upper Gastrointestinal:** The treatment of these tumours requires a multi-disciplinary collaboration between medical and surgical disciplines (radiation oncologists, gastroenterologists, medical oncologists, surgeons). 2007-2008: oesophago-gastric dataset is tested to collect patient data and then a full comprehensive dataset will follow. Future: Collaboration with

\(^{49}\)http://www.australianearcomagroup.org/
Western & Central Melbourne Integrated Cancer Service\(^{50}\) to establish an optimal follow-up program.

- **Tumour stream Uro-oncology (Renal and Prostate):** New therapies for urological cancers involve urologists, radiation oncologists, medical oncologists and palliative care physicians. The data might be difficult to capture. 2007-2008: A renal cell carcinoma dataset has been developed and tested at the Austin and Royal Melbourne Hospital. The data will be complemented by a tissue collection. A prostate dataset has been developed and tested at the Royal Melbourne Hospital and Epworth and will be complemented by a tissue collection. Future: population-based prostate cancer registry. The renal database will be developed as a web based module with the functionality to document multi-disciplinary meetings, treatment plans and pull data from existing hospital systems.

- **Technical:** Originally with ASP.NET and C# on a .NET v1.1 framework, web server: IIS 5.0 and SQL Server 7.0 For speeding up the responses: Anthem .NET AJAX library technology (for front end).
  
  There is a user manual\(^{62}\) for the database and it can be tested online\(^{63}\).

### 4.2.2 CART-WHEEL (Center for Analysis of Rare Tumours)

Cart-Wheel\(^{64}\) is a project which started in 2009 to pool data for the study of rare cancers. Patients with a rare tumour or their proxies fill in the data through a website. It will be used for a database to enable and support ethically approved research projects. Patients all over the world can supply data with a consent to different levels of data and/or tissue access.

- **Data:** The database records information about personal data like name, age, sex, born in, mostly lived, background, culture and contact details. Details about diagnosis cover tumour type, date of diagnosis, primary and secondary tumour sites. Treatment details include surgery date, what surgery was performed, who performed the surgery, the kind of the treatment, date of start and end, radiotherapy details and sideeffects to any treatment. There are some questions about general health, smoking habits and alcohol consumption. If there were other tumours the kind and date of diagnosis is recorded, furthermore the family history of cancer, family gene tests and tumour gene tests.

\(^{50}\)http://www.wcmics.org/
5 Conclusion

Depending on the aim there are many different kinds of databases. They might be designed for molecular research, therapy decision/evaluation, support of clinical routine and quality assurance, integrated patient's care, for clinical trials or epidemiological analysis. Their function can vary from being a pure information repository to being a tool which helps to extract additional insights. The examples of tumour databases in this report only review the latter. For the sake of completeness it should be mentioned here that there are also databases that contain pure bibliographic repositories like MEDLINE[65] or information on oncology topics for cancer patients like NCI's Comprehensive Cancer Database PDQ [66] which can be accessed through the internet.

An epidemiological cancer registry requires just scarce clinical information on data which describe the tumour and the progress of the therapy. This serves the purpose of characterising the behaviour of cancer in a population with its survival sufficiently. To evaluate therapies in an clinical evidence based manner a more detailed record of clinical data is necessary. On the other hand, Tissue banks are a necessary tool to improve the molecular science of oncogenetics and tumour progression[67, 68]. However, to describe a tumour disease of an individual in detail, sufficient data on both, biomedical and clinical data is needed.

But not only the datasets vary depending on their objective but also the technical solutions. The range comprises from a password secured, PC-held database with a couple of relational database tables to a multi-source, database-federated infrastructure with complex service and security architecture.

The simple solutions are often used by physicians in a local clinical environment to do "conventional" evaluation on treatment outcome of their treated patients. The capability of data mining is very limited.

Formerly, tumour data used to be stored in individual databases with proprietary data formats. This made integrational research on different resources impossible. The big volume of data which is created in the cancer community obviously is exploited most effectively for progress through close interaction. Such a interdisciplinary medical infrastructure will have a variety of users with different professional backgrounds and needs not only to provide role based access but must also make it feasible to search, collect, manage and analyse data from multiple sources. That kind of data sharing requires semantic and syntactic interoperability[52].

[51]The Medical Literature Analysis and Retrieval System Online is a bibliographic database including information for articles from academic journals covering life science and biomedicine.
[52]ability to communicate, exchange and interpretate data.
In current projects there has been an effort to realise an infrastructure for accessing and seamless integrate heterogeneous medical databases with the help of GRID services. Besides providing resources for high performance computing the GRID enables a secure and privacy preserving data access.
Appendix

Summary of a Basic Tumour Documentation

In the following as an example there is a draft of the German basic tumour documentation[69].

Demographic Data

They are important for the identification of the patient and his tumour and therefore enable the communication of involved parties. These data should be collected at the first contact in the hospital:

- Name, first name, birth name, place of birth, marital status, next family member
- No. for aftercare, tumour registration data co-treating hospitals (name, address, department), co-treating medical practitioners (name, address)
- For population-based tumour registries: nationality, municipality and job-“anamnesis”

Pre-treatment Data

all information should include the date, respectively.

- Date of admission and reason for consultation
- Anamnesis, pre-neoplastic conditions and lesions
- Date of the first tumour diagnosis
- Other primary tumours
- Performance status\(^{53}\)
- Serious secondary disorders and surgical risks\(^{54}\)
- Diagnostics (imaging, tumour marker)
- Location of primary tumour (ICD-O)
- Side location (only for pairs of organs)
- Localisation of distant metastasis
- TNM-classification and clinical stage

\(^{53}\) For adults mainly Karnofsky scoring or ECOG score is used.
\(^{54}\) Usually done with ASA physical status classification system.
Data about Therapy and Progress

all information should include the date

○ Intended and conducted therapies: radiotherapy, surgical treatment, drug therapy

○ Clinical R-classification

○ Non-surgical therapy: overall assessment of the tumour (see WHO)

○ secondary surgical interventions

○ early therapy complications

Pathological Data

○ histological typing (ICD-O-3)

○ histopathological grading (see WHO, UICC)

○ pTNM-classification

○ analysis of isolated tumour cells

○ Classification of invasion of local lymphatic vessels and veins

○ tumour biological and molecular biological

○ classification of resection boundaries

○ methods of classification of resection

○ definitive grading

○ definitive R-classification (Tumour status after therapy, gives result of therapy)

Quality of Life

To comprise the QOL the questionnaire suggested by the EORTC should be used. It collects the most important aspects of the health-related QOL (global health, physical, role, emotional, cognitive and social functioning), most frequent tumour-related symptoms (fatigue, anorexia, nausea, emesis, pain, gastrointestinal symptoms, insomnia, breathlessness) and financial strain in a reliable way.

○ EORTC-QLQ C30[70]
**Follow-Up**

Follow-up data should be taken after each contact with the patient (eg. aftercare, incidence of progression or relapse, conclusion of a multimodal therapy, drop-out not death related). Nevertheless, the examination should be done at least once a year.

**Conclusion**

If the patient dies or leaves aftercare, a final outcome data should be taken.
References


http://www.med.uni-giessen.de/akkk/info/15/abstracts/vandenbergh.html.


http://www.med.uni-giessen.de/akkk/gtds/.

http://www.med.uni-giessen.de/akkk/gtds/grafisch/doku/bd5f.htm.


[18] College of Life Science and Medicine of the University. Database design guidelines. (accessed 05/2010). Available from:

[19] College of Life Science and Medicine of the University. Confidentiality guidelines. (accessed 05/2010). Available from:

http://www.abdn.ac.uk/clsm/dbase/pdf/CIT.pdf.


[34] NATCANSAT. Homepage of NATCANSAT about RTDS. (accessed 05/2010). Available from: http://www.canceruk.net/rtservices/rtds/.


[38] James B Yu. NCI SEER public-use data: Applications and limitations in oncology research. Oncology, 23(3), March 2009.


[66] NCI. PDQ comprehensive cancer database. (accessed 05/2010). Available from: 


[69] DKG. Basis Tumordokumentation. (accessed 05/2010). Available from: 
http://eliph.klinikum.uni-heidelberg.de/OTD/teil/teil-b.

[70] EORTC. QLQ-C30. (accessed 05/2010). Available from: 
Nomenclature

AJCC  The American Joint Committee on Cancer
ASA  American Society of Anesthesiologists
BDT  Behandlungsdatentransfer (treatment data transfer)
CCR  Clinical cancer registry
CoC  Commission on Cancer
CTCAE  Common Terminology Criteria for Adverse Events
CWT  Cancer Waiting Times
DBMS  Database Management System
ECR  epidemiological cancer registry
GP  A general practitioner: doctor who provides primary care. He is generally not specialized and treats acute and chronic illnesses.
GUI  Graphic User Interface
HIS  Hospital Information System
HL7  Health Level 7
IARC  International Agency for Research on Cancer
ICD-O  International Classification of Diseases for Oncology
ICPM  International Classification of Procedures in Medicine
MDT  Multi-disciplinary Team System: A team which is made up of a number of health care professionals (i.e. Doctors, Nurses, Radiologists and Histopathologists) who have experience in a particular cancer and decide about the proceeding of a patient treatment.
NAACCR  North American Association of Central Cancer Registries
NATCANSAT National Cancer Services Analysis Team: in-house provider of medical informatic services to the NHS ranging from application and website development to geographical data analysis.

NCSI National Cancer Information Services

PAS Patient Administration System

PATH Pathology

PSA Prostate-Specific Antigen

QOL Quality of Life

RDBMS Relational Database Management System

RT Radiotherapy

SPSS Statistical Package for Social Scientists

TURP Transurethral resection of the prostate

UICC International Union Against Cancer

UKACR Kingdom Association of Cancer Registries

UML Unified Modeling Language

WHO World Health Organization

ICD International Statistical Classification of Diseases and Related Health Problems
Note: WP24, D.2: Inventory of Existing Databases in Collaboration with Centres of the PARTNER Consortium

By now, there are no databases in Hadron therapy centres within the PARTNER consortium that could be reported.

The only facility in operation treating patients is the *Heidelberger Ionenstrahl-Therapie Zentrum (HIT)*\(^55\).

\(^{55}\) [http://www.hit-centrum.de](http://www.hit-centrum.de)