ICT-2011-288048

EURECA

Enabling information re-Use by linking clinical Research and CAre

IP
Contract Nr: 288048

Deliverable: D9.2 Canonical Models of EHRs and Clinical Trial Systems

Due date of deliverable: (02-01-2013)
Actual submission date: (04-03-2013)

Start date of Project: 01 February 2012 Duration: 42 months

Responsible WP: Philips

Revision: <outline, draft, proposed, **accepted**>

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
	Dissemination level	
PU	Public	X
PP	Restricted to other programme participants (including the Commission Service	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (excluding the Commission Services)	

0 DOCUMENT INFO

0.1 Author

Author	Company	E-mail
Monique Hendriks	Philips Research	monique.hendriks@philips.com
Andre Dekker	MAASTRO	andre.dekker@maastro.nl
M. Scott Marshall	MAASTRO	m.scott.marshall@maastro.nl
David Pérez Rey	UPM	dperez@infomed.dia.fi.upm.es
Keyur Mehta	GBG	Keyur.Mehta@germanbreastgroup.de
Raúl Alonso	UPM	ralonso@infomed.dia.fi.upm.es
Juan Manuel Moratilla	UPM	jmmoratilla@infomed.dia.fi.upm.es
Gema Molina	UPM	gmolina@infomed.dia.fi.upm.es
Cyril Krykwinski	IJB	cyril.krykwinski@bordet.be
Sheng Yu	UOXF	sheng.yu@oncology.ox.ac.uk
Norbert Graf	UdS	Norbert.Graf@uniklinikum-saarland.de

0.2 Documents history

Document version #	Date	Change
V0.1		Starting version, template
V0.2		Definition of ToC
V0.3		First complete draft
V0.4		Integrated version (send to WP members)
V0.5		Updated version (send PCP)

0.3 Document data

Keywords	
Editor Address data	Name: Monique Hendriks Partner: Philips Research Address: High Tech Campus 34, 5656AE Eindhoven, The Netherlands Phone: +31611230156 Fax: E-mail: monique.hendriks@philips.com
Delivery date	04-03-2013

0.4 Distribution list

Date	Issue	E-mailer

Table of Contents

1 Introduction

The purpose of this document is to provide a common information model (CIM) for the EURECA services, which enables at least the implementation of the scenarios described in WP1, but is designed in such a way that it is easily extendable; this common information model will contain generic concepts and relations which can encompass a broad set of specific concepts and relations. As such, the CIM's concepts and relations can be instantiated for a diverse set of scenarios in the clinical domain.

Since EURECA services are aimed at the entire domain of clinical research as well as care, there is a necessity to provide a common information model. A variety of standard ontologies and models are already in use and several partners use their own data and information models. The information model provided here will encompass a data model consisting of concepts which can be obtained from systems which are already in use by defining a mapping. For the clinical partners in EURECA, these mappings will be provided in deliverable D4.3. Therefore, the services become applicable to any clinical site by developing a mapping from the clinical site's data models to the EURECA common data model.

1.1 Structure of the deliverable

Section 2 describes in detail the available information systems and the data models used by those systems at the clinical sites. Section 3 provides a detailed description of the scenarios proposed by the clinical partners and shows which data is required to execute those scenarios and where that data can be found in the systems at the clinical sites. Section 4.1 presents some previously used common information models in the clinical domain. Section 4.2 presents the common information model that will be used in the EURECA project. Section 4.3 describes the core data set, the terminology that will be used in the data model. Finally, Section 5 presents some conclusions.

2 Clinical Sites and Available Systems

This section describes the internal data models of the systems used by the clinical partners in the project. The purpose of these descriptions is to understand which data is available in existing systems and how it can be retrieved for mapping and storage in the EURECA common data and information model. Furthermore we aim to reveal in this section, which subset of the data described in D9.1 is relevant and how it is used.

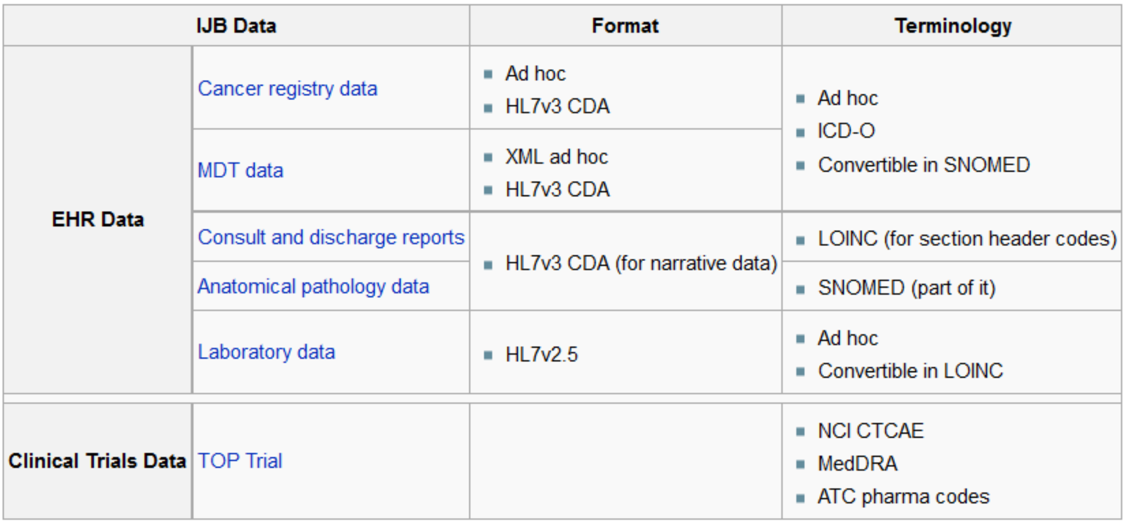
2.1 Institut Jules Bordet (IJB)

Institut Jules Bordet is involved in trials and treatment of breast cancer. Scenarios of IJB are focused on trial feasibility, selection and recruitment, linking care data to the cancer registry, treatment side effect retrospective study, pre-filling of CRF, safety reporting and long-term follow-up.

At IJB, patient care data is stored in an internally developed EHR system called Oribase. The language of Oribase is French.
Patient care data is also stored in:
- Cancer registry
- Multidisciplinary Team (MDT)
- Consultation reports
- Laboratory data
- Anatomopathology
- Chemotherapy prescriptions
- Day Hospital consultation reports
- Radiology and nuclear medicine
- PACS

IJB uses the clinical trial management system Oracle Clinical to manage clinical trials.

Table 1 lists the data sources used at IJB.



IJB Data		Format	Terminology
EHR Data	Cancer registry data	Ad hoc HL7v3 CDA	Ad hoc ICD-O Convertible in SNOMED
	MDT data	XML ad hoc HL7v3 CDA	
	Consult and discharge reports	HL7v3 CDA (for narrative data)	LOINC (for section header codes)
	Anatomical pathology data		SNOMED (part of it)
	Laboratory data	HL7v2.5	Ad hoc Convertible in LOINC
Clinical Trials Data	TOP Trial		NCI CTCAE MedDRA ATC pharma codes

Table 1: Patient care and clinical data provided by IJB

For detailed descriptions of the data schemas used in each of these systems, see EURECA Deliverable D9.1 [8].

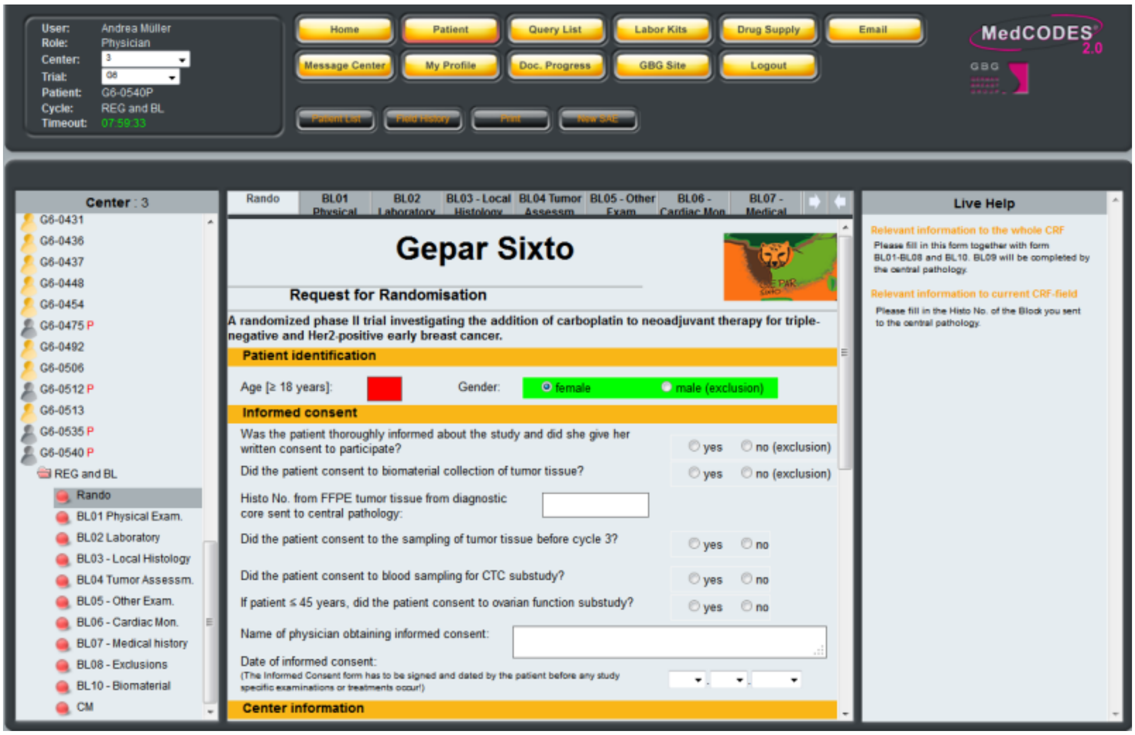
2.2 German Breast Group (GBG)

The German Breast Group is an Academic Research Organisation. GBG is involved in trials of breast cancer and will be validating the scenarios regarding trial selection and recruitment.

GBG has a group of heads of hospitals (oncologists) that meet regularly and discuss new possible clinical oncology trials. The heads of hospitals are the advertisers of the trials and find patients in their respective hospitals. Other clinics who are interested contact us and receive information of trials through our GBG Annual Meeting.

At GBG, for trial management, an in-house system called MedCODES is used. Figures Figure 1-Figure 3 show some screenshots of this system.

The data formats used in MedCODES are ad hoc, no standards are used.



User: Andrea Müller
Role: Physician
Center: 3
Trial: G6
Patient: G6-0540P
Cycle: REG and BL
Timeout: 07:59:33

Center : 3

G6-0431
G6-0436
G6-0437
G6-0448
G6-0454
G6-0475 P
G6-0492
G6-0506
G6-0512 P
G6-0513
G6-0535 P
G6-0540 P
REG and BL
Rando
BL01 Physical Exam.
BL02 Laboratory
BL03 - Local Histology
BL04 Tumor Assessm.
BL05 - Other Exam.
BL06 - Cardiac Mon.
BL07 - Medical history
BL08 - Exclusions
BL10 - Biomaterial
CM

Gepar Sixto

Request for Randomisation

A randomized phase II trial investigating the addition of carboplatin to neoadjuvant therapy for triple-negative and Her2-positive early breast cancer.

Patient identification

Age [≥ 18 years]: 🟥 Gender: ● female ○ male (exclusion)

Informed consent

Was the patient thoroughly informed about the study and did she give her written consent to participate? ○ yes ○ no (exclusion)

Did the patient consent to biomaterial collection of tumor tissue? ○ yes ○ no (exclusion)

Histo No. from FFPE tumor tissue from diagnostic core sent to central pathology: ____

Did the patient consent to the sampling of tumor tissue before cycle 3? ○ yes ○ no

Did the patient consent to blood sampling for CTC substudy? ○ yes ○ no

If patient ≤ 45 years, did the patient consent to ovarian function substudy? ○ yes ○ no

Name of physician obtaining informed consent: ____

Date of informed consent:
(The Informed Consent form has to be signed and dated by the patient before any study specific examinations or treatments occur!)

Center information

Live Help

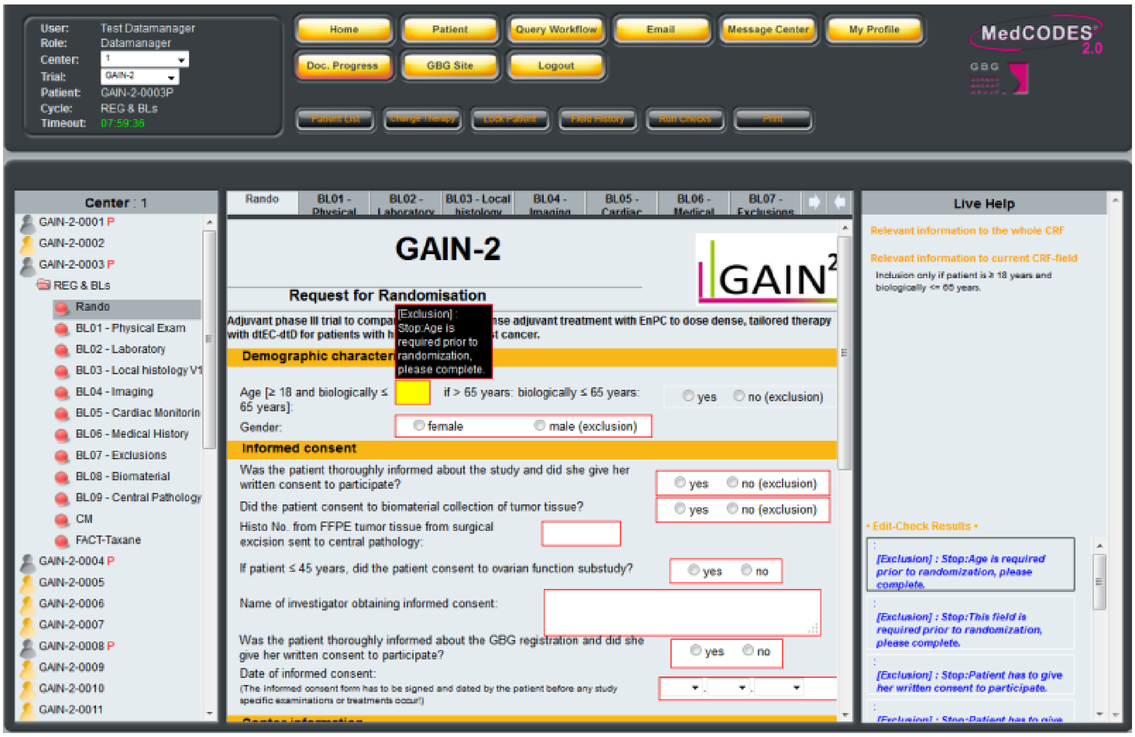
Relevant information to the whole CRF

Please fill in this form together with form BL01-BL08 and BL10. BL09 will be completed by the central pathology.

Relevant information to current CRF-field

Please fill in the Histo No. of the Block you sent to the central pathology.

Figure 1: MedCODES screenshot of data entry for a new patient to be included in randomization of a trial.



User: Test Datamanager
Role: Datamanager
Center: 1
Trial: GAIN-2
Patient: GAIN-2-0003P
Cycle: REG & BLs
Timeout: 07:59:36

Center : 1

GAIN-2-0001 P
GAIN-2-0002
GAIN-2-0003 P
REG & BLs
Rando
BL01 - Physical Exam
BL02 - Laboratory
BL03 - Local histology V1
BL04 - Imaging
BL05 - Cardiac Monitorin
BL06 - Medical History
BL07 - Exclusions
BL08 - Biomaterial
BL09 - Central Pathology
CM
FACT-Taxane
GAIN-2-0004 P
GAIN-2-0005
GAIN-2-0006
GAIN-2-0007
GAIN-2-0008 P
GAIN-2-0009
GAIN-2-0010
GAIN-2-0011

GAIN-2

Request for Randomisation

Adjuvant phase III trial to compa nse adjuvant treatment with EnPC to dose dense, tailored therapy with dtEC-dtD for patients with h t cancer.

[Exclusion] :
Stop:Age is required prior to randomization, please complete.

Demographic character

Age [≥ 18 and biologically ≤ 65 years]: 🟨 if > 65 years: biologically ≤ 65 years: ○ yes ○ no (exclusion)

Gender: ○ female ○ male (exclusion)

Informed consent

Was the patient thoroughly informed about the study and did she give her written consent to participate? ○ yes ○ no (exclusion)

Did the patient consent to biomaterial collection of tumor tissue? ○ yes ○ no (exclusion)

Histo No. from FFPE tumor tissue from surgical excision sent to central pathology: ____

If patient ≤ 45 years, did the patient consent to ovarian function substudy? ○ yes ○ no

Name of investigator obtaining informed consent: ____

Was the patient thoroughly informed about the GBG registration and did she give her written consent to participate? ○ yes ○ no

Date of informed consent:
(The informed consent form has to be signed and dated by the patient before any study specific examinations or treatments occur!)

Live Help

Relevant information to the whole CRF

Relevant information to current CRF-field

Inclusion only if patient is ≥ 18 years and biologically <= 65 years.

• Edit-Check Results •

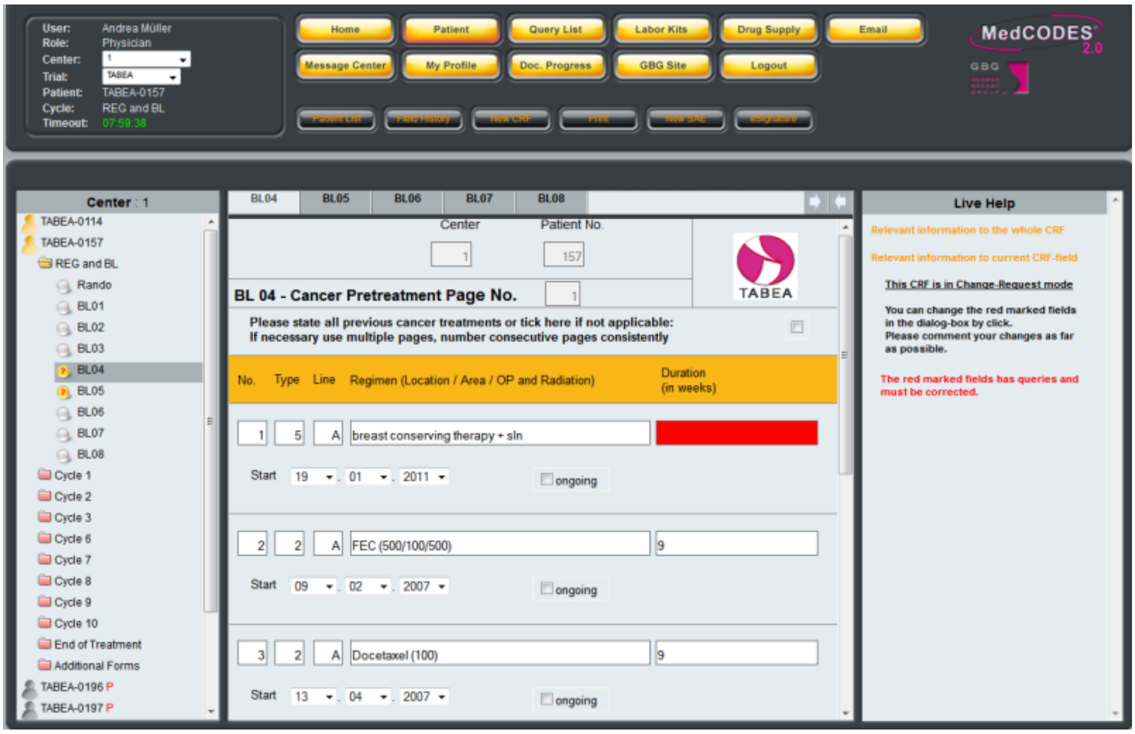
[Exclusion] : Stop:Age is required prior to randomization, please complete.

[Exclusion] : Stop:This field is required prior to randomization, please complete.

[Exclusion] : Stop:Patient has to give her written consent to participate.

[Exclusion] : Stop:Patient has to give

Figure 2: MedCODES screenshot of data management of inclusion/exclusion criteria.



User: Andrea Müller
Role: Physician
Center: 1
Trial: TABEA
Patient: TABEA-0157
Cycle: REG and BL
Timeout: 07:59:38

Center 1
Patient No. 157

BL 04 - Cancer Pretreatment Page No. 1

Please state all previous cancer treatments or tick here if not applicable: □
If necessary use multiple pages, number consecutive pages consistently

No.	Type	Line	Regimen (Location / Area / OP and Radiation)	Duration (in weeks)
1	5	A	breast conserving therapy + sln	🟥
Start 19 . 01 . 2011			□ ongoing	
2	2	A	FEC (500/100/500)	9
Start 09 . 02 . 2007			□ ongoing	
3	2	A	Docetaxel (100)	9
Start 13 . 04 . 2007			□ ongoing	

Relevant information to the whole CRF

Relevant information to current CRF-field

This CRF is in Change-Request mode

You can change the red marked fields in the dialog-box by click.
Please comment your changes as far as possible.

The red marked fields has queries and must be corrected.

Figure 3: MedCODES screenshot of a CRF.

2.3 MAASTRO

MAASTRO clinic is involved in radiation oncology (trials and treatment) in different diseases. Scenarios of MAASTRO are focussed on decision support; helping treating physicians to find the right literature, possible trials, prediction of outcome of treatment and adverse events related to treatment options. There is also a scenario on data reuse. Figure 4 visualizes the care process at MAASTRO and how MAASTRO’s data sources are involved in this care process.

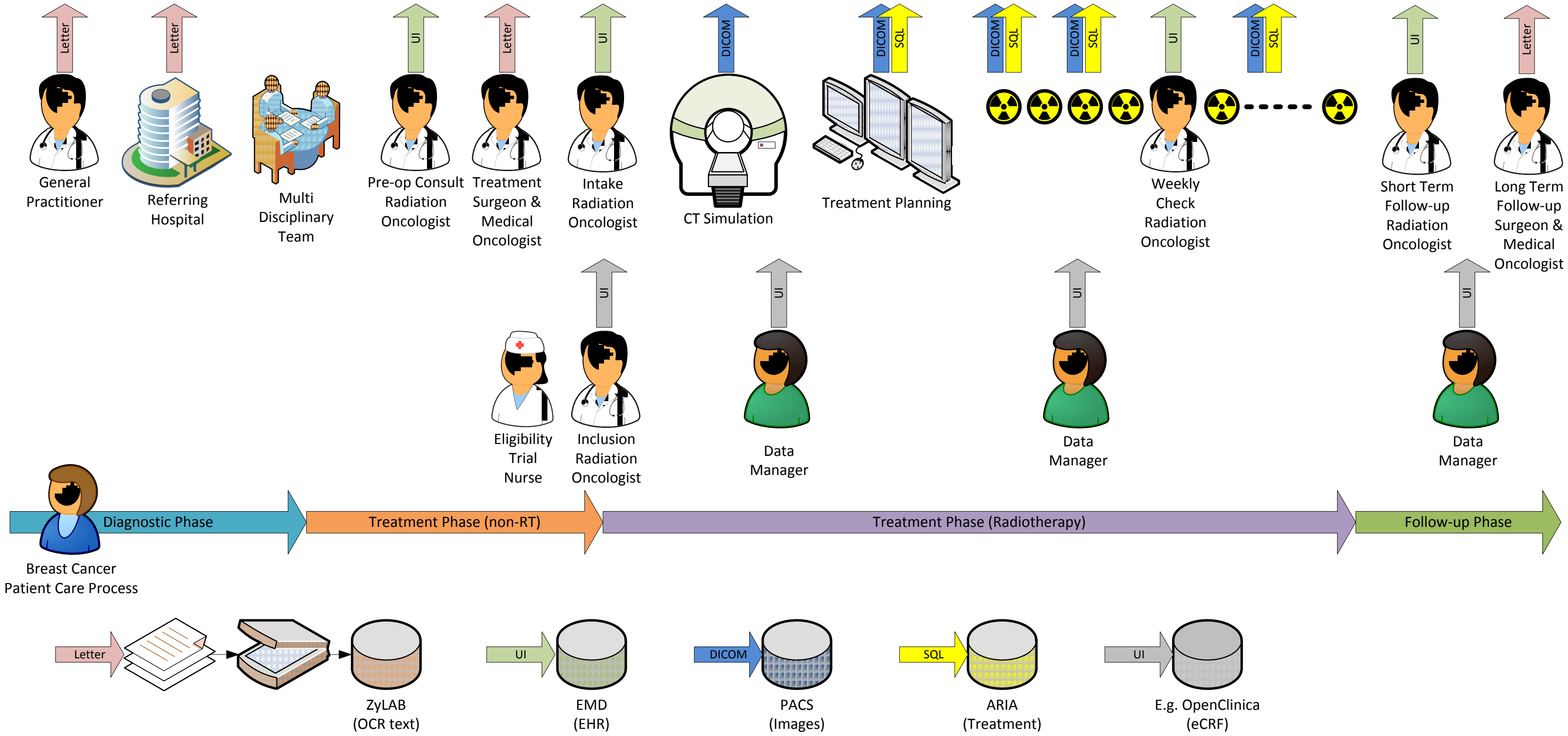


Figure 4: MAASTRO care process and data sources. Top: Radiotherapy care process including data generation (arrows). Top-Middle: Additional data generation if patient is included in a clinical trial. Bottom-Middle: Chronological treatment phases (not to scale). Bottom: Clinical data sources to be used in EURECA.

CARE PROCESS

For a "typical" breast cancer patient treated with a breast conserving care plan including radiation therapy at MAASTRO the care process is as follows:

Diagnostic phase

After examination, the patient is sent to the referring hospital by the general practitioner for a series of diagnostic tests which usually include mammography, echography, PET-CT, MR and ultimately a biopsy which confirms the diagnosis. At a multidisciplinary board meeting (which often includes the radiation oncologist), a decision is made to treat the patient with, in this example, breast conserving therapy. Depending on stage and/or the need to reduce tumor load, neo-adjuvant chemotherapy may be indicated. The referring hospital and sometimes general practitioner send information on this patient in the form of letters (printed on paper) to MAASTRO. These letters are scanned, OCR(Optical Character Recognition)-ed and put into the ZyLAB database (see below).

Treatment phase (non-RT)

In this phase the patient is operated to remove the lesion. Before the operation, the radiation oncologist sees the patient and fills in information on the patient and her cancer in the EMD database (see below). The surgical report, sentinel node procedure report and the pathology report of the tumor are sent to MAASTRO in the form of letters which are scanned and imported into ZyLAB.

Treatment phase (Radiotherapy)

After recovering from surgery, the patient is invited to MAASTRO for the intake. The following information is present at intake:

- Report sentinel node scintigraphy (required)
- Operation Report on with, if available, letter from surgeon (lumpectomy / sentinel node / amputation)
- PA-report
  - Biopsy breast (before surgery)
  - Biopsy or cytology armpit (if done) (before surgery)
  - Lumpectomy or mastectomy mammary
  - Axillary sentinel node (if done) and / or axillary node dissection
- Report oncology discussion (MDB)
- Medical imaging:
  - Report mammograms last 2 years (if necessary the images of specimen radiography or CT, postoperative mammogram, ultrasound)
  - MRI from last year

At intake, the radiation oncologist enters information into the EMD. A simulation CT scan, tumour and normal tissue delineation and treatment plan are subsequently made. The CT scan and treatment plan are stored as DICOM objects in the PACS (see below) and in the ARIA database (see below).

Once the treatment plan is approved, the patient is usually treated with a fractionation schedule that might vary depending on the requirement of a radiation boost to the primary tumour (e.g. the shortest schedule is 16 daily (on workdays) fractions of 2,67 Gray per fraction). During the radiation treatment the patient is seen weekly by a radiation oncologist, during these "on treatment" visits, information on especially acute toxicities are noted in the EMD. All fractions are captured in the ARIA database. Imaging done during therapy is archived in the PACS.

Follow-up

After treatment the patient is usually referred back to the referring hospital. Adjuvant chemotherapy may be indicated in this phase depending on the reduced risk of disease free and overall survival. MAASTRO has agreed with the referring hospitals that only one physician does the follow-up. Depending on the chemotherapy component and the wishes of the patient, the patient is followed-up once a year for five years either by a medical oncologist or a surgeon or a radiation oncologist. The follow-up physician sends letters to the other members of the treatment team (which is put into the MAASTRO ZyLAB database).

TRIALS

The trial eligibility is usually considered at intake. A trial nurse checks which patients have an intake appointment the next day and compares the trial criteria with the information present at that time in ZyLAB and EMD. If he estimates that the patient might be eligible, he supplies the radiation oncologist with the required information to estimate if the patient is eligible and also to perform an informed consent procedure. If the patient is included in a trial, the data management department will supply the radiation oncologist, during various patient interactions, with the required information and case report forms and/or approach the patient directly for information (e.g. questionnaires). Depending on the trial, the data managers enter this information onto paper or into an electronic clinical trial system. For MAASTRO initiated trials the OpenClinica system (see below) is preferred.

DATA SOURCES

- EMD: A homemade electronic health record system. Figure 5 shows the structure of the MAASTRO EMD
- ZyLAB (www.zylab.com): A document management system that contains scanned documents which or OCR-ed and all text put into an index. When scanned each document is annotated with basic metadata (e.g. which patient the document belongs to)
- ARIA (http://www.varian.com/us/oncology/radiation_oncology/aria/): An oncology information system that holds information on administered radiotherapy (often called a Record & Verify system)
- OpenClinica (www.openclinica.com): An open source clinical trial management system.
- PACS (www.clearcanvas.ca): An open source image archive.

The data in the EMD and the data linked to the EMD via PACS, ARIA or ZyLAB consists of structured patient, tumour and treatment features, the trial the patient has participated in (if any) and free text answers to the questions Medical History, Oncological History and Medication (often in short-hand notation).

The high level structure of the EMD is shown in **Error! Reference source not found.**. Intake questions are questions like Medical and Oncological History, Medication, etc. Treatments are linked to ARIA. Referral letters can be linked to the patient via ZyLAB.

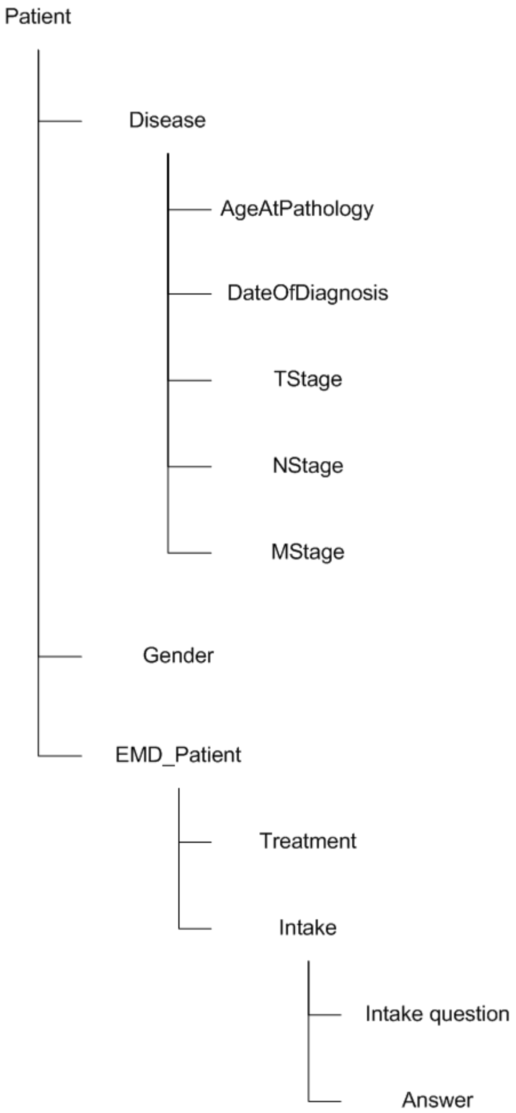


Figure 5: Structure of MAASTRO EMD

MAASTRO Data			Format	Terminology
EHR Data	EMD	Electronic Medische Dossier	• Ad hoc XML (via euroCAT) • HL7v3 CDA	• Ad hoc • ICD-9 • CTCAE • NCI Thesaurus (via euroCAT)
	Aria	Oncology Information System	• DICOM • Ad-hoc	• None
	PacsOne/ClearCanvas	PACS	• DICOM	• None
	Zylab	Clinical reports - OCR scans	• TXT	• None

Table 2: Patient care and clinical data provided by MAASTRO

2.4 University of Oxford (UOXF)

The University of Oxford (UOXF) is involved in trials and treatment of breast cancer and sarcoma (soft tissue). Scenarios of UOXF are focussed on automated diagnosis, increasing clinical trial enrolment and collecting care data relevant for research automatically.

There are a large number of clinical systems and databases running in the Oxford University Hospitals trust. Some databases are part of the operational system in the hospitals to support daily healthcare activities while others support clinical research. An overview has been given in the deliverable D9.1 [8] to cover the primary data sources and systems that could potentially provide clinical data for the Eureca project. Actual accessibility, and related conditions, are being assessed and is mentioned below.

2.4.1 The Electronic Patient Record system

Cerner Millennium EPR is the primary electronic patient record system of the main hospitals in Oxford. Its main responsibilities include patient administration, scheduling and identity management etc.

As a proprietary clinical system the internal information model of the Cerner Millennium is not completely open. However it has the capability to expose HL7 interfaces and send HL7 messages. In the current setting, an HL7 integration engine called MirthConnect is used to connect with the Cerner Millennium EPR system to forward information to different operational clinical systems.

2.4.2 PACS (imaging)

The NHS Picture Archiving and Communications Systems (PACS) store images that are generated in many clinical areas. The PACS system in Oxford supports standard DICOM based communication. However it is still under investigation whether an interface will be available for data querying. The NHS has published a set of specifications to standardise the implementation of PACS system in the UK. The internal model for communication should be compliant with the DICOM standard.

2.4.3 Pathology systems and databases

The NHS Oxford University Hospitals trust has a number of pathology databases that are operating across different sites. The pathology reports that are stored in these databases are most likely to contain free text and a limited amount of structured data. It is under investigation and negotiation to obtain clinical data from the operational system. The schema of each database is slightly different from each other. Most data models of these databases were constructed by internal NHS staff.

2.4.4 Aria (oncology information system)

The department of oncology in Oxford uses Aria to manage cancer patients. It handles the patient chemotherapy scheduling and following-ups. The system stores cancer related information such as images and clinical notes. The live system has the capability to deploy HL7 interfaces to connect with other systems. Currently the internal data model and schema is not present. However standard HL7 templates could be defined to interact with the interfaces. It is also under negotiation that whether the system could provide a live link or regular dumps of data.

2.4.5 BioBanking software

The Oxford Musculoskeletal Biobank (OMB) provides the facility and lab inventory management system for researchers to retrieve and manage samples and tissues. Currently the OMB uses Sapphire, a biobanking solution from Labvantage to manage information that associated with the samples. The internal data model is not known, however users could define their own information templates that could be stored in Sapphire, which means the data model for the sample information is known. It is under close investigation to obtain access to data and data exporting functionality.

2.4.6 Retrospective clinical studies databases

There are also retrospective clinical studies databases in Oxford that have been built over years to fit specific clinical purposes. One example is the Sarcoma research database, which consists of data that were manually gathered from different information systems in the trust. A summary of the schema of the database is shown in Figure 6.

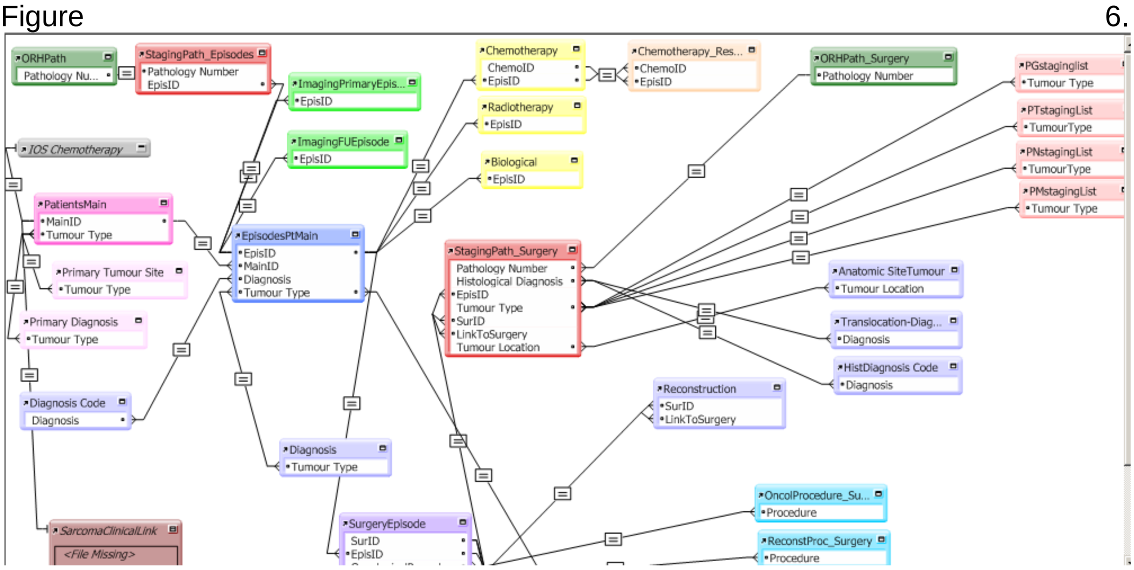


Figure 6: Sarcoma research database schema

The database has been exported from the FileMaker pro database application therefore it is fully available as a data resource.

2.4.7 Cancer Outcomes and Services Dataset (COSD)

The Cancer Outcomes and Services Dataset is a data standard that is being mandated through the UK to be used as a structured data specification for cancer reporting. It has a comprehensive information model that covers a large area in cancer healthcare. Therefore it could be seen as a harmonised model for different cancer related data sources. All NHS trusts need to be compliant with the standard when submitting clinical data about cancer treatment. Therefore all IT departments in the trusts are undergoing processes to adopt and implement the data standard.

In Oxford this work has gone under way through various tracks. It is envisioned that the use of the COSD standard is a more reliable way of gathering cancer related information. For this reason, we are investigating the effort of implementing the COSD in OUH trust and the possibility of using a developmental service that provides the COSD compliant data as a primary data source instead of connecting and pulling data from each individual clinical system.

The COSD standard contains a large data model[1]. What is relevant in our research and the Eureca project are a subset of the model. Figure 8 shows a SarcomaCorePathology type from the model, which is covering the related information for sarcoma cancer. As a result, all clinical data will be collected with respect to this specification and the final data output will be available a primary data source.

The cancer outcome service dataset (COSD) will be the primary standardised data specification for all cancer related clinical data. This dataset standard specifies the structure for reporting health information about cancer patients. The clinical data that are compliant with the standard will be collected by the NHS information department of the vast majority of hospitals in the UK and submitted to the National Cancer Data Repository.

The specification covers a wide range of cancer related information. The data that is defined in the standard are covering the following cancer categories: *Breast, CNS (site specific Central Nervous System), Colorectal, CTYA (Children, Teenager and Young Adult), Gynaecological, Haematology, Head and Neck, Lung, Sarcoma, Skin , Upper GI,* and *Urology*.

The general structure of a COSD compliant report will contain the following elements shown in Table 3.

[1] More details of the COSD dataset: http://www.ncin.org.uk/collecting_and_using_data/data_collection/cosd.aspx

Core (generic structure)	LinkagePatientID	Patient identity details
	LinkageDiagnosis	To carry diagnostic details for linkage
	Demographics	Where this information is exchanged, the appropriate data item name should be used to identify the particular instance of the data
	Referrals	To carry patient referral details to the Trust that receives the first referral
	Imaging	To carry imaging details
	Diagnosis	To carry diagnostic details
	Cancer care plan	To carry cancer care plan details
	Clinical trials	To carry clinical trial details for a patient who is eligible for a cancer clinical trial
	Staging	To carry the cancer staging details
	Treatment	To carry the cancer treatment details
	Surgery and other procedures	Surgery details and other procedures including interventional radiology, laser treatment, endoscopies etc
	Radiotherapy	Radiotherapy details
	Active monitoring	To carry active monitoring details
	Pathology details	Contains all data from the Royal College of Pathologists dataset
	Cancer recurrence	To carry cancer recurrence/secondary details
	Death details	Details about the death

Table 3: Structure of a COSD compliant cancer report

Each cancer patient of a particular cancer type will have the generic structure with additional cancer specific information. Please see the full standard specification for all data elements[2].

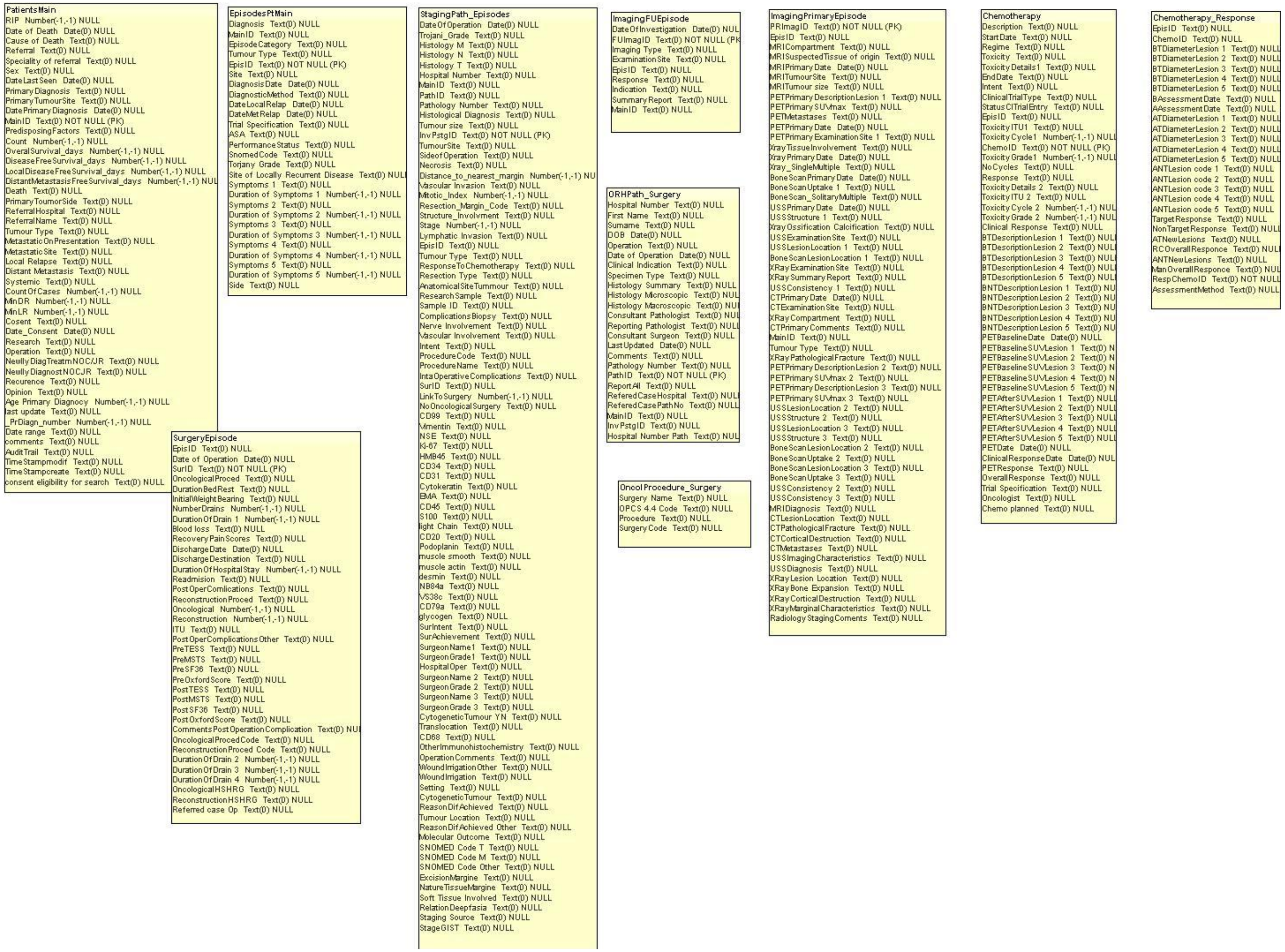
The implementation of the COSD standard in the Oxford university hospitals trust is currently in the development phase. The information service department is responsible for building a service that collects and submits clinical data in a COSD standard compliant fashion. The data is collected from a variety of clinical systems: PACS, pathology databases, Cerner Millennium EPR, MDT database etc. Once the service is ready for testing, a copy of the COSD dataset can be obtained from the MS SQL server in the information service department.

2.4.8 SarcDB – a retrospective clinical study database

The sarcoma research database (SarcDB) is a manually curated database with data entered by a NHS data manager. The data manager logs on to numerous clinical systems such as EPR, PACS, pathology report database, to look for relevant information of the new sarcoma patients that were discussed at the MDT meeting. The content of SarcDB covers a wide range of information including the patient outcome data, chemotherapy and pathology. The coverage of this existing dataset is similar to what the COSD dataset aims to cover. In the future a possible transformation may be carried out to make the dataset COSD compliant.

The main data model of the research database is shown in Figure 7.

[2] The complete specification is available on http://www.ncin.org.uk/collecting_and_using_data/data_collection/cancer_outcomes_and_services_dataset_cosd_latest_downloads.aspx



PatientsMain
RIP Number(-1,-1) NULL
Date of Death Date(0) NULL
Cause of Death Text(0) NULL
Referral Text(0) NULL
Speciality of referral Text(0) NULL
Sex Text(0) NULL
DateLastSeen Date(0) NULL
PrimaryDiagnosis Text(0) NULL
PrimaryTumourSite Text(0) NULL
DatePrimaryDiagnosis Date(0) NULL
MainID Text(0) NOT NULL (PK)
PredisposingFactors Text(0) NULL
Count Number(-1,-1) NULL
OveralSurvival_days Number(-1,-1) NULL
DiseaseFreeSurvival_days Number(-1,-1) NULL
LocalDiseaseFreeSurvival_days Number(-1,-1) NULL
DistantMetastasisFreeSurvival_days Number(-1,-1) NUL
Death Text(0) NULL
PrimaryToumorSide Text(0) NULL
ReferralHospital Text(0) NULL
ReferralName Text(0) NULL
Tumour Type Text(0) NULL
MetastaticOnPresentation Text(0) NULL
MetastaticSite Text(0) NULL
Local Relapse Text(0) NULL
Distant Metastasis Text(0) NULL
Systemic Text(0) NULL
CountOfCases Number(-1,-1) NULL
MinDR Number(-1,-1) NULL
MinLR Number(-1,-1) NULL
Cosent Text(0) NULL
Date_Consent Date(0) NULL
Research Text(0) NULL
Operation Text(0) NULL
NewllyDiagTreatmNOC/JR Text(0) NULL
NewllyDiagnostNOCJR Text(0) NULL
Recurence Text(0) NULL
Opinion Text(0) NULL
Age Primary Diagnocy Number(-1,-1) NULL
last update Text(0) NULL
_PrDiagn_number Number(-1,-1) NULL
Date range Text(0) NULL
comments Text(0) NULL
AuditTrail Text(0) NULL
TimeStampmodif Text(0) NULL
TimeStampcreate Text(0) NULL
consent eligibility for search Text(0) NULL

EpisodesPtMain
Diagnosis Text(0) NULL
MainID Text(0) NULL
EpisodeCategory Text(0) NULL
Tumour Type Text(0) NULL
EpisID Text(0) NOT NULL (PK)
Site Text(0) NULL
DiagnosisDate Date(0) NULL
DiagnosticMethod Text(0) NULL
DateLocalRelap Date(0) NULL
DateMetRelap Date(0) NULL
Trial Specification Text(0) NULL
ASA Text(0) NULL
PerformanceStatus Text(0) NULL
SnomedCode Text(0) NULL
Torjany Grade Text(0) NULL
Site of Locally Recurrent Disease Text(0) NULL
Symptoms 1 Text(0) NULL
Duration of Symptoms 1 Number(-1,-1) NULL
Symptoms 2 Text(0) NULL
Duration of Symptoms 2 Number(-1,-1) NULL
Symptoms 3 Text(0) NULL
Duration of Symptoms 3 Number(-1,-1) NULL
Symptoms 4 Text(0) NULL
Duration of Symptoms 4 Number(-1,-1) NULL
Symptoms 5 Text(0) NULL
Duration of Symptoms 5 Number(-1,-1) NULL
Side Text(0) NULL

SurgeryEpisode
EpisID Text(0) NULL
Date of Operation Date(0) NULL
SurID Text(0) NOT NULL (PK)
OncologicalProced Text(0) NULL
DurationBedRest Text(0) NULL
InitialWeightBearing Text(0) NULL
NumberDrains Number(-1,-1) NULL
DurationOfDrain 1 Number(-1,-1) NULL
Blood loss Text(0) NULL
RecoveryPainScores Text(0) NULL
DischargeDate Date(0) NULL
DischargeDestination Text(0) NULL
DurationOfHospitalStay Number(-1,-1) NULL
Readmision Text(0) NULL
PostOperComlications Text(0) NULL
ReconstructionProced Text(0) NULL
Oncological Number(-1,-1) NULL
Reconstruction Number(-1,-1) NULL
ITU Text(0) NULL
PostOperComplicationsOther Text(0) NULL
PreTESS Text(0) NULL
PreMSTS Text(0) NULL
PreSF36 Text(0) NULL
PreOxfordScore Text(0) NULL
PostTESS Text(0) NULL
PostMSTS Text(0) NULL
PostSF36 Text(0) NULL
PostOxfordScore Text(0) NULL
CommentsPostOperationComplication Text(0) NUI
OncologicalProcedCode Text(0) NULL
ReconstructionProced Code Text(0) NULL
DurationOfDrain 2 Number(-1,-1) NULL
DurationOfDrain 3 Number(-1,-1) NULL
DurationOfDrain 4 Number(-1,-1) NULL
OncologicalHSHRG Text(0) NULL
ReconstructionHSHRG Text(0) NULL
Referred case Op Text(0) NULL

StagingPath_Episodes
DateOfOperation Date(0) NULL
Trojani_Grade Text(0) NULL
Histology M Text(0) NULL
Histology N Text(0) NULL
Histology T Text(0) NULL
Hospital Number Text(0) NULL
MainID Text(0) NULL
PathID Text(0) NULL
Pathology Number Text(0) NULL
Histological Diagnosis Text(0) NULL
Tumour size Text(0) NULL
InvPstgID Text(0) NOT NULL (PK)
TumourSite Text(0) NULL
SideofOperation Text(0) NULL
Necrosis Text(0) NULL
Distance_to_nearest_margin Number(-1,-1) NU
Vascular Invasion Text(0) NULL
Mitotic_Index Number(-1,-1) NULL
Resection_Margin_Code Text(0) NULL
Structure_Involvment Text(0) NULL
Stage Number(-1,-1) NULL
Lymphatic Invasion Text(0) NULL
EpisID Text(0) NULL
Tumour Type Text(0) NULL
ResponseToChemotherapy Text(0) NULL
Resection Type Text(0) NULL
AnatomicalSiteTummour Text(0) NULL
ResearchSample Text(0) NULL
Sample ID Text(0) NULL
ComplicationsBiopsy Text(0) NULL
Nerve Involvement Text(0) NULL
Vascular Involvement Text(0) NULL
Intent Text(0) NULL
ProcedureCode Text(0) NULL
ProcedureName Text(0) NULL
IntaOperativeComplications Text(0) NULL
SurID Text(0) NULL
LinkToSurgery Number(-1,-1) NULL
NoOncologicalSurgery Text(0) NULL
CD99 Text(0) NULL
Vimentin Text(0) NULL
NSE Text(0) NULL
Ki-67 Text(0) NULL
HMB45 Text(0) NULL
CD34 Text(0) NULL
CD31 Text(0) NULL
Cytokeratin Text(0) NULL
EMA Text(0) NULL
CD45 Text(0) NULL
S100 Text(0) NULL
light Chain Text(0) NULL
CD20 Text(0) NULL
Podoplanin Text(0) NULL
muscle smooth Text(0) NULL
muscle actin Text(0) NULL
desmin Text(0) NULL
NB84a Text(0) NULL
VS38c Text(0) NULL
CD79a Text(0) NULL
glycogen Text(0) NULL
SurIntent Text(0) NULL
SurAchievement Text(0) NULL
SurgeonName1 Text(0) NULL
SurgeonGrade1 Text(0) NULL
HospitalOper Text(0) NULL
SurgeonName 2 Text(0) NULL
SurgeonGrade 2 Text(0) NULL
SurgeonName 3 Text(0) NULL
SurgeonGrade 3 Text(0) NULL
CytogeneticTumour YN Text(0) NULL
Translocation Text(0) NULL
CD68 Text(0) NULL
OtherImmunohistochemistry Text(0) NULL
OperationComments Text(0) NULL
WoundIrrigationOther Text(0) NULL
WoundIrrigation Text(0) NULL
Setting Text(0) NULL
CytogeneticTumour Text(0) NULL
ReasonDifAchieved Text(0) NULL
Tumour Location Text(0) NULL
ReasonDifAchieved Other Text(0) NULL
Molecular Outcome Text(0) NULL
SNOMED Code T Text(0) NULL
SNOMED Code M Text(0) NULL
SNOMED Code Other Text(0) NULL
ExcisionMargine Text(0) NULL
NatureTissueMargine Text(0) NULL
Soft Tissue Involved Text(0) NULL
RelationDeepfasia Text(0) NULL
Staging Source Text(0) NULL
StageGIST Text(0) NULL

ImagingFUEpisode
DateOfInvestigation Date(0) NUL
FUImagID Text(0) NOT NULL (PK
Imaging Type Text(0) NULL
ExaminationSite Text(0) NULL
EpisID Text(0) NULL
Response Text(0) NULL
Indication Text(0) NULL
SummaryReport Text(0) NULL
MainID Text(0) NULL

ORHPath_Surgery
Hospital Number Text(0) NULL
First Name Text(0) NULL
Surname Text(0) NULL
DOB Date(0) NULL
Operation Text(0) NULL
Date of Operation Date(0) NULL
Clinical Indication Text(0) NULL
Specimen Type Text(0) NULL
Histology Summary Text(0) NULL
Histology Microscopic Text(0) NUL
Histology Macroscopic Text(0) NUI
Consultant Pathologist Text(0) NUL
Reporting Pathologist Text(0) NULL
Consultant Surgeon Text(0) NULL
LastUpdated Date(0) NULL
Comments Text(0) NULL
Pathology Number Text(0) NULL
PathID Text(0) NOT NULL (PK)
ReportAll Text(0) NULL
ReferedCaseHospital Text(0) NULL
ReferedCasePathNo Text(0) NULL
MainID Text(0) NULL
InvPstgID Text(0) NULL
Hospital Number Path Text(0) NUL

Oncol Procedure_Surgery
Surgery Name Text(0) NULL
OPCS 4.4 Code Text(0) NULL
Procedure Text(0) NULL
SurgeryCode Text(0) NULL

ImagingPrimaryEpisode
PRImagID Text(0) NOT NULL (PK)
EpisID Text(0) NULL
MRICompartment Text(0) NULL
MRISuspectedTissue of origin Text(0) NULL
MRIPrimaryDate Date(0) NULL
MRITumourSite Text(0) NULL
MRITumour size Text(0) NULL
PETPrimaryDescriptionLesion 1 Text(0) NULL
PETPrimarySUVmax Text(0) NULL
PETMetastases Text(0) NULL
PETPrimaryDate Date(0) NULL
PETPrimaryExaminationSite 1 Text(0) NULL
XrayTissueInvolvement Text(0) NULL
XrayPrimaryDate Date(0) NULL
Xray_SingleMultiple Text(0) NULL
BoneScanPrimaryDate Date(0) NULL
BoneScanUptake 1 Text(0) NULL
BoneScan_SolitaryMultiple Text(0) NULL
USSPrimaryDate Date(0) NULL
USSStructure 1 Text(0) NULL
XrayOssification Calcification Text(0) NULL
USSExaminationSite Text(0) NULL
USSLesionLocation 1 Text(0) NULL
BoneScanLesionLocation 1 Text(0) NULL
XRayExaminationSite Text(0) NULL
XRaySummaryReport Text(0) NULL
USSConsistency 1 Text(0) NULL
CTPrimaryDate Date(0) NULL
CTExaminationSite Text(0) NULL
XRayCompartment Text(0) NULL
CTPrimaryComments Text(0) NULL
MainID Text(0) NULL
Tumour Type Text(0) NULL
XRayPathologicalFracture Text(0) NULL
PETPrimaryDescriptionLesion 2 Text(0) NULL
PETPrimarySUVmax 2 Text(0) NULL
PETPrimaryDescriptionLesion 3 Text(0) NULL
PETPrimarySUVmax 3 Text(0) NULL
USSLesionLocation 2 Text(0) NULL
USSStructure 2 Text(0) NULL
USSLesionLocation 3 Text(0) NULL
USSStructure 3 Text(0) NULL
BoneScanLesionLocation 2 Text(0) NULL
BoneScanUptake 2 Text(0) NULL
BoneScanLesionLocation 3 Text(0) NULL
BoneScanUptake 3 Text(0) NULL
USSConsistency 2 Text(0) NULL
USSConsistency 3 Text(0) NULL
MRIDiagnosis Text(0) NULL
CTLesionLocation Text(0) NULL
CTPathologicalFracture Text(0) NULL
CTCorticalDestruction Text(0) NULL
CTMetastases Text(0) NULL
USSImagingCharacteristics Text(0) NULL
USSDiagnosis Text(0) NULL
XRayLesion Location Text(0) NULL
XRayBone Expansion Text(0) NULL
XRayCorticalDestruction Text(0) NULL
XRayMarginalCharacteristics Text(0) NULL
RadiologyStagingComents Text(0) NULL

Chemotherapy
Description Text(0) NULL
StartDate Text(0) NULL
Regime Text(0) NULL
Toxicity Text(0) NULL
ToxicityDetails1 Text(0) NULL
EndDate Text(0) NULL
Intent Text(0) NULL
ClinicalTrialType Text(0) NULL
StatusClTrialEntry Text(0) NULL
EpisID Text(0) NULL
ToxicityITU1 Text(0) NULL
ToxicityCycle1 Number(-1,-1) NULL
ChemoID Text(0) NOT NULL (PK)
ToxicityGrade1 Number(-1,-1) NULL
NoCycles Text(0) NULL
Response Text(0) NULL
ToxicityDetails 2 Text(0) NULL
ToxicityITU 2 Text(0) NULL
ToxicityCycle 2 Number(-1,-1) NUL
ToxicityGrade 2 Number(-1,-1) NUL
Clinical Response Text(0) NULL
BTDescriptionLesion 1 Text(0) NULI
BTDescriptionLesion 2 Text(0) NULI
BTDescriptionLesion 3 Text(0) NULI
BTDescriptionLesion 4 Text(0) NULI
BTDescriptionLesion 5 Text(0) NULI
BNTDescriptionLesion 1 Text(0) NU
BNTDescriptionLesion 2 Text(0) NU
BNTDescriptionLesion 3 Text(0) NU
BNTDescriptionLesion 4 Text(0) NU
BNTDescriptionLesion 5 Text(0) NU
PETBaselineDate Date(0) NULL
PETBaselineSUVLesion 1 Text(0) N
PETBaselineSUVLesion 2 Text(0) N
PETBaselineSUVLesion 3 Text(0) N
PETBaselineSUVLesion 4 Text(0) N
PETBaselineSUVLesion 5 Text(0) N
PETAfterSUVLesion 1 Text(0) NULL
PETAfterSUVLesion 2 Text(0) NULL
PETAfterSUVLesion 3 Text(0) NULL
PETAfterSUVLesion 4 Text(0) NULL
PETAfterSUVLesion 5 Text(0) NULL
PETDate Date(0) NULL
ClinicalResponseDate Date(0) NUL
PETResponse Text(0) NULL
OverallResponse Text(0) NULL
Trial Specification Text(0) NULL
Oncologist Text(0) NULL
Chemo planned Text(0) NULL

Chemotherapy_Response
EpisID Text(0) NULL
ChemoID Text(0) NULL
BTDiameterLesion 1 Text(0) NULL
BTDiameterLesion 2 Text(0) NULL
BTDiameterLesion 3 Text(0) NULL
BTDiameterLesion 4 Text(0) NULL
BTDiameterLesion 5 Text(0) NULL
BAssessmentDate Text(0) NULL
AAssessmentDate Text(0) NULL
ATDiameterLesion 1 Text(0) NULL
ATDiameterLesion 2 Text(0) NULL
ATDiameterLesion 3 Text(0) NULL
ATDiameterLesion 4 Text(0) NULL
ATDiameterLesion 5 Text(0) NULL
ANTLesion code 1 Text(0) NULL
ANTLesion code 2 Text(0) NULL
ANTLesion code 3 Text(0) NULL
ANTLesion code 4 Text(0) NULL
ANTLesion code 5 Text(0) NULL
TargetResponse Text(0) NULL
NonTargetResponse Text(0) NULL
ATNewLesions Text(0) NULL
RCOverallResponce Text(0) NULI
ANTNewLesions Text(0) NULL
ManOverallResponce Text(0) NUL
RespChemoID Text(0) NOT NULL
AssessmentMethod Text(0) NULL

Figure 7: Data model of the SarcDB research database

Each patient has a number of episodes of investigation and each may contain the pathology report, surgical information, imaging and treatment. Please see the full data model of the research database.

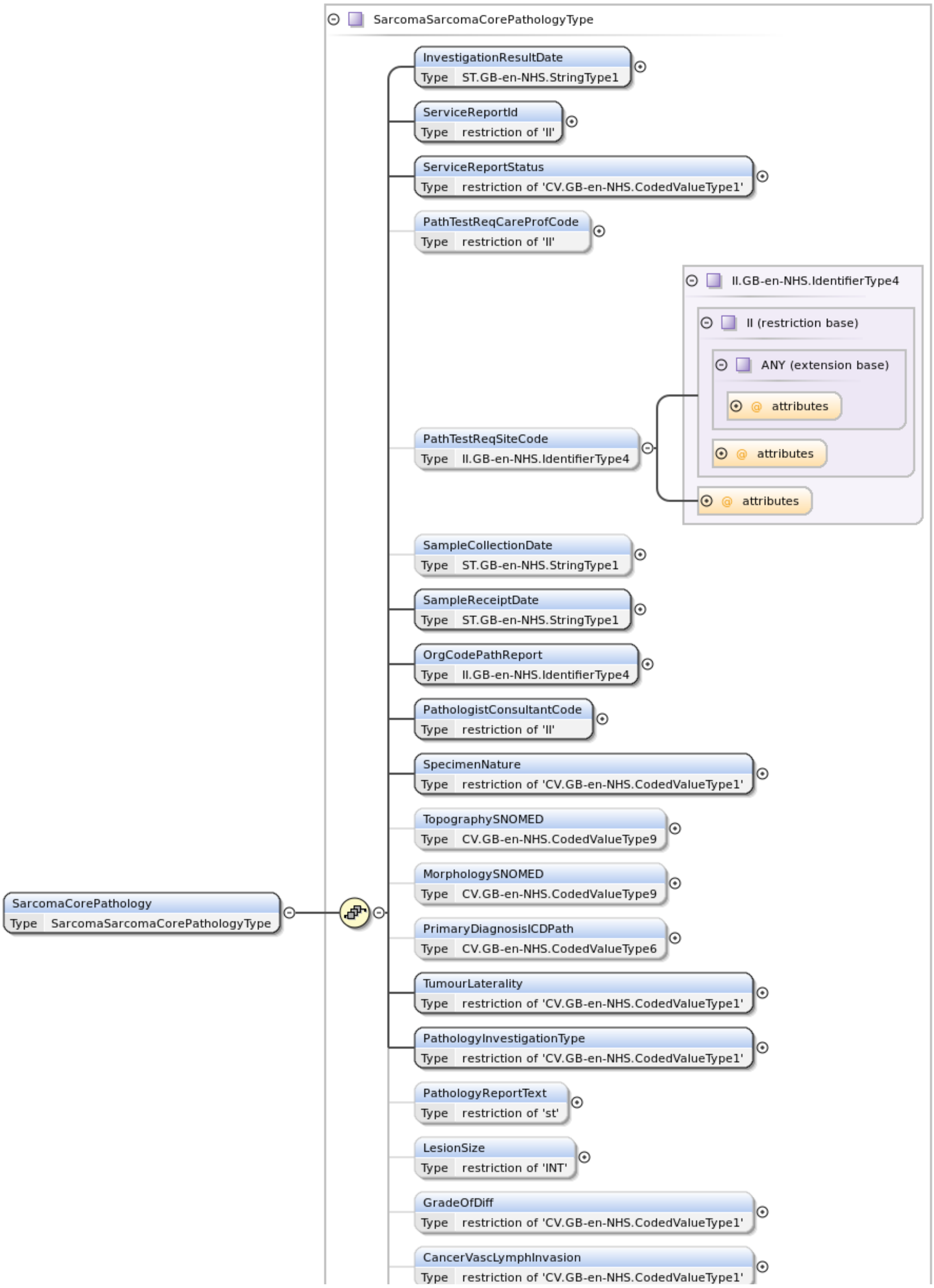

SarcomaSarcomaCorePathologyType
InvestigationResultDate
Type ST.GB-en-NHS.StringType1
ServiceReportId
Type restriction of 'II'
ServiceReportStatus
Type restriction of 'CV.GB-en-NHS.CodedValueType1'
PathTestReqCareProfCode
Type restriction of 'II'
II.GB-en-NHS.IdentifierType4
II (restriction base)
ANY (extension base)
attributes
attributes
attributes
PathTestReqSiteCode
Type II.GB-en-NHS.IdentifierType4
SampleCollectionDate
Type ST.GB-en-NHS.StringType1
SampleReceiptDate
Type ST.GB-en-NHS.StringType1
OrgCodePathReport
Type II.GB-en-NHS.IdentifierType4
PathologistConsultantCode
Type restriction of 'II'
SpecimenNature
Type restriction of 'CV.GB-en-NHS.CodedValueType1'
TopographySNOMED
Type CV.GB-en-NHS.CodedValueType9
MorphologySNOMED
Type CV.GB-en-NHS.CodedValueType9
PrimaryDiagnosisICDPath
Type CV.GB-en-NHS.CodedValueType6
TumourLaterality
Type restriction of 'CV.GB-en-NHS.CodedValueType1'
PathologyInvestigationType
Type restriction of 'CV.GB-en-NHS.CodedValueType1'
PathologyReportText
Type restriction of 'st'
LesionSize
Type restriction of 'INT'
GradeOfDiff
Type restriction of 'CV.GB-en-NHS.CodedValueType1'
CancerVascLymphInvasion
Type restriction of 'CV.GB-en-NHS.CodedValueType1'
SarcomaCorePathology
Type SarcomaSarcomaCorePathologyType

Figure 8: A type of SarcomaCorePathology in the COSD data model

Many interfaces could be developed to work with COSD such as SOAP and RESTful web services, data dumps etc.

2.5 Universität des Saarlandes (UdS)

UdS is involved in paediatrics, clinical trials and treatment of kidney cancer in children.

2.5.1 Hospital Information system

At UdS, patient care data is stored in an internally developed EHR system called HIS (Hospital Information System). HIS is based on SAP R/3 with the Patient Management (IS-H), the Clinical Data Management (i.s.h.med), and the modules for surgery (i.s.h.med OP)and Radiology (i.s.h.med RAD). In administration and logistics of the hospital SAP modules are implemented. For different departments of the hospital (e.g. Laboratory, Radiology, Infectiology, etc.) specific subsystems are used, Data exchange between the systems are done via a central Communication server. The structure of the HIS is shown in Figure 9.

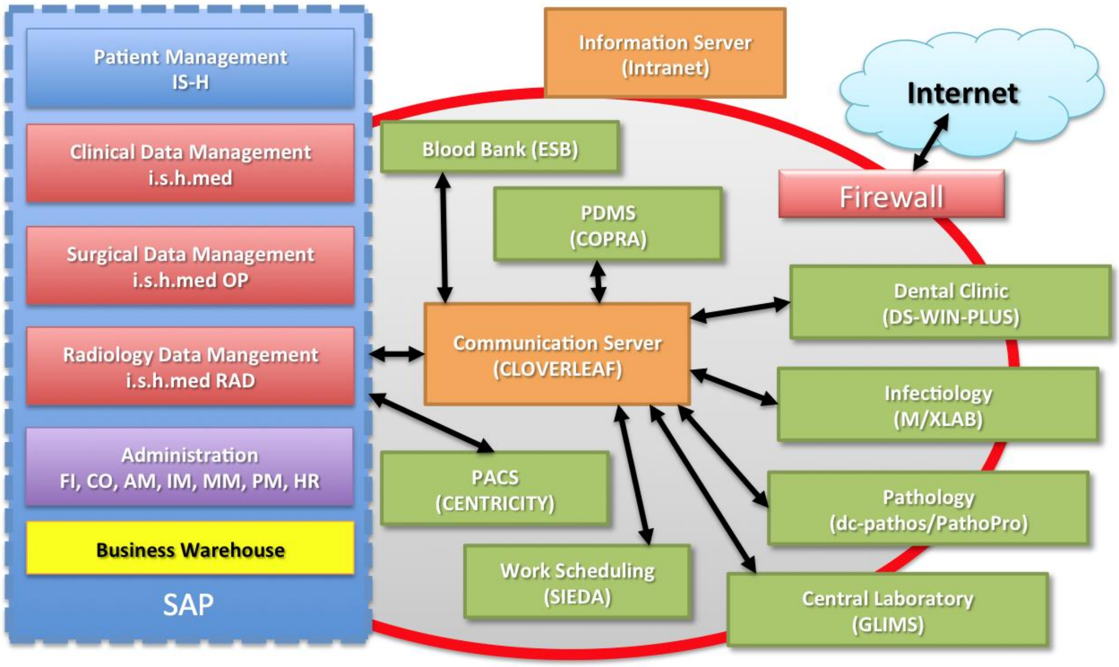
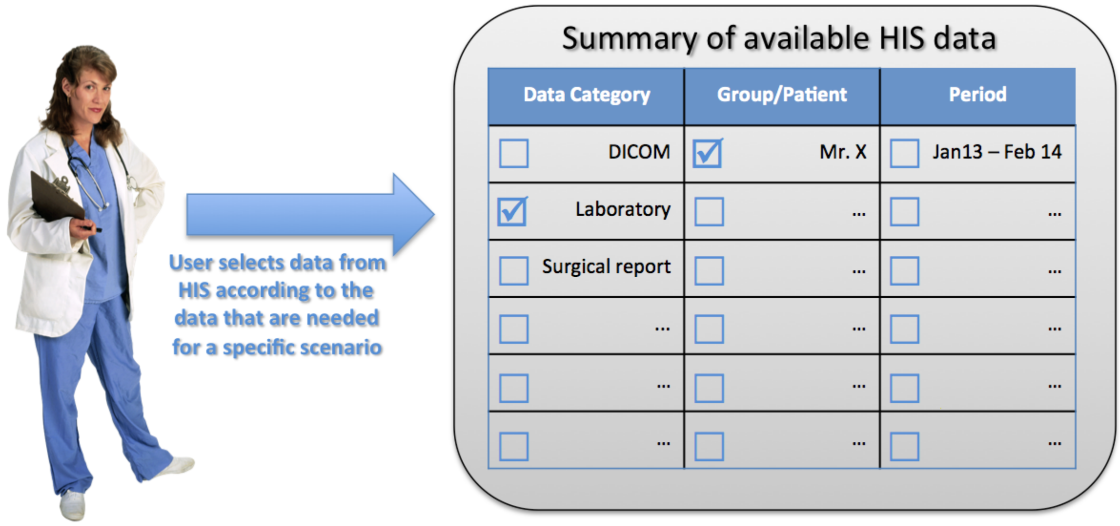


Figure 9: The structure of HIS at UdS. Names in brackets are the names of the specific software in use.

HIS is a complex system consisting of over 200 databases, including patient demographics, diagnosis information, imaging data (PACS), lab data and little treatment information. This information is recorded in structured or unstructured, free text format. The following main data are provided in the SAP system:

Master data of patients:	structured
Diagnosis:	ICD 10 plus text
Medical Procedures:	OPS301 and text
DRGs:	structured according the DRG catalogue with text
Medical Reports:	unstructured, mainly Word or PDF documents
Laboratory data:	structured data

To re-use the data from HIS it is important to know, which data are needed for a specific tool. A download of all data is not possible by regulations given by the local data protection officer. For that reason, to retrieve data, an end-user of the HIS needs to define which data, from which patient(s) and over which time period is needed. After pseudonymisation in the HIS system these data can be exported to a communication server, from which other tools can get access to the data. Such a push service will be developed by Fraunhofer IBMT in close cooperation with UdS, Custodix and UPM. For that purpose a graphical user interface is currently developed that will allow data extraction from this complex system in a structured way (XML format) (Figure 10).


User selects data from HIS according to the data that are needed for a specific scenario

Summary of available HIS data

Data Category	Group/Patient	Period
□ DICOM	☑ Mr. X	□ Jan13 – Feb 14
☑ Laboratory	□ ...	□ ...
□ Surgical report	□ ...	□ ...
□ ...	□ ...	□ ...
□ ...	□ ...	□ ...
□ ...	□ ...	□ ...

Figure 10: A possibility of a graphical user interface.

2.5.2 Clinical trial database for the SIOP Nephroblastoma trial

Data from the SIOP (Sociétè Internationale d'Oncologie Pédiatrique, or International society of peadiatric oncology) nephroblastoma clinical trial is recorded in a specific SIOP database that is shown with all tables in D9.1 (Figure 11).

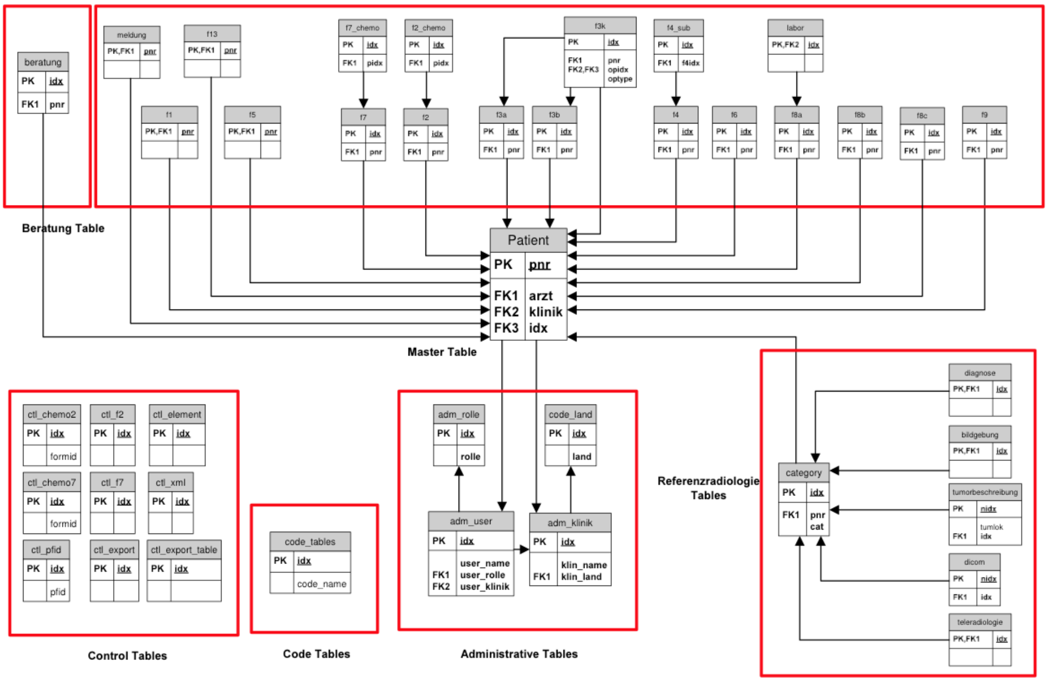
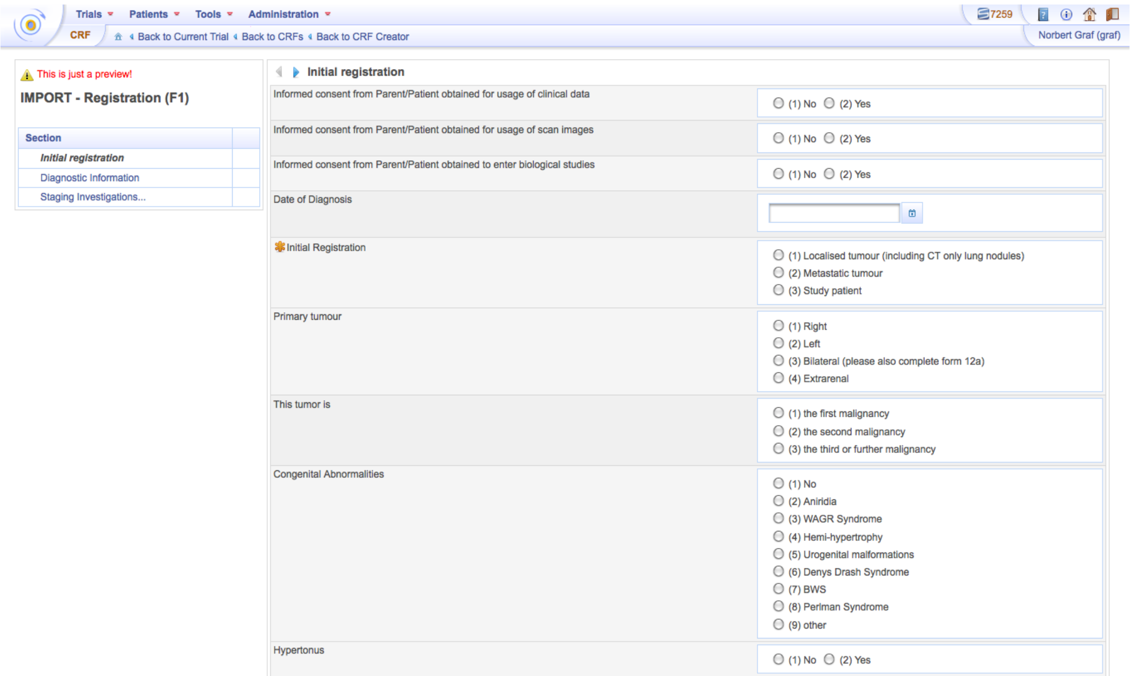


Figure 11: Schema of the SIOP nephroblastoma database

This database contains in addition to trial data also consultation data in CRFs. All CRFs contain structured information. The consultation data consists in addition of free text data. This database is anonymized for usage within EURECA. The anonymization was carried out by CUSTODIX and anonymity was controlled by USAAR.

These clinical trial data will be transferred to ObTiMA in the near future. ObTiMA is an ontology based trial builder and management system, developed in the ACGT and p-medicine project[3]. For that purpose the corresponding CRFs for the SIOP nephroblastoma are already created (Figure 12).

[3] http://www.obtima.org



This is just a preview!

IMPORT - Registration (F1)

Section
Initial registration
Diagnostic Information
Staging Investigations...

Initial registration

Informed consent from Parent/Patient obtained for usage of clinical data	○ (1) No ○ (2) Yes
Informed consent from Parent/Patient obtained for usage of scan images	○ (1) No ○ (2) Yes
Informed consent from Parent/Patient obtained to enter biological studies	○ (1) No ○ (2) Yes
Date of Diagnosis	
Initial Registration	○ (1) Localised tumour (including CT only lung nodules) ○ (2) Metastatic tumour ○ (3) Study patient
Primary tumour	○ (1) Right ○ (2) Left ○ (3) Bilateral (please also complete form 12a) ○ (4) Extrarenal
This tumor is	○ (1) the first malignancy ○ (2) the second malignancy ○ (3) the third or further malignancy
Congenital Abnormalities	○ (1) No ○ (2) Aniridia ○ (3) WAGR Syndrome ○ (4) Hemi-hypertrophy ○ (5) Urogenital malformations ○ (6) Denys Drash Syndrome ○ (7) BWS ○ (8) Perlman Syndrome ○ (9) other
Hypertonus	○ (1) No ○ (2) Yes

Figure 12: Part of the registration CRF of the SIOP nephroblastoma trial in ObTiMA

3 Data sources / available datasets

Deliverable D9.1 described the data that is available from the clinical sites in the project. Two main categories of data are provided; data from clinical trials and patient care data. One of the goals of the EURECA project is to enable information exchange between research and practice. This means that the data from systems which are used to gather data in clinical trials and the data from the different systems which are used in managing patient information needs to be integrated into one common information model. This information model is provided in Section 4.2. In order to arrive at this information model, we specify here the data that is used by the clinical partners in their clinical trials (Section 3.2) and in their patient care (Section 3.3) that is relevant for the scenarios developed in this project. The scenarios are described in Section 3.1.

3.1 Scenarios

This section gives a short description of the relevant scenarios per clinical partner. These scenarios were linked to 6 technical use cases. Figure 13 shows 6 categories of scenarios and how they are interrelated. Figures Figure 14 and Figure 15 show the technical use cases and how they are linked to the scenarios from the clinical partners.

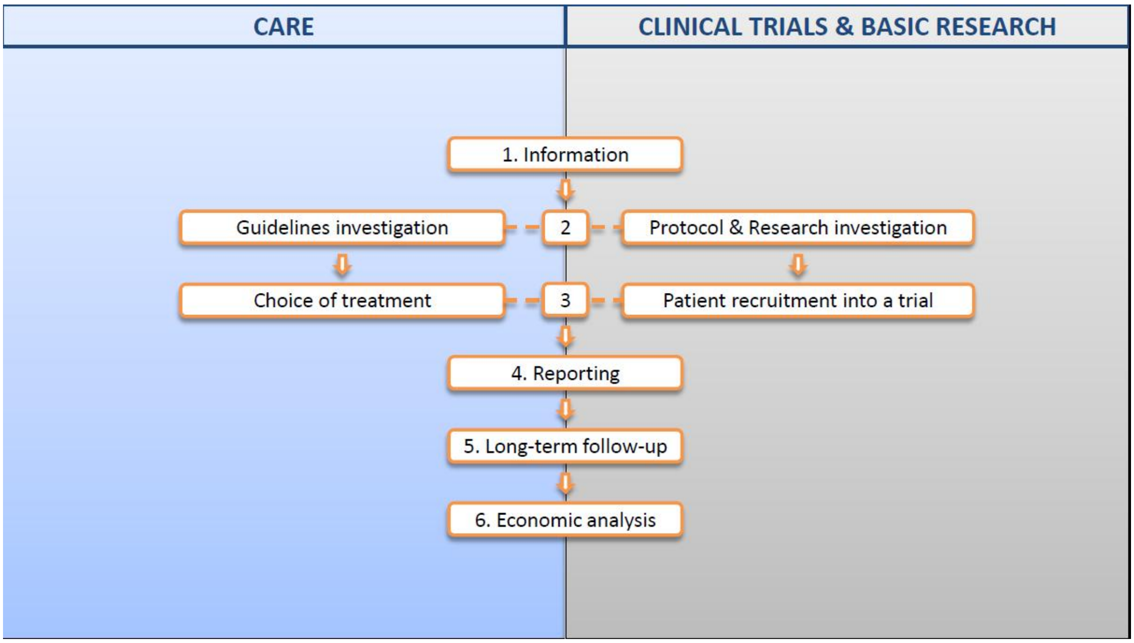
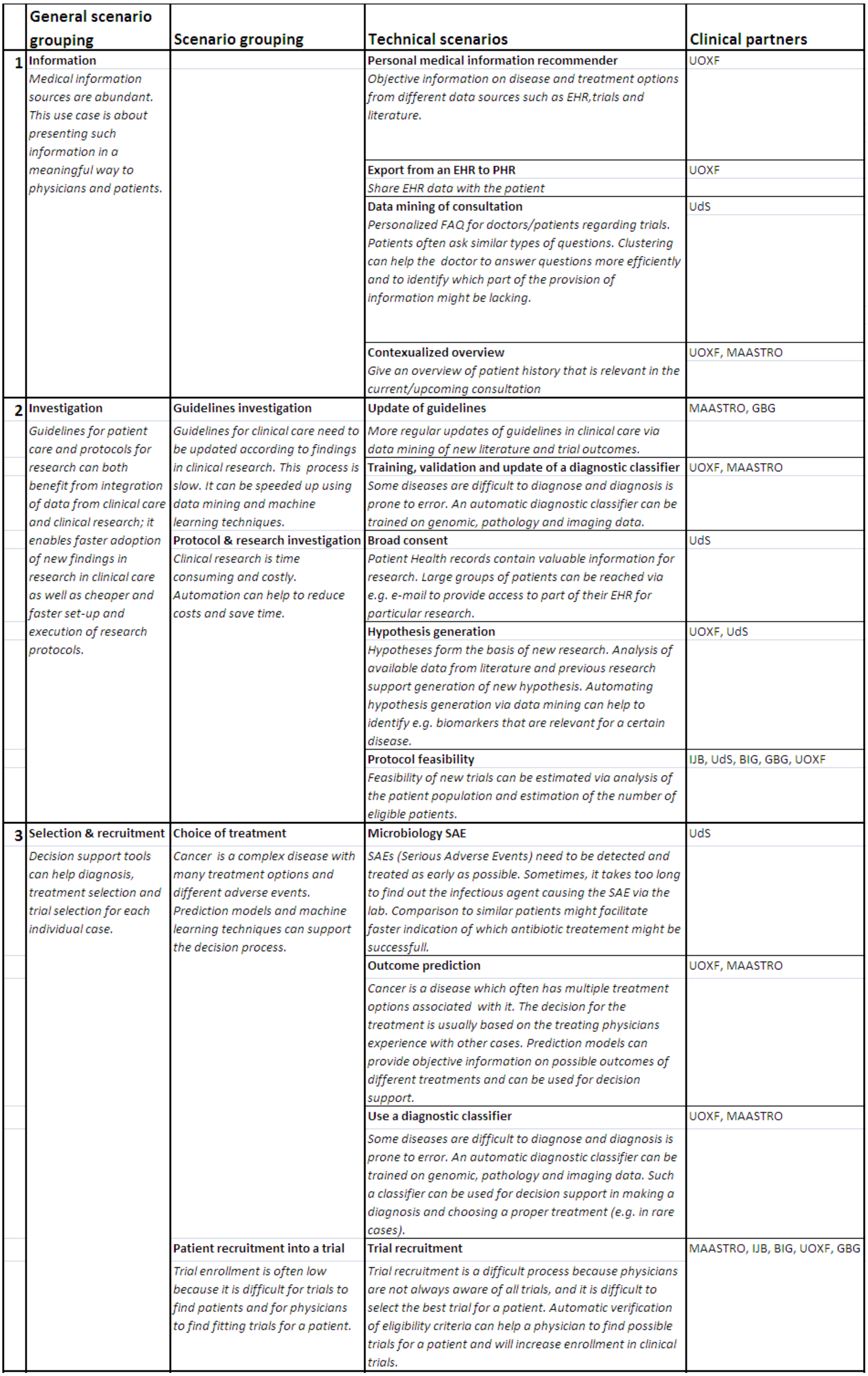
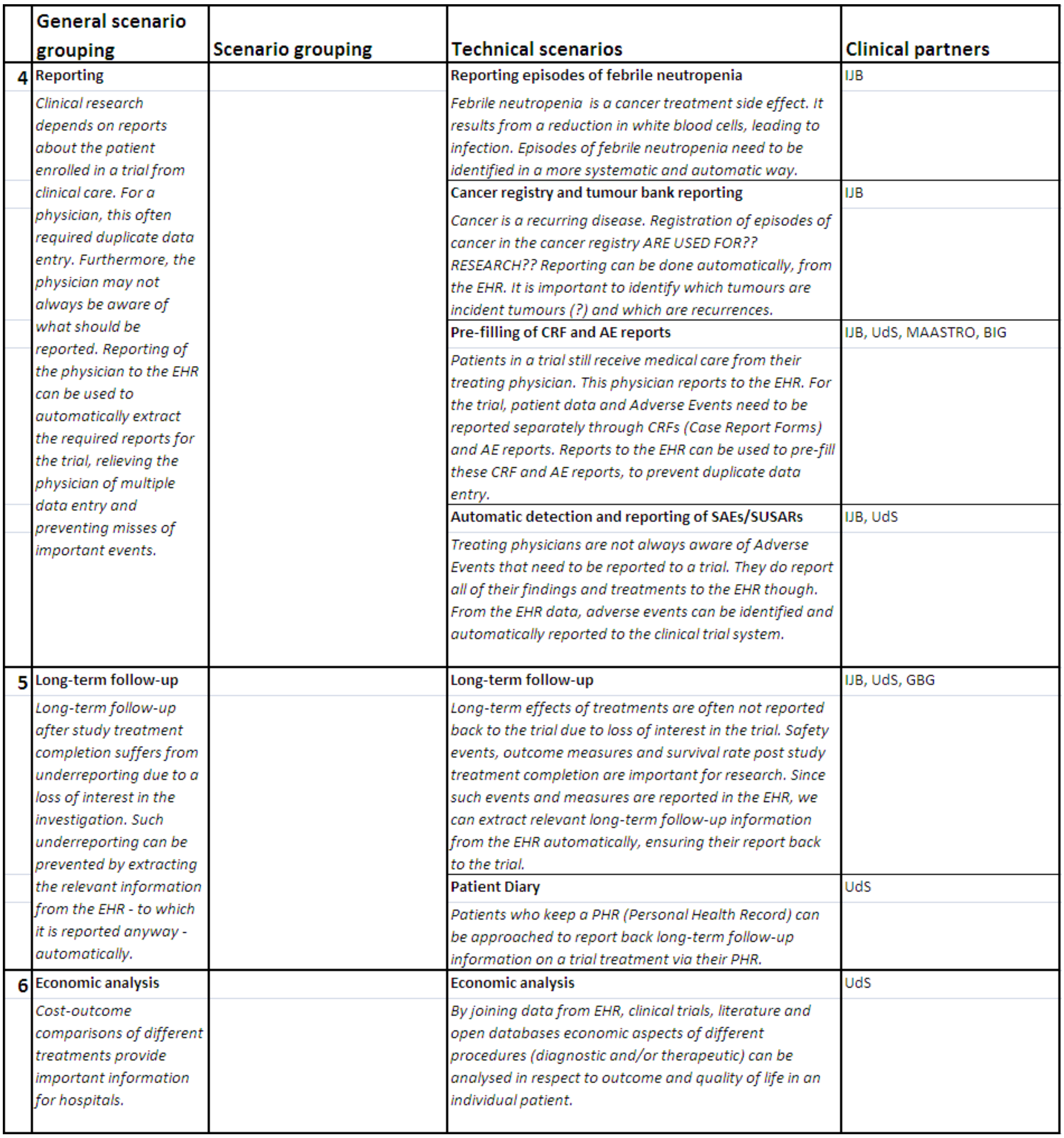


Figure 13: Interrelation of categories of scenarios



	General scenario grouping	Scenario grouping	Technical scenarios	Clinical partners
1	**Information** *Medical information sources are abundant. This use case is about presenting such information in a meaningful way to physicians and patients.*		**Personal medical information recommender** *Objective information on disease and treatment options from different data sources such as EHR,trials and literature.*	UOXF
			Export from an EHR to PHR *Share EHR data with the patient*	UOXF
			Data mining of consultation *Personalized FAQ for doctors/patients regarding trials. Patients often ask similar types of questions. Clustering can help the doctor to answer questions more efficiently and to identify which part of the provision of information might be lacking.*	UdS
			Contexualized overview *Give an overview of patient history that is relevant in the current/upcoming consultation*	UOXF, MAASTRO
2	**Investigation** *Guidelines for patient care and protocols for research can both benefit from integration of data from clinical care and clinical research; it enables faster adoption of new findings in research in clinical care as well as cheaper and faster set-up and execution of research protocols.*	**Guidelines investigation** *Guidelines for clinical care need to be updated according to findings in clinical research. This process is slow. It can be speeded up using data mining and machine learning techniques.*	**Update of guidelines** *More regular updates of guidelines in clinical care via data mining of new literature and trial outcomes.*	MAASTRO, GBG
			Training, validation and update of a diagnostic classifier *Some diseases are difficult to diagnose and diagnosis is prone to error. An automatic diagnostic classifier can be trained on genomic, pathology and imaging data.*	UOXF, MAASTRO
		Protocol & research investigation *Clinical research is time consuming and costly. Automation can help to reduce costs and save time.*	**Broad consent** *Patient Health records contain valuable information for research. Large groups of patients can be reached via e.g. e-mail to provide access to part of their EHR for particular research.*	UdS
			Hypothesis generation *Hypotheses form the basis of new research. Analysis of available data from literature and previous research support generation of new hypothesis. Automating hypothesis generation via data mining can help to identify e.g. biomarkers that are relevant for a certain disease.*	UOXF, UdS
			Protocol feasibility *Feasibility of new trials can be estimated via analysis of the patient population and estimation of the number of eligible patients.*	IJB, UdS, BIG, GBG, UOXF
3	**Selection & recruitment** *Decision support tools can help diagnosis, treatment selection and trial selection for each individual case.*	**Choice of treatment** *Cancer is a complex disease with many treatment options and different adverse events. Prediction models and machine learning techniques can support the decision process.*	**Microbiology SAE** *SAEs (Serious Adverse Events) need to be detected and treated as early as possible. Sometimes, it takes too long to find out the infectious agent causing the SAE via the lab. Comparison to similar patients might facilitate faster indication of which antibiotic treatement might be successfull.*	UdS
			Outcome prediction *Cancer is a disease which often has multiple treatment options associated with it. The decision for the treatment is usually based on the treating physicians experience with other cases. Prediction models can provide objective information on possible outcomes of different treatments and can be used for decision support.*	UOXF, MAASTRO
			Use a diagnostic classifier *Some diseases are difficult to diagnose and diagnosis is prone to error. An automatic diagnostic classifier can be trained on genomic, pathology and imaging data. Such a classifier can be used for decision support in making a diagnosis and choosing a proper treatment (e.g. in rare cases).*	UOXF, MAASTRO
		Patient recruitment into a trial *Trial enrollment is often low because it is difficult for trials to find patients and for physicians to find fitting trials for a patient.*	**Trial recruitment** *Trial recruitment is a difficult process because physicians are not always aware of all trials, and it is difficult to select the best trial for a patient. Automatic verification of eligibility criteria can help a physician to find possible trials for a patient and will increase enrollment in clinical trials.*	MAASTRO, IJB, BIG, UOXF, GBG

Figure 14: Technical use cases 1-3 and their links to scenarios from the partners



	General scenario grouping	Scenario grouping	Technical scenarios	Clinical partners
4	**Reporting**		**Reporting episodes of febrile neutropenia**	IJB
	Clinical research depends on reports about the patient enrolled in a trial from clinical care. For a physician, this often required duplicate data entry. Furthermore, the physician may not always be aware of what should be reported. Reporting of the physician to the EHR can be used to automatically extract the required reports for the trial, relieving the physician of multiple data entry and preventing misses of important events.		*Febrile neutropenia is a cancer treatment side effect. It results from a reduction in white blood cells, leading to infection. Episodes of febrile neutropenia need to be identified in a more systematic and automatic way.*	
			Cancer registry and tumour bank reporting	IJB
			Cancer is a recurring disease. Registration of episodes of cancer in the cancer registry ARE USED FOR?? RESEARCH?? Reporting can be done automatically, from the EHR. It is important to identify which tumours are incident tumours (?) and which are recurrences.	
			Pre-filling of CRF and AE reports	IJB, UdS, MAASTRO, BIG
			Patients in a trial still receive medical care from their treating physician. This physician reports to the EHR. For the trial, patient data and Adverse Events need to be reported separately through CRFs (Case Report Forms) and AE reports. Reports to the EHR can be used to pre-fill these CRF and AE reports, to prevent duplicate data entry.	
			Automatic detection and reporting of SAEs/SUSARs	IJB, UdS
			Treating physicians are not always aware of Adverse Events that need to be reported to a trial. They do report all of their findings and treatments to the EHR though. From the EHR data, adverse events can be identified and automatically reported to the clinical trial system.	
5	**Long-term follow-up**		**Long-term follow-up**	IJB, UdS, GBG
	Long-term follow-up after study treatment completion suffers from underreporting due to a loss of interest in the investigation. Such underreporting can be prevented by extracting the relevant information from the EHR - to which it is reported anyway - automatically.		*Long-term effects of treatments are often not reported back to the trial due to loss of interest in the trial. Safety events, outcome measures and survival rate post study treatment completion are important for research. Since such events and measures are reported in the EHR, we can extract relevant long-term follow-up information from the EHR automatically, ensuring their report back to the trial.*	
			Patient Diary	UdS
			Patients who keep a PHR (Personal Health Record) can be approached to report back long-term follow-up information on a trial treatment via their PHR.	
6	**Economic analysis**		**Economic analysis**	UdS
	Cost-outcome comparisons of different treatments provide important information for hospitals.		*By joining data from EHR, clinical trials, literature and open databases economic aspects of different procedures (diagnostic and/or therapeutic) can be analysed in respect to outcome and quality of life in an individual patient.*	

Figure 15: Technical use cases 4-6 and their links to scenarios from the partners

3.2 Case Report Forms (CRFs)/clinical trial data

Data collection in clinical trials takes place via Case Report Forms. The data collected on a Case Report Form can consist of the relevant patient data, eligibility criteria and a report of adverse events.

Relevant patient data consists of demographic data (anonymised) and data related to the disease such as tumor size, tumor grade, biomarker expression status. Clinical endpoints might also be included. During consultations, physicians fill in reports to verify eligibility criteria and adverse events. This patient data is collected via a user interface which shows the CRF and is typically stored in an ad hoc manner, using Oracle Clinical, Open Clinica, or similar software, a self constructed database or excel sheets.

Besides patient data which is reported in CRFs, pathological data relevant for the trial is also collected, consisting of for example genomic data.

In the following, we describe per clinical partner which data is available from clinical trials and which data is required for the implementation of a few example scenarios.

3.2.1 Institut Jules Bordet (IJB)

IJB is involved in the following scenarios:

- Protocol feasibility
- Trial recruitment
- Reporting episodes of febrile neutropenia
- Cancer registry and tumour bank reporting
- Pre-filling of CRF and AE reports
- Automatic detection and reporting of SAEs/SUSARs
- Long-term follow up

TOP trial data will be useful within the following scenarios:

- Protocol feasibility
- Trial recruitment and patient screening
- Pre-filling of CRF and AE reports
- Automatic detection and reporting of SAEs/SUSARs
- Long-term follow-up

For IJB, a description is provided of the data which is needed for the scenario “Reporting episodes of febrile neutropenia”. In this scenario, no clinical trial data is used, only patient care data.

3.2.2 German Breast Group (GBG)

GBG is involved in the following scenarios:

- Update of guidelines
- Protocol feasibility
- Trial recruitment
- Long term follow up

Below, a description is provided of the data which is needed for the “Trial recruitment” scenario.

As an example, we take a metastatic study, the patient details needed to verify eligibility criteria are listed below.

All inclusion/exclusion criteria are checked in a data entry form in the MedCODES system via radiobuttons, so data on eligibility criteria is all available in structured format in the form of an answer “yes” or “no” to the question whether this criterium is met.

Required data	Instantiation of required data type for the specific type of research executed at IJB	Available data
Eligibility criteria	index	structured
	date_rando	structured
	Treatment arm	structured
	Menopausal status	structured
	Height	structured
	Weight	structured
	Karnofsky index, %	structured
	Race	structured
	All inclusion criteria fulfilled	structured
	At least one exclusion criteria fulfilled	structured
	pT at 1st diagnosis	structured
	pN at 1st diagnosis	structured
	M at 1st diagnosis	structured
	Grade	structured
	Hormone receptor status	structured
	Her2	structured
	locoregional_BL	structured
	liver_BL	structured
	lung_BL	structured
	bone_BL	structured
	CNS_BL	structured
	other_BL	structured
	number_sites	structured
	Metastatic site binary	structured
	Haemoglobin, baseline	structured
	Leucocytes, baseline	structured
	Neutrophils, baseline	structured
	Thrombocytes, baseline	structured
	AP, baseline	structured
	SGOT, baseline	structured
	SGPT, baseline	structured
	Bilirubin, baseline	structured
	Serum Creatinine, baseline	structured
	ECG	structured
	Echocardiography	structured
	LVEF	structured
	age	structured
	Any chemotherapy	structured
	Chemotherapy adjuvant or neo-adjuvant	structured
	Chemotherapy palliative	structured
	Anthracycline-containing chemotherapy	structured
	Anthracycline-containing chemotherapy adjuvant or	structured

	neo-adjuvant	
	Anthracycline-containing chemotherapy palliative	structured
	Taxane-containing chemotherapy	structured
	Taxane-containing chemotherapy adjuvant or neo-adjuvant	structured
	Taxane-containing chemotherapy palliative	structured
	Endocrine therapy	structured
	Endocrine therapy adjuvant	structured
	Endocrine therapy palliative	structured
	Radiotherapy	structured
	Radiotherapy adjuvant	structured
	Radiotherapy palliative	structured
	Trastuzumab	structured
	Trastuzumab adjuvant	structured
	Trastuzumab palliative	structured
	Bisphosphonate treatment	structured
	Bisphosphonate treatment adjuvant	structured
	Bisphosphonate treatment palliative	structured
	Other treatments	structured
	Other treatments adjuvant	structured
	Other treatments palliative	structured
Consent form for screening	All patients are referred from external sites, so all have already provided consent for screening to their treating physician at that external site.	N.A.
Description of trial	Available via heads of hospitals or GBG Annual Meeting.	N.A.
Organization organizing the trial (for prospective recruitment)	Available via heads of hospitals or GBG Annual Meeting.	N.A.

3.2.3 MAASTRO

MAASTRO is involved in the following scenarios:

- Contextualized overview
- Update of guidelines
- Training, validation and update of a diagnostic classifier
- Outcome prediction
- Use a diagnostic classifier

- Trial recruitment
- Pre-filling of CRF and AE reports

Below, a description is provided of the data which is needed for the “Trial recruitment” scenario.

Required data	**Instantiation of required data type for the specific type of research executed at MAASTRO**	**Available data**
Eligibility criteria	Available from website/clinicaltrials.gov	Unstructured/free text
Consent form for screening	Available in the EMD (opt-out solution for research as regulated in the Netherlands)	Structured
Description of trial	Available from website/clinicaltrials.gov	Unstructured/free text
Organization organizing the trial (for prospective recruitment)	Available from clinicaltrials.gov	Unstructured/free text

3.2.4 University of Oxford (UOXF)

UOXF is involved in the following scenarios:

- Personal medical information recommender
- Export from an EHR to PHR
- Contextualized overview
- Training, validation and update of a diagnostic classifier
- Hypothesis generation
- Protocol feasibility
- Outcome prediction
- Use a diagnostic classifier
- Trial recruitment

Below, a description is provided of the data which is needed for the following scenarios:

- Training, validation and update of a diagnostic classifier
- Trial recruitment

<u>Available data for “training, validation and update of a diagnostic classifier”</u>

This scenario requires the following data from the clinical research systems:

- Oxford Sarcoma data (used to train and test the classifier):
    - Manually collected data from various NHS clinical systems across many hospitals in Oxford. It contains many aspects of the sarcoma patient information including demographics, patient outcome, pathology, chemotherapy, surgery information, biopsy information etc. Please see the detailed database schema in Section 2.4.
- EU-wide Sarcoma data (used to validate the classifier):
    - Similar to our sarcoma research database but from other European sites. The data schema will be similar to the sarcoma database. It will be gradually collected and become available in one year.

For this scenario, Oxford Sarcoma dataset is used to train and test the classifier. EU-wide Sarcoma data will be used to validate the classifier. This data will be available in one year.

Required data	**Instantiation of required data type for the specific type of research executed at UOXF**	**Available data**
Reported cases of sarcoma	Sarcoma research database:	All data, mostly structured data. A small amount of free text.

Available data for "trial recruitment"

Required data	**Instantiation of required data type for the specific type of research executed at UOXF**	**Available data**
Eligibility criteria	Openclinica	To be discussed depending on exact application. Example data can be provided, however further ethics application/approval might be needed. This can be investigated once the use case is fully defined and the scope of the data use is defined. A specific use case description is needed for each dataset.
Consent form for screening	Example consent form from the biobank that stores the tissue sample from patients	Paper example consent form

3.2.5 Universität des Saarlandes (UdS)

UdS is involved in the following scenarios:

- Data mining of consultation
- Broad consent
- Hypothesis generation
- Protocol feasibility
- Microbiology SAE
- Pre-filling of CRF and AE reports
- Automatic detection and reporting of SAEs/SUSARs
- Long-term follow-up
- Patient diary
- Economic analysis

Below, a description is provided of the data, which is needed for the “Data mining of consultation” scenario. This data is taken from the Wilms tumour study (SIOP 2001/GPOH). The same data will be available in the future via ObTiMA. In that case specific CRFs for consultation will be available.

Required data	**Instantiation of required data type for the specific type of research executed at UdS**	**Available data**
Stage	“Stadium” (stage) on CRF form f1	Structured, numerical (1-4)
Histology	“Sicherung der Diagnose durch” (securing of diagnosis through) on CRF form f1	Structured, list
Preoperative chemo or primary surgery	“Vortherapie in anderer Klinik” (pretreatment in a different clinic) on CRF form f1	Semi-structured; “nein” (no), or “ja, wie:” (yes, how) followed by free text.
Age	“Geburtsdatum” (date of birth) on CRF form f1	Structured, numerical; day-month-year
Child has a syndrome?	“Syndrome/hereditäre Grunderkrankungen/assoziierte Fehlbildungen” (Syndrome / Hereditary underlying diseases / associated anomalies) on CRF form f1	Structured, choice
Metastatic?	“Metastasen gefunden” (metastases found) on CRF form f1	Structured, choice
Uni-/bilateral disease	“Seite” (site) on CRF form f1	Structured, choice
Patient condition	“Allgemeinzustand bei Diagnosestellung” (overall condition at the time of diagnosis) on CRF form f1	Structured, choice
Consultation question	“Beratung” (consultation) table in SIOP	Free text
Response	“Beratung” (consultation) table in SIOP	Free text

For the following scenarios data from HIS are needed:

- Microbiology SAE
- Pre-filling of CRF and AE reports
- Automatic detection and reporting of SAEs/SUSARs
- Economic analasys

These data will be made available via a push service as described briefly in Chapter 2.5.1 of this deliverable.

For the long-term follow-up scenario, data from the cancer registry of the Saarland are needed. At the moment access to this data is under negotiation. A data structure is not available before agreement to use these data.

3.3 EHR data/patient care data

Patient care data is gathered in a variety of systems. Most hospitals use an EHR system, either self-developed (Oribase, developed at Institut Jules Bordet, the MAASTRO EMD and the UdS Hospital Information System), or an existing system (Cerner Millenium EPR (www.cerner.com), used at UOXF). These systems typically store demographic data, appointments, and hospital stay information. Laboratory data, clinical data (consults), treatment data, pathology data and imaging data are usually stored in separate systems, linking to the EHR.

In the following, we describe for a few example scenarios, what data is required and where that data is localized in the EHR or other systems used at the site.

3.3.1 Institut Jules Bordet (IJB)

IJB is involved in the following scenarios:

- Protocol feasibility
- Trial recruitment
- Reporting episodes of febrile neutropenia
- Cancer registry and tumour bank reporting
- Pre-filling of CRF and AE reports
- Automatic detection and reporting of SAEs/SUSARs
- Long-term follow up

Below, a description is provided of the data which is needed for the scenario "Reporting episodes of febrile neutropenia"

Available data for "reporting episodes of febrile neutropenia"

For this scenario, and in the context of looking for patients who have had an episode of febrile neutropenia, we need structured or unstructured (textual) data in French to extract the following concepts. The following data from IJB will be used to this end:

- Cancer registry data
- MDT data
- Consult and discharge reports
- Anatomical pathology data
- Laboratory data

Febrile neutropenia is the event we would like to identify (with negative predictive value, the positive predictive value is less crucial). This is the conjunction the same day of neutropenia and fever. In clinical research, most often, we are interested in chemotherapy-induced febrile neutropenia.

- Fever:

Fever (also known as pyrexia) is defined as an oral temperature ≥ 38.5°C or ≥ 38°C (2 measurements in a 12-hours period separated by 3-4 hours). Fever can be documented in the hospital (in case of a patient hospitalised or when the patient comes to a consultation) but can also be documented at home by the patient himself. Identification of febrile episodes is by far the most relevant information to extract from the medical chart.

Data	Standard phrases	Analysis
Consult and discharge report	CONCLUSION(S): Syndrome inflammatoire et neutropénie sans point d'appel infectieux à l'interrogatoire et à l'examen clinique, recommandation au patient de surveiller la température 2 x par jour et se présenter aux urgences au moindre pic fébrile ou au moindre signe infectieux apparaissant.	Surveillance ⇒ neutropenic without fever

- <u>Neutropenia:</u>

Neutropenia is an abnormally low level of neutrophils in the blood (neutrophils are a type of white blood cells, and are also known as polymorphonuclear leukocytes or PMNs). Neutropenia is defined as an absolute neutrophil count (ANC) < 500/mm3 (measured in mm3 of blood). The information about neutropenia can be found as structured data in the lab or as text data in a consultation note (or in an external scanned document).

Data	Standard phrases	Analysis
Consult and discharge report	La biologie qui date d'hier montre une hémoglobine à 8,1 g/dl, des plaquettes à 36.000/mm³, des GB à 940/mm³ avec GOT et GPT majorés à 76 et 148 U/l ainsi qu'une bilirubine totale à 1,3 mg/dl, en cours de normalisation. A noter une CRP à 43,2 mg/l en majoration.	White blood cells at 940 and apyretic ⇒ we do not know yet if the patient is neutropenic

- <u>Determine whether the patient received antibacterials (antibiotics) and/or prophylactic antifungals:</u>

The development of febrile neutropenia without any other obvious explanation than infection in a patient requires prompt antibiotic initiation.

At Institut Jules Bordet, the empiric antibiotic is chosen according to the risk estimated by the MASCC score:

- MASCC >=21 means low risk of febrile neutropenia and moxifloxacin is normally used
- MASCC < 21 means high risk of febrile neutropenia and cefepime or tazocin are normally used.

Data	Standard phrases	Analysis
Hospital summary	EVOLUTION AU COURS DE L'HOSPITALISATION: Durant l'hospitalisation, le patient a présenté une neutropénie fébrile avec un score MASCC évalué à 19.	At this level only suspiscion of febrile neutropenia

- <u>Treatment and treatment drugs' name:</u>

By the past, patients were all hospitalised for the treatment of febrile neutropenia. However, following clinical research, it has become obvious that febrile neutropenia is a heterogeneous syndrome and now the management is done according to risk of serious medical complication : intravenous antibiotics versus oral antibiotics, hospitalisation versus early or immediate discharge.

We are also interested by drugs' name, both in prophylaxis (preventive) and within treatment (empirical or not).

Date	Standard phrases	Analysis
Hospital summary	EVOLUTION AU COURS DE L'HOSPITALISATION: Le patient a été mis sous antibiothérapie à large spectre par Tazocin 4 g x 4 qui l'a reçue pendant 11 jours.	Empirical treatment of febrile neutropenia

- Determine wether the neutropenia is chemotherapy-induced:

A chemotherapy-induced febrile neutropenia means that chemotherapy should have been administered within a period of 30 days before onset of febrile neutropenia.

Date	Standard phrases	Analysis
Hospital summary From 19/01 to 05/02	CONCLUSIONS: Le patient est actuellement en 2ème ligne thérapeutique par Folfiri. Il a reçu durant l'hospitalisation la deuxième cure le 20.01.	Febrile neutropenia seems chemotherapy-induced
Hospital summary	EVOLUTION EN COURS D'HOSPITALISATION: Induction selon le protocole GRAALL 2005 le 1er janvier 2006 avec cortico-sensibilité mais chimio-résistance au J8 (blastes à 6.9 %).	Indication of chemotherapy prescription
Hospital summary	EVOLUTION EN COURS D'HOSPITALISATION: Neutropénie fébrile au J16 du traitement avec hémoculture positive pour un staphylocoque traité avec succès par Tazocin, puis shift en Meronem.	Suspiscion of chemotherapy-induced febrile neutropenia, no notion of fever and ANC

- Date of admission:

The date of onset or date of admission is the first date of documentation of both fever and neutropenia (blood samples for documenting neutropenia in ambulatory patients are most often driven by the development of fever).

- Prior episodes of febrile neutropenia:

We want to know any previous episodes of febrile neutropenia in patient's life.

Date	Standard phrases	Analysis
Hospital summary From 19/01 to 05/02	CONCLUSION: Antécédent de prostatite à Entérococcusfaecalis traité début février.	This patient had an MDI before but nothing says that it was in the context of febrile neutropenia

- <u>Outcome of episodes of febrile neutropenia</u>

We classify it into 3 categories :

- Resolution without serious medical complication development (resolution means recovery from neutropenia and a period of five days without fever, serious medical complication is defined in the literature)
- Resolution with serious medical complication development
- Death before resolution of the febrile neutropenic episode

Data	Standard phrases	Analysis
Hospital summary	CONCLUSIONS: Evolution favorable sous antibiothérapie à large spectre (Tazocin).	Favorable development
Hospital summary	EVOLUTION EN COURS D'HOSPITALISATION: Vu la bonne évolution, le patient est sorti d'hospitalisation le 2 janvier 2012	Favorable development

- <u>Clinical and biological documentation of the infection:</u>

At onset of a febrile neutropenic episode, some actions will be taken to try to document the presence of an infection. These actions are physical examination to identify a clinical site of infection, radiological investigations (most often a chest X ray) also for a clinical documentation of infection and samples (blood, urine, stools) looking at a microbiologically documentation of the infection (the germ might be bacterial, viral or fungal, is is possible that several germs are responsible for the febrile neutropenia).

Once the results of the investigations are known, the episode can be classified as follows :

- Microbiologically documented infection (MDI) :
    - Bacteremia if a bacterial germ is found in the blood
    - Viremia if a viral germ is found in the blood
    - Fungemia if a fungal germ is found in the blood
- Microbiologically documented infection without bacteremia/fungemia/viremia : when a responsible pathogen is found but not in the blood (note that fungemias and viremias are rare and therefore that category of documentation is often called MDI without bacteremia). Bacterias are the most frequent responsible pathogens, this justifies why the first treatment is antibiotherapy (and not antifungal agents or antiviral agents)

- Clinically documented infection when a possible site of infection is identified but without identification of a pathogen (example : digestive tract, mucositis, respiratory tract, ...)
- of unknown origin : when nothing is documented but no other retrospectively assessed plausible explanation for the fever is found
- Fever not related to infection : when a non-infectious cause is identified retrospectively (for instance due to a drug but fever disappearance when the drug is stopped – rare because when fever disappears, it is difficult to attribute the resolution due to antibiotherapy or due to another reason).

Data	Standard phrases	Analysis
Hospital summary	EVOLUTION AU COURS DE L'HOSPITALISATION: Il a présenté à ce moment des selles liquides hémorragiques avec un tableau de colite au CT abdominal. Les coprocultures sont revenues négatives ainsi que les hémocultures.	• Negative blood cultures ⇒ no bacteremia • Negative stool cultures⇒ no other microbiological documentation • But clinical documentation (digestive)

3.3.2 GBG

GBG is not involved in regular treatment of patients. Therefore, they have no explicit EHR/patient data, except for the data which is available in their CRFs.

3.3.3 MAASTRO

MAASTRO is involved in the following scenarios:

- Contextualized overview
- Update of guidelines
- Training, validation and update of a diagnostic classifier
- Outcome prediction
- Use a diagnostic classifier
- Trial recruitment
- Pre-filling of CRF and AE reports

Below, a description is provided of the data which is needed for the “Trial recruitment” scenario.

As an example of data needed to verify possible criteria, we take the list of data provided by the ADPG (Advanced Data Patient Generator).

Required data	**Instantiation of required data type for the specific type of research executed at MAASTRO**	**Available data**
Contact information of patient (for retrospective recruitment)	EMD	Structured
Contact information of treating physician (for retrospective recruitment)	EMD	Structured
Informed consent for contacting the patient for research	EMD, lookup table	Structured
Care data:		
Menopausal Status	This information is usually not of interest, because they mainly treat older patients, if it is necessary to note, it is noted somewhere separately, it is not structurally recorded	Not available
Currently Pregnant	MAASTRO never treats pregnant patients	Per guideline
Currently Nursing	If this is the case, it is noted somewhere separately, but it is so rare that it is not structurally recorded	Not Available
Hispathology	EMD, pathology lookup table	Structured
HER2	EMD, Oncological history	Unstructured/free text
ER	EMD, Oncological history	Unstructured/free text
PR	EMD, Oncological history	Unstructured/free text
Stage	EMD, TNM lookup table	Structured
Tumor Size	ZyLAB, Surgical PA report	Unstructured, OCR scan
Lymph Nodes	EMD, Pre-op N Stage	Structured
Distant Metastases	EMD, M Stage	Structured

3.3.4 University of Oxford (UOXF)

UOXF is involved in the following scenarios:

- Personal medical information recommender
- Export from an EHR to PHR
- Contextualized overview
- Training, validation and update of a diagnostic classifier
- Hypothesis generation
- Protocol feasibility
- Outcome prediction
- Use a diagnostic classifier
- Trial recruitment

Below, a description is provided of the data which is needed for the following scenarios:

- Training, validation and update of a diagnostic classifier
- Trial recruitment

Available data for "training, validation and update of a diagnostic classifier"

Required data	**Instantiation of required data type for the specific type of research executed at UOXF**	**Available data**
Patient clinical information	EPR	Via cloned data warehouse, Mirth HL7 messaging engine etc (under investigation). Mostly structured
Pathology	Several pathology databases	Under investigation (Free text)
Imaging	PACS	Under investigation (structured)
Genomic data	Raw files	

Available data for "trial recruitment"

Required data	**Instantiation of required data type for the specific type of research executed at UOXF**	**Available data**
Patient data at consulting clinical centre	EPR	Via cloned data warehouse, Mirth HL7 messaging engine etc (under investigation). Mostly structured
Patient record	EPR	Via cloned data warehouse, Mirth HL7 messaging engine etc (under investigation). Mostly structured

3.3.5 Universität des Saarlandes (UdS)

UdS is involved in the following scenarios:

- Data mining of consultation
- Broad consent
- Hypothesis generation
- Protocol feasibility
- Microbiology SAE
- Pre-filling of CRF and AE reports
- Automatic detection and reporting of SAEs/SUSARs

- Long-term follow-up
- Patient diary
- Economic analysis

Below, a description is provided of the data, which is needed for the “Data mining of consultation” scenario.

For the “Data mining of consultation” scenario, the data required can be extracted from the CRF forms and the “Beratung” (consultation) table in the SIOP database (see Section 3.2.5).

The data of the SIOP database are already anonymized as described in Section 2.5 and are available for usage to all partners, who have signed the contracts of the legal and ethical framework of EURECA.

For the other scenarios data from the EHR of the hospital (HIS) are needed. For this purpose a graphical interface is developed (see Section 2.5) to allow the download of needed data on a communication server, where these data can be used for further usage.

4 EURECA Common Information Model

The EURECA Common Information Model has to fulfill two main objectives: (i) to provide a common structure capable of gathering any kind of medical data provided by the clinicians, in order to facilitate the reasoning and querying on the information; and (ii) to use a common vocabulary ensuring that data from different sources is going to be stored using the same terminology easing the work of clinicians and researchers.

4.1 Overview of Common Data Models

The aim of the Common Data Model in the EURECA project is to provide common storage of data from different Clinical Trials and EHR systems in a homogeneous way. Additionally, for the project it is very necessary to provide semantic interoperability. In the following sections describe data models used in similar projects: i2b2, OMOP, HL7 RIM and HDOT.

4.1.1 I2B2 (Informatics for Integrating Biology and the Bedside)

i2b2 Center[4] is developing the i2b2 framework. It aims to facilitate the design of targeted therapies for individual patients with diseases having genetic origins when combined with IRB-approved genomic data. It is also scalable and will enable researchers to use the clinical data for discovery research.

i2b2 framework is designed as a set of *cells*, in an environment named *Hive*, the set of all the cells, (SOA services). The i2b2 Hive consists of a number of core cells that establish basic services, as well as any number of additional cells to provide enhanced services.
The core services[5] of the Hive are:

- **Project management**, used to provide user authentication and manage group and role information. It also keeps track of what cells are parts of the Hive.
- **File repository**, holding large files of data including radiological images and genetic sequences. The files are generally referenced from the Data Repository cell.
- **Identity management**, used to manage a patient's protected health information in a manner consistent with the HIPAA privacy rule. Patient data is available only as a HIPAA defined "Limited Data Set" to most of the i2b2 services.
- **Workflow framework**, used to process information in steps through various parts of the hive. Most processed information will come to reside in the Data Repository Cell or as a display to the user.
- **Data repository** (CRC), containing the clinical data (phenotypic and genotypic data) in a structured format. Data queries and visualizations are available through this service.
- **Ontology management**, managing the terminology and knowledge information typically used in the hive. It is contacted for, or distributes knowledge to, cells during most of the hives transactions.

[4] https://www.i2b2.org/
[5] https://www.i2b2.org/software/index.html

The i2b2 Hive is centered around two concepts. The first concept is the existence of services provided by applications that are “wrapped” into functional units, such that their functionality are exposed as messages that travel to and from the various cells of the hive. The second concept is that of persistent data storage, which is managed by the cell named the “Clinical Research Chart” 6, but the Hive doesn't consist only in this core services (dark blue cells,

Figure 16). It also has other cells like Web Application (collection of client-side components that communicate with i2b2 Cells and allow the investigator to query and display the data of the hive), CRC plug-in (this server-side CRC plug-in calculates patient count breakdown for the children of a given concept), 8 workbench and 2 optional cells, one for the language of processing and another for PFT Processing for a pulmonary specific function.

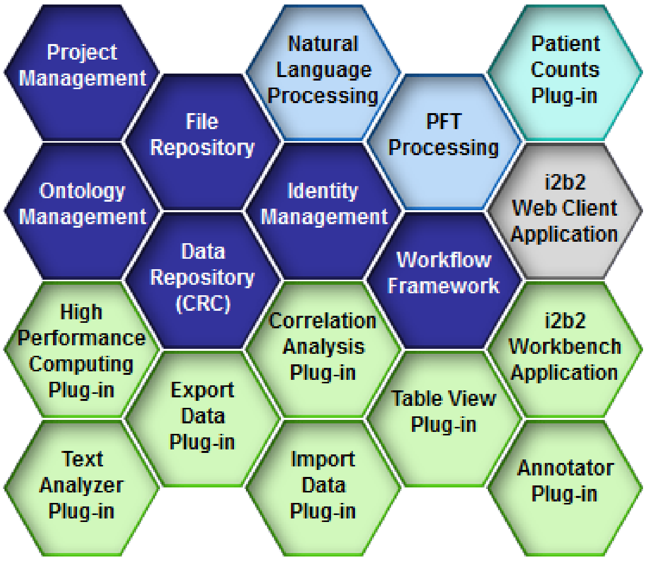


Figure 16 – i2b2 Hive

The cells that have a particular interest for EURECA project are the ontology, cell *Ontology Management service* and the design of the clinical data warehouse, cell *Data Repository (CRC)*.

4.1.1.1 Ontology Management (ONT)

The Ontology Management cell contains concepts and information about relationships between concepts for the entire hive. It also contains the definitions of i2b2 vocabulary. i2b2 data is stored in a relational database, with a star schema format, ie, star schema contains one fact (observation_fact) and many dimension tables (visit_dimension, patient_dimension, concept_dimension and provider_dimension) as discussed in the following section. i2b2 ontology (metadata) is formed by concept codes and the hierarchical structure of these codes together with their descriptive terms. The vocabulary in the ONT cell is organized in a hierarchical way, representing relationships between the "parents" (top levels) and "children" (lower levels) nodes. Note that elements occurring on the same level are known as “siblings”. Figure

[6] https://www.i2b2.org/software/files/PDF/current/HiveIntroduction.pdf

17shows the categories of the ONT cell on the left. In the picture on the right, a number of categories displayed until the nodes, among themselves are *siblings*.

This workbench can be downloaded, as a demo version to your computer, or accessed online at https://www.i2b2.org/webclient/.

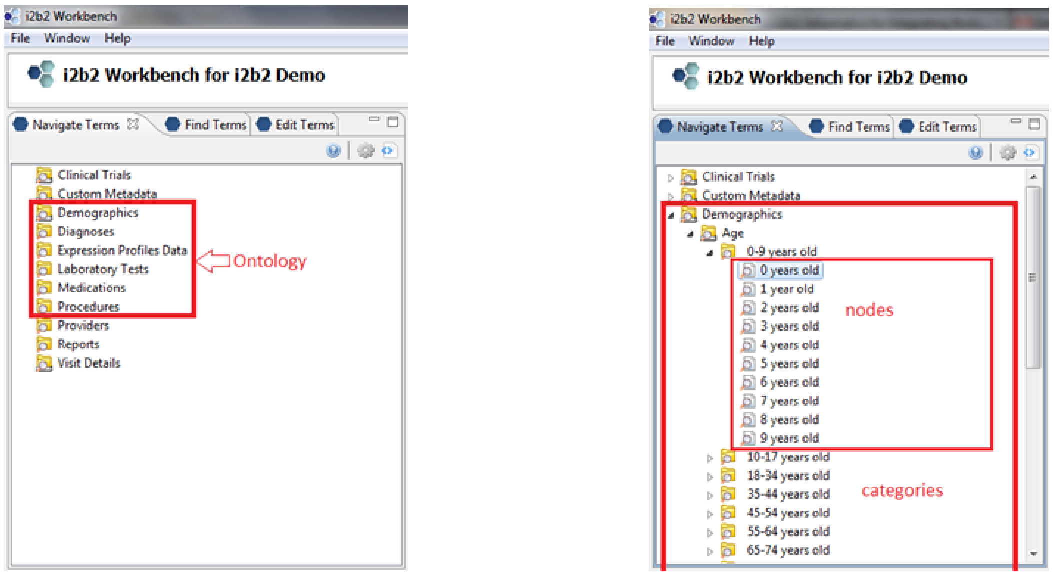


Figure 17 – Ontology tree structure and nodes and categories

It is important to note that *vocabularies in the ONT cell may originate from different sources, and the codes from each source are distinguished from the others by a unique prefix which is appended to the source code. Each distinct vocabulary and its associated codes is called a scheme*[7].

4.1.1.2 Data Repository (CRC)

Data Repository, also known as CRC (Clinical Research Chart), is one of the core cells of the hive. It is designed as a *data warehouse of patient phenotypic and genotypic data that interacts with other cells of the hive to provide information for users*.
Its main requirements are:

1. It should allow fast queries and hold healthcare information from different sources
2. Must be able to easily join to other repositories to form a large one
3. CRC should store objects present in the genomic data

The data in the CRC cell is de-identified, except for encrypted patient notes (notes from hospitals). The CRC relies on the Project Management cell for authentication and on the Ontology cell for metadata management (metadata management module that has the concepts to define the CRC requests).

The CRC provides two services, a setFinder web service, and a PDO (Patient Data Object) web service (see Figure 18[8]).

[7] https://www.i2b2.org/software/files/PDF/current/Ontology_Messaging.pdf
[8] https://www.i2b2.org/software/files/PDF/current/CRC_Architecture.pdf

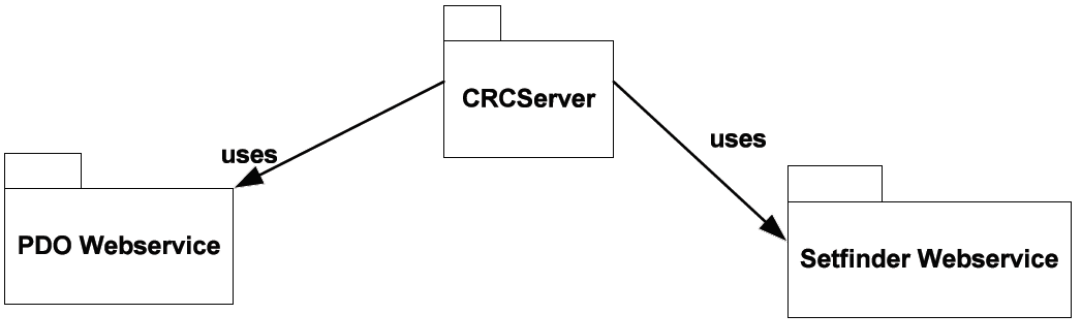


Figure 18 – Part of the Context Diagram

- **Setfinder** manages a user's setFinder queries. The queries are used to create sets of patients that satisfy the specified criteria. The API closely mimics the way the graphical user interface is designed, and setFinder queries are composed of query constraints, a list of panels and its items. The criteria put constraints on concepts from the i2b2 ontology in order to select instances from the data mart.

- The **PDO** web service exposes the clinical data, providing access to patient information such as clinical observations, demographics and provider data.

The CRC cell is designed to keep data from clinical trials, laboratory tests and medical record systems, and many other types of clinical data from different heterogeneous sources.
As it was said before, the CRC is designed as a data warehouse in terms of design. The data mart (subparts of the data warehouse) uses a star schema as information model. This model can be seen in Figure 19. The structure of this model has the observation table as fact table and visits, patient, concept and provider tables as dimensions, which provide additional information about fields in the fact table.

In i2b2 model, each observation has an associated visit (a particular patient encounter), patient, concept and provider instance (referenced with the PK, relation between fact table with the dimension tables). An observation also contains various attributes to store details such as the confidence interval, the begin date and end date of the observation, the units, the value of the observation (measurement), etc. The CRC stores data in 3 different tables: data tables, lookup tables and mapping tables. A short description of each group of tables is presented below[9]:

- Data Tables:

  - *Observation_Fact:* represents the intersection of the dimension tables. In this table, each row describes an observation about a patient made during a visit.

  - *Visit_Dimension*: this table represents sessions where observations were made and stores the events. Each row represents a visit (a session). In a

[9] https://www.i2b2.org/software/files/PDF/current/CRC_Design.pdf

visit, you may have more than one observation. The observations made during the visit of a patient are aggregated using the visit table. The visit record also holds specifics data about the location of the session (hospital or clinic where the session occurred, whether the patient was an inpatient or outpatient, etc)

- *Patient_Dimension*: represents a unique patient in the database and contains patient specific data such as demographics (gender, age, race, etc).

- Lookup Tables:

  - *Concept_Dimension*: in this table, each row represents one concept. Different concepts can be diagnoses, procedures, medications or lab tests. It can also hold any concept type, such as genetic data.

  - *Provider_Dimension*: the provider table represents the physician or provider at the institution.

- Mapping Tables:

  - *Patient_mapping*: maps the i2b2 patient_num to an encrypted number, patient_ide.
  - *Encounter_mapping*: maps the i2b2 encounter_number to an encrypted number, encounter_ide.

PDO (Patient Data Object, the XML representation of patient data) returns data according to this information model.

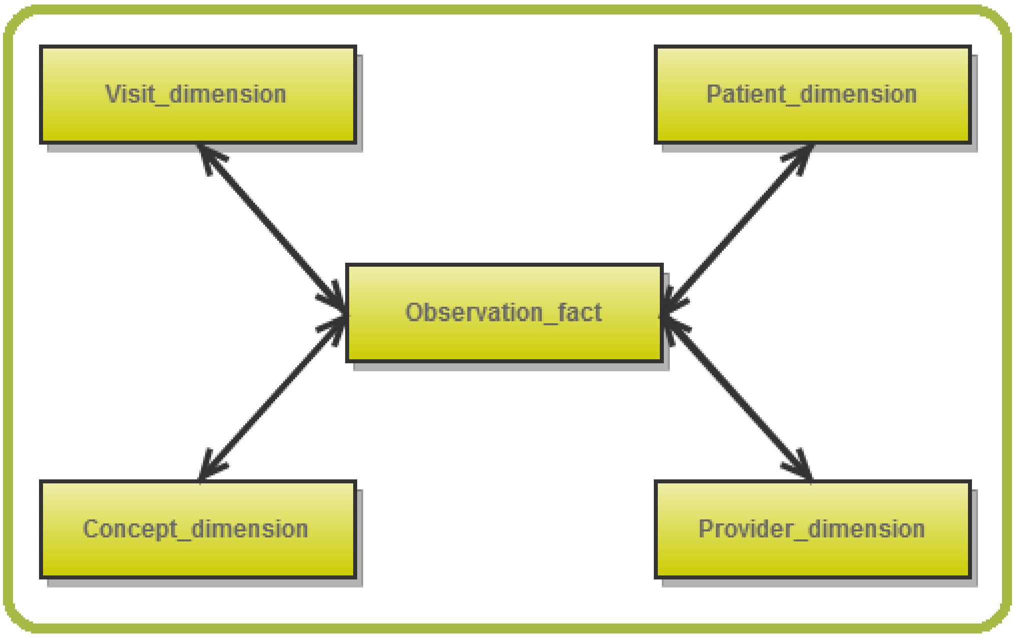


Figure 19 – CRC star schema

4.1.2 OMOP (Observational Medical Outcomes Partnership)

OMOP is a “*public-private partnership designed to help improve the monitoring of drugs for safety. The partnership is conducting a multi-year initiative to research methods that are feasible and useful to analyze existing healthcare databases to identify and evaluate safety and benefit issues of drugs already on the market*” [10].

OMOP provides a public website to inform about the research and keep awareness for consumers, patients, and providers.

The OMOP common data model (V2.0) consists of 2 models, a conceptual data model (which describes the overall development approach) and a logical data model (relational model) with two components:

- Logical Entities and Attributes of the Conceptual Data Model (CDM) Core Module
- Logical Entities and Attributes of the Dictionary

CDM will ensure that research methods can be systematically applied to produce meaningfully comparable results. However, it is not intended to be an integration point for multiple source data sets.

4.1.2.1 Logical Data Model

The logical data model is formed by two modules:

1. the Dictionary component, composed of seven entities (Concept, Concept_Synonym, Concept_Relationship, Concept_Ancestor, Vocabulary_Ref, Source_To_Concept_Map and Relationship_Type) and the
2. CDM Core Module, with thirteen entities (Person, Drug_Exposure, Drug_Era, Drug_Exposure_Ref, Condition_Ocurrence, Condition_Era, Condition_Ocurrence_Ref, Visit_Ocurrence, Procedure_Ocurrence, Proc_Ocurrence_Ref, Observation, Observation_Type_Ref and Observation_Period).

These two components (CDM Core Module and Dictionary) interact with each other to provide a data source mapping.
Figure 20 shows the Logical data model view of the OMOP CDM (CDM Core Module and Dictionary).

[10] http://omop.fnih.org/researchoverview

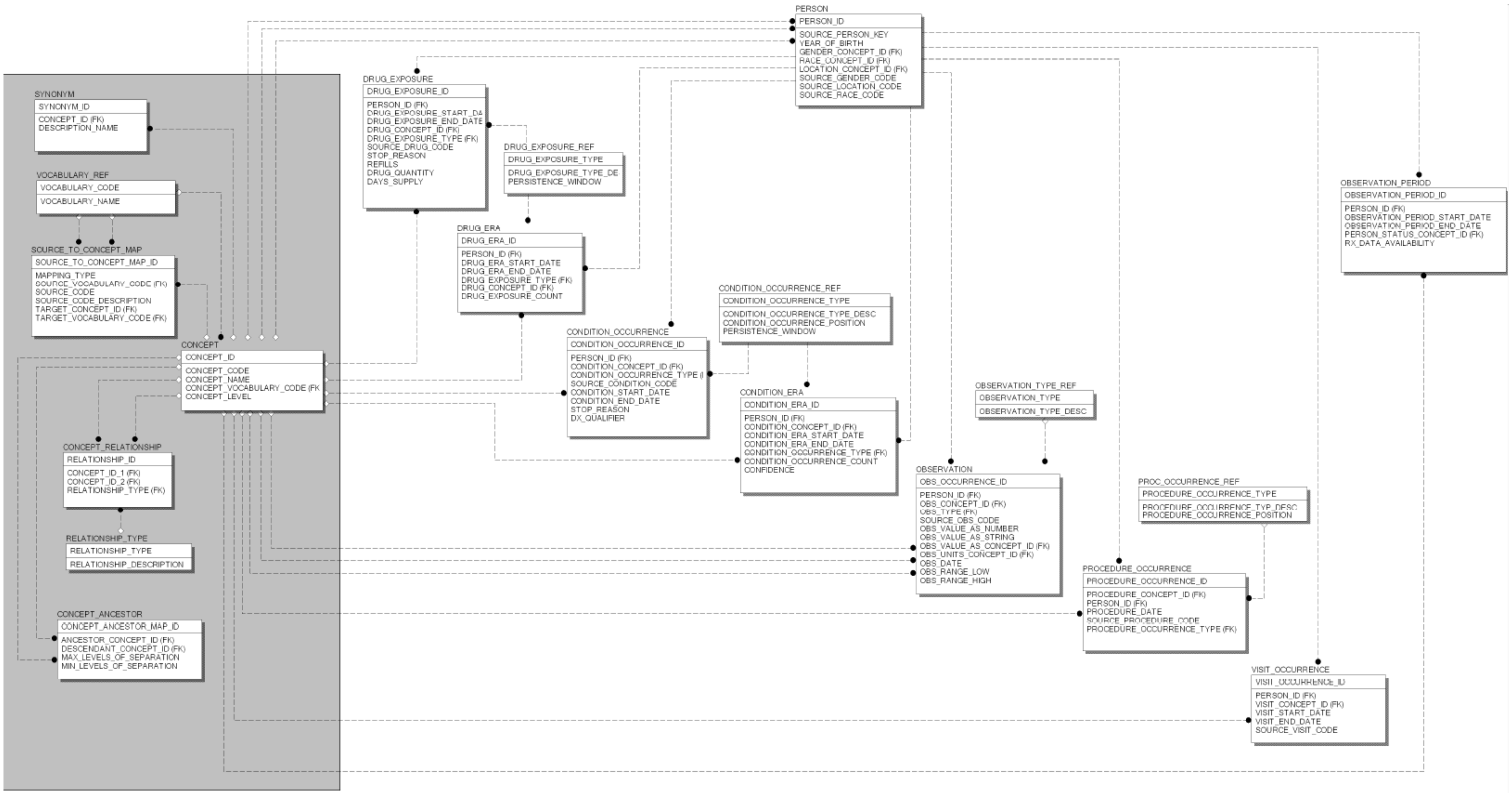


Figure 20 – Logical View of the OMOP Common Data Model

The thirteen tables that compose the CDM Core Module are described below:

- Person:
  Store data from demographics, e.g. gender, race, location, etc. It has a fixed value *person_id*, which is the Primary Key.

- Drug_Exposure:
  "Contains individual records that reflect drug utilization from within the observational source"[11]

- Drug_Era:
  Refers to a time interval in which a person is exposed to a particular drug of a particular strength. Keep in mind that *Drug_Era* and *Drug_Exposure* are not the same. *Drug_Era* is the combination of successive *Drug_Exposure* entities.

- Drug_Exposure_Ref:
  A list of reference codes for the different types of *Drug_Exposure*. It is used to define the data source and the type of representation of the Drug utilization recorded.

- Condition_Ocurrence:
  It records individual instances of *Person* conditions. These conditions are recorded in different data sources in different ways (various levels of standarization).

- Condition_Era:
  It's similar to *Drug_Era*. *Condition_Era* is the combination of successive *Condition_Occurrence* entities.

- Condition_Ocurrence_Ref:
  "*Reflects the indicator(s) from which the Condition_Occurrence was drawn or inferred, and indicates whether a condition (diagnosis) was primary or secondary and the relative positioning within a Person's condition record*" [12].

- Visit_Ocurrence:
  It contains all the visits a *Person* has made to health care providers.

- Procedure_Ocurrence:
  "*Record individual instances of Person procedures extracted from source data*"[13]. These instances are recorded in different data sources in different ways (various levels of standarization).

- Proc_Ocurrence_Ref:
  Reflects the indicator/s from which the *Procedure_Occurrence* was drawn or inferred, and indicates whether a procedure was primary or secondary and the relative positioning within a *Person's* procedure record.

[11] http://75.101.131.161/download/loadfile.php?docname=CDM%20Specification%20V4.0
[12] http://75.101.131.161/download/loadfile.php?docname=CDM%20Specification%20V4.0
[13] http://75.101.131.161/download/loadfile.php?docname=CDM%20Specification%20V4.0

- Observation:
  Contains all general observations (Lab observations from Medical Claims, EHR (Electronic Health Records), Person chiefs, others…).

- Observation_Type_Ref:
  It represents the type of observation to be made in *Person*

- Observation_Period:
  This entity is used to track the status of a *Person* during a period of study. The status can be: Active, Inactive, Obsolete, Deceased, Unknown or Other.

The Dictionary logical data model also has seven tables, listed above, all related to Concept (except Relationship_Type that is related to Concept_Relationship). The connection between the Dictionary and the Core Module is done via this table (Concept_Relationship) that stores the relationship between two concepts.

4.1.3 HL7 RIM

Health Level 7 Reference Information Model (HL7 RIM) is a standard defined by the HL7 organization. The main goal of this organization is to develop international healthcare informatics standards, thus they define themselves as an organization created in order to *“create the best and most widely used standards in healthcare”*[14].

Specifically, the HL7 RIM uses a UML to show a model that is able to contextualize any situation related with the health care services environment, from a fact observed in a patient until the cost of certain treatment.

[14] http://www.hl7.org/about/index.cfm?ref=common

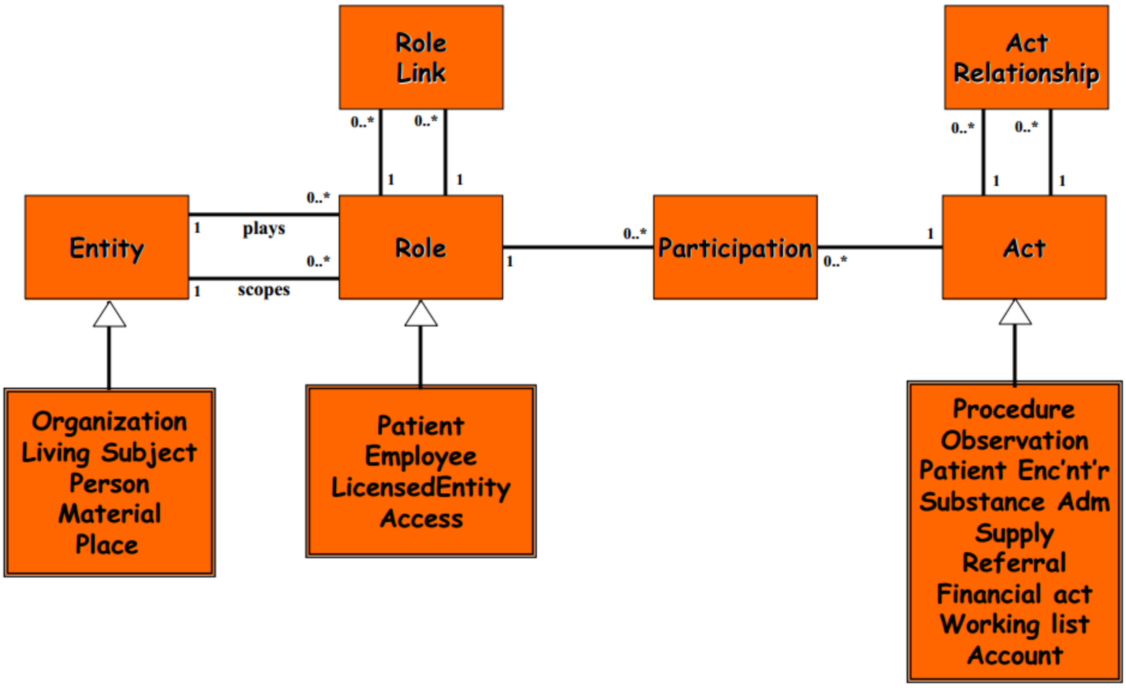


Figure 21 – RIM Core Classes[15]

An overview of the structure of the RIM is shown in Figure 21. As can be appreciated in the figure, the distribution of the RIM is based on 3 main classes:

- Act
- Role
- Entity

The main classes can be specified or generalized by the subclasses shown in Figure 21. The main classes are related to each other through 3 association classes:

- ActRelationship
- Participation
- RoleLink

All the classes are compounded of a set of attributes that give meaning to the classes, allowing the main objective as had been mentioned before: build a standard that will be able to include any medical situation. Some of these attributes are considered as structural attributes. They are:

- Act: classCode, moodCode, negationInd, levelCode.
- Role: classCode, negationInd.
- Entity: classCode, determinerCode.
- ActRelationship: typeCode, contextControlCode.
- Participation: typeCode, contextControlCode.
- RoleLink: typeCode

In Figure 22 the full structure of the HL7 RIM standard is shown. The subclasses of the three main classes are assigned the same colour as the main class.

[15] http://www.hl7.org/documentcenter/public_temp_82FA0ADD-1C23-BA17-0CE002E6D5DB7961/calendarofevents/himss/2009/presentations/Reference%20Information%20Model_Tue.pdf

The main classes (and some of their most important specifications) are defined as follows [1]:

- **ACT:** "*Act is a record of something that has happened or may happen*" [16].It represents any action that has happened, is happening, is requested to happen or is intended to happen. As a comment of usage of this class, just mention two characteristics, (i) any instance of the Act class just represents an action in a point in the time, it cannot be changed. Instead of that, a new instance should be created and then both should be linked using an ActRelationship instance; (ii) acts have participants, which can be actors or targets. Additionally, as it has been mentioned before, there exists some classes that specify the main ones with extra specific attributes, the principal sub-classes of Act are:
    - **Observation** specifies the Act class and is described as "An Act of recognizing and noting information about the subject, and whose immediate and primary outcome (post-condition) is new data about a subject. Observations often involve measurement or other elaborate methods of investigation, but may also be simply assertive statements"[16].
    - **Procedure.** It is an act specialization that includes some form of intervention in a patient, a planned alteration or the manipulation of the structure of the body.
        - **SubstanceAdministration** specifies the sub-class Procedure. It refers to the action of introducing (or other kind of applying) any substance to a subject.
    - **Document.** This sub-class is focuses on stored the specific characteristics of a document associated to a patient.

- **ENTITY** class represents any living or non-living thing that exists or will not exist. Obviously, the common entity instances will be person or even animal. However, this class is designed to store also organisations or other types of things that can participate in an Act. The main sub-classes of Entity are:
    - **LivingSubject** stores the specific characteristics of an Entity instance in the case that it refers to a living subject, organisms or complex animal; e.g. person, dog, microorganism or plants. So, this class can encompass mammals, birds, fishes, bacteria, parasites, fungi and viruses.
        - **Person** is a sub-class of LivingSubject that offers the specific attributes that characterizes a person, such as educational level, marital status or address.
    - **Organization.** This sub-class of Entity represents a formalized group of people with a common purpose and the infrastructure to carry out that purpose.

- **ROLE:** This class establishes the role that an entity plays when it participates in an act.

- **PARTICIPATION** is the "association between an Act and a Role with an Entity playing that Role". As the description says, any Entity that is implied in an Act is linked with it through a Participation instance. The type of

[16] Benson, Tim "Principles of Health Interoperability HL7 and SNOMED". 1st Edition, 2010, Health Informatics Series.

participation that an entity does is specified within the *typeCode* attribute in Participation class. An examples of the kind of the involvement in the action is, e.g. performers of the act: surgeons, observers, practitioners…

- **ACT RELATIONSHIP.** This class establishes the existence of a relation between two different instances of the Act class, specifying the meaning and purpose of the relationship through the attribute *typeCode*.

- **ROLE LINK** connects two different instances of the class Role, and expresses the dependency between them. E.g. employer-employee, doctor-patient, etc.

Following the description of the standards HL7 v3 detailed before, in the European project INTEGRATE a relational database model following the HL7 RIM standard has been created. The INTEGRATE data model contains only those classes (and their attributes) needed for the scenarios of the project. This model has been tested and has been deemed suitable for different types of data from different data sources. Additionally, although the HL7 RIM guarantees that it can represent almost any medical situation, the INTEGRATE HL7 RIM based model is open for minor changes ensuring the possibility to adapt it for new requirements, such as adding new (non-clinical) data sources such as genomic data, or to increase the performance.

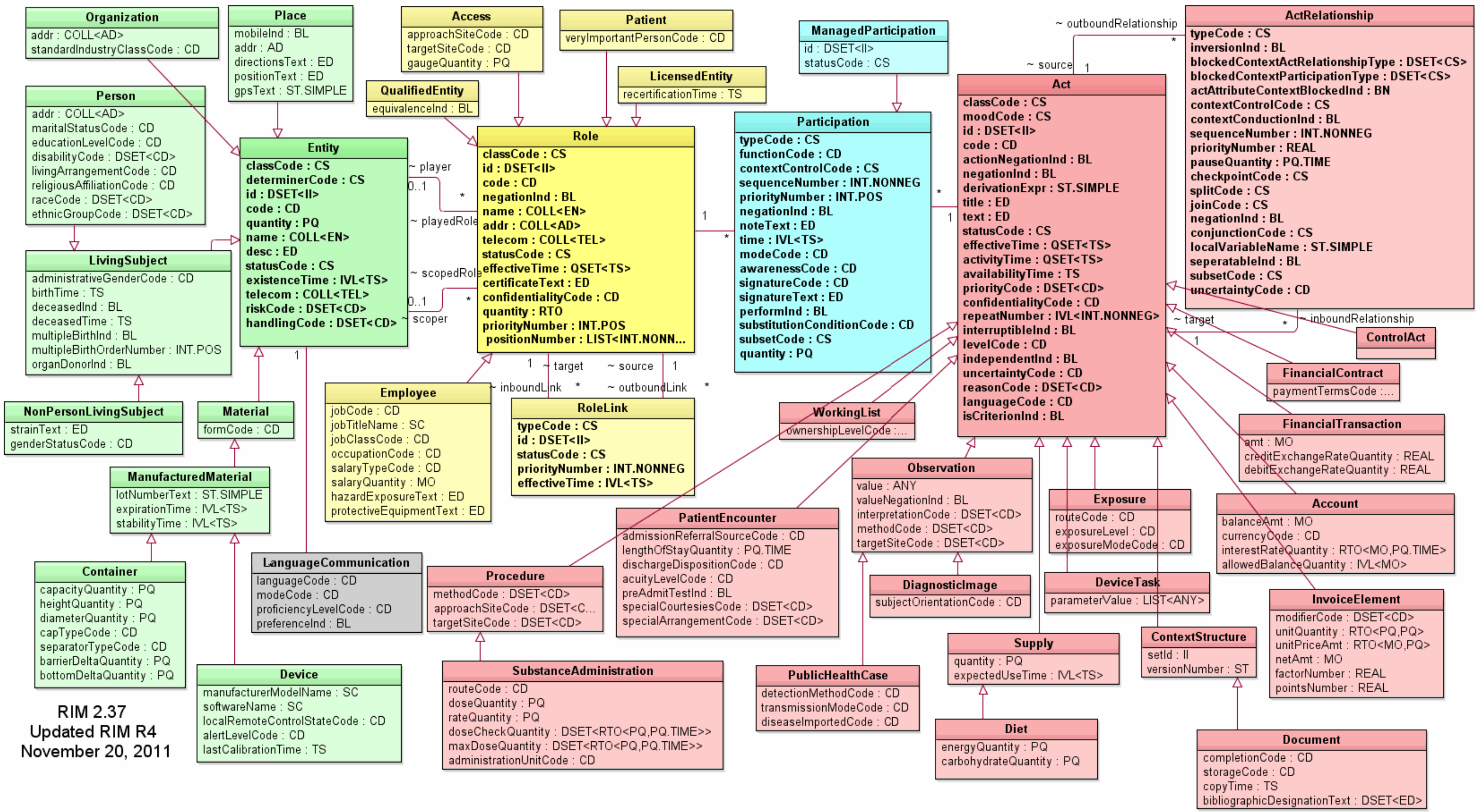


Figure 22 – HL7 RIM backbone clas

4.1.4 HDOT (p-medicine)

The Health Data Ontology Trunk (HDOT) [2] is conceived as a modular middle-layer ontology (ontology with domain-driven classes but application-independent) as it specifies upper-level classes from the biomedical domain while keeping a very general semantic structure that can be further developed by the introduction of several modules with different specializations.

HDOT is broad enough to contain all general classes under which any more specific semantic content can be subsumed allowing enriching the semantic content of different resources, e.g. integrating parts of biomedical terminologies like SNOMED CT, NCI Thesaurus, ICD...

HDOT is currently being developed in the P-medicine [4] project. As it hasn’t been finished it was discarded for the time being.

4.1.5 BRIDG Model

The Biomedical Research Integrated Domain Group (BRIDG) Model [5] is an information model, specifically an instance of a Domain Analysis Model (DAM), representing a shared view of the concepts of protocol-driven clinical research. This structured information model is being used to support development of data interchange standards and technology solutions that will enable semantic (meaning-based) interoperability within the biomedical/clinical research arena and between research and the healthcare arena.

This model has been developed in conjunction with several stakeholders including the Clinical Data Interchange Standards Consortium (CDISC) [6], the HL7 Regulated Clinical Research Information Management Technical Committee (RCRIM) Work Group, the US National Cancer Institute (NCI), and the US Food and Drug Administration (FDA).

It also provides a mapping to represent the information of this model as HL7 RIM (HL7 is one of the developers). Due to its complexity and specificity and the fact that most information can be represented in HL7 RIM it was decided not to use this model. The following UML diagram shows the extent of this model.

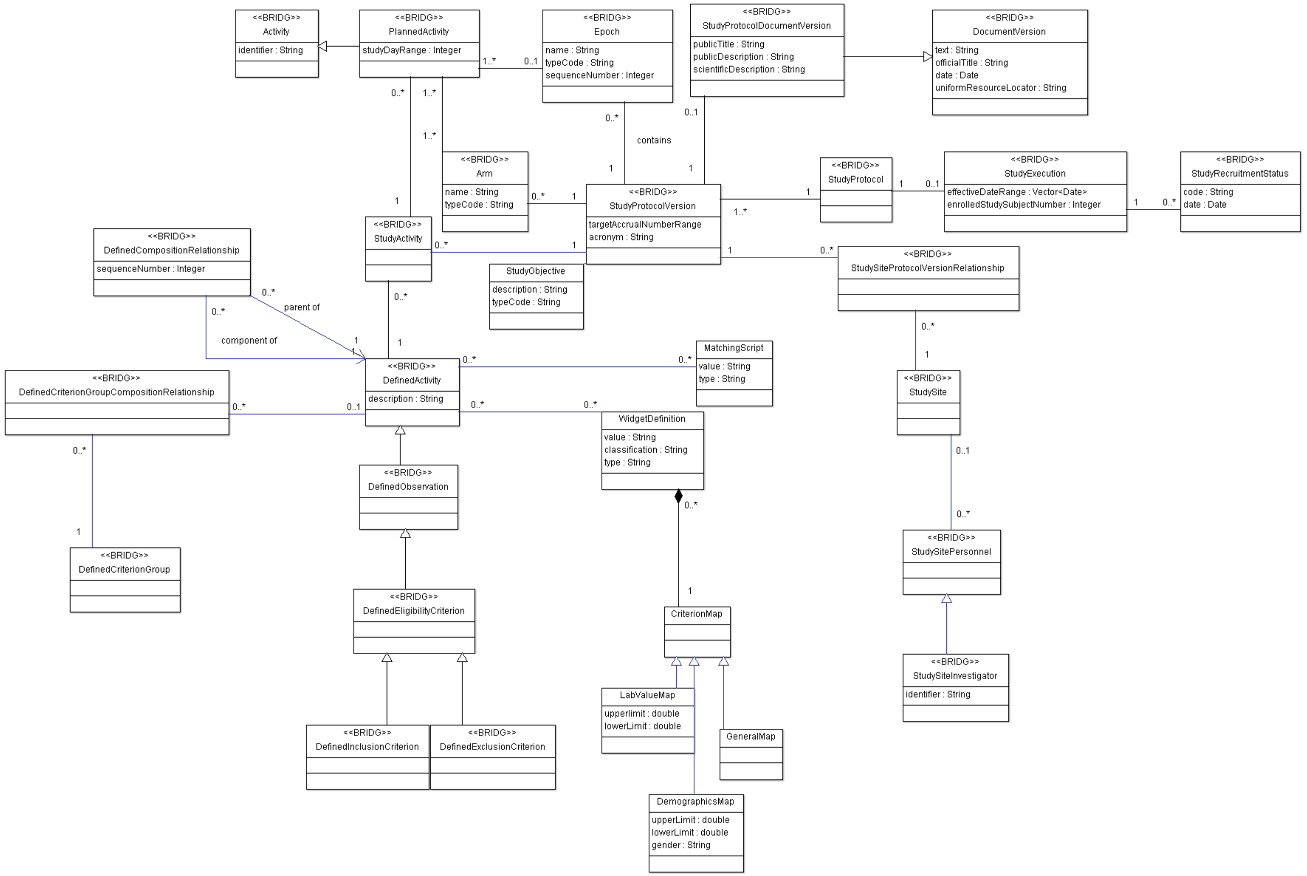


Figure 23 – BRIDG Model UML Diagram [7]

4.1.6 openEHR Reference Model

The openEHR Foundation is defined as a virtual community that aims to provide interoperability and computability in e-health, focused in electronic health records (EHRs) and systems.

The openEHR Foundation provide a set of specifications defining a health information reference model. In openEHR, archetypes are the keystone of the openEHR architecture. An archetype is a definition of a pattern used for capturing clinical data and storing patient data into openEHR Reference Model.

There are 4 main categories of archetypes: Composition, Section, Entry and Cluster; each of these classes defines a part of the openEHR Reference Model[17]

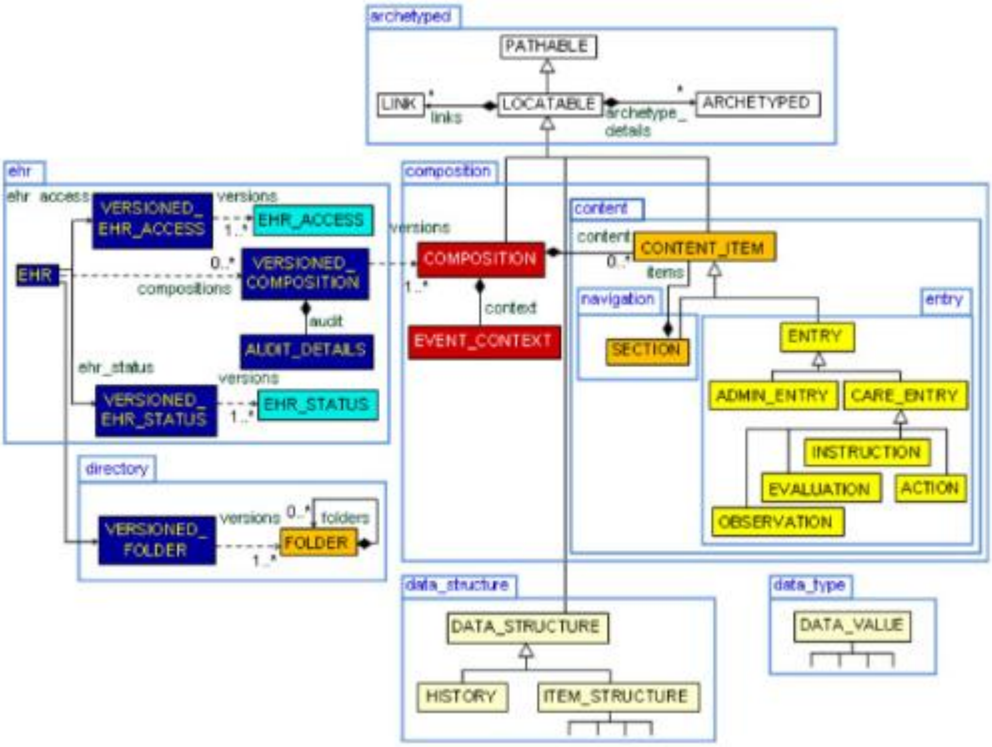


Figure 24 – simplified openEHR Reference Model

4.1.7 CDISC2RDF

The goal of the CDISC2RDF[18] consortium is to “make CDISC[6] standards linkable, computable, and queryable”. CDISC has established standards which are widely used to support the acquisition, exchange, submission and archive of clinical research data and metadata. For this reason, creating tooling that is based on an authoritative representation in RDF of CDISC standards would increase interoperability among systems that employ those tools. The first deliverable of CDISC2RDF, published January 20, 2013, includes SDTM 1.2, Implementation Guideline (IG) 3.1.2 and

[17] http://www.openehr.org/wiki/display/healthmod/Introduction+to+Archetypes+and+Archetype+classes
[18] http://cdisc2rdf.com/

Controlled Terminology (CT), plus CT for data capture standards (CDASH) and analysis standards (ADaM).

The NCI EVS team that provides the official version of CDISC terminology[19] has confirmed plans to publish an RDF/OWL version with the April 2013 publication, building on the CDISC2RDF deliverable to provide a linkable, computable, and queryable version in addition to the existing pdf, html, and odm/xml versions of CDISC terminology. The CDISC2RDF team members will continue to work on OWL representations of the CDASH and ADaM data standards for discussion and application with partners and stakeholders.

[19] http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/cdisc

4.2 EURECA Common Data Model

As has been mentioned in Section 4.1.3, in the European project INTEGRATE a HL7 RIM is being used, based on the Common Data Model (CDM). The INTEGRATE CDM is a relational model that has been created following the HL7 RIM specifications. As it has been demonstrated to be effective in INTEGRATE project, it has been decided to reuse the scheme, adapting it to the needs of the EURECA project.

4.2.1 Schema

Due to the wideness of the RIM, only some of its classes have been modelled for the EURECA project. Selected classes have been those which are used in the shared clinical data of breast cancer clinical trials. Similarly, not all attributes of each class in the RIM are needed, so only a subset of them is included in the relational model of the CDM.

Some attributes have been simplified in the relational model compared with those defined by HL7 v3 standard. The classes and relationships from the RIM in the EURECA project are:

- *Act*, with the sub-classes *Observation*, *Procedure*, *SubstanceAdministration*, and *Exposure*.
- *Role*
- *Entity*, with the sub-classes *LivingSubject*, *Person*, and *Device*

Also two main relationships classes from the RIM have been implemented: *ActRelationship* and *Participation*.
In addition, the relational model has some tables that do not correspond to a class in the RIM. These additional tables are:

- *ActProcedureApproachSiteCode*
- *ActMethodCode*
- *ActTargetSiteCode*
- *ActObservationInterpretationCode*
- *ActObservationValues*

Figure 25 shows the implemented schema.

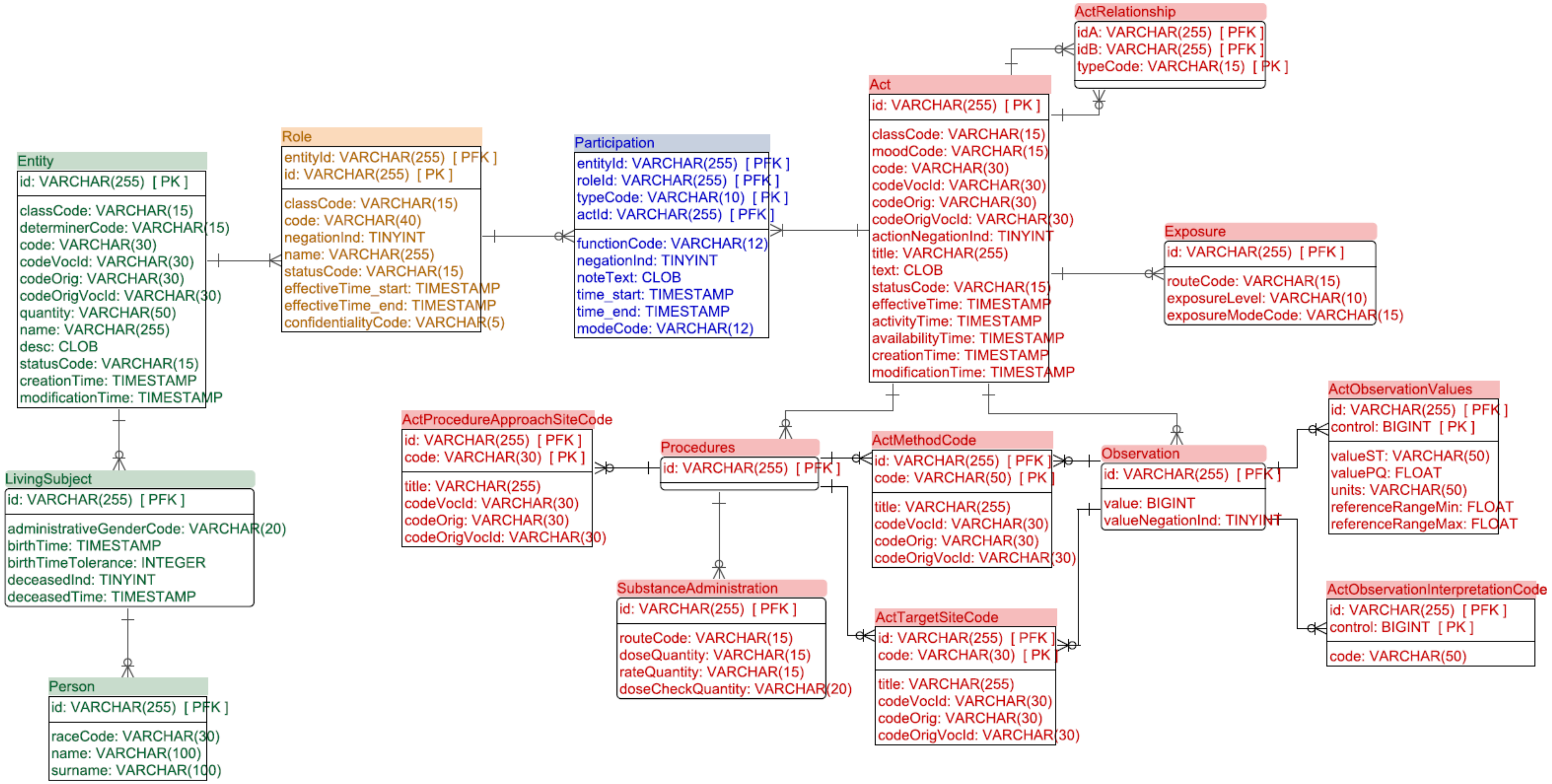


Figure 25 – HL7 RIM-based relational model schema

4.2.2 Database Views

A view is a stored query that returns a result set. It is a "virtual table", it appears like any other table of the schema, but it really isn't.

We have created a set of views because in most queries we need to retrieve information accessing to the same set of tables. By using these views, it is easier to query the data model because it is not needed to write joins between tables in the query itself, as the view handles it internally.

Listed below are the different views and the tables that compose them:

- *Living Subject view:*
    - Entity
    - Living Subject
- *Person view:*
    - Entity
    - Living Subject
    - Person
- *Procedure view:*
    - Act
    - ActMethodCode
    - ActTargetSiteCode
    - ActProcedureApproachSiteCode
    - Procedures
- *Substance Administration view:*
    - Act
    - ActMethodCode
    - ActTargetSiteCode
    - ActProcedureApproachSiteCode
    - Procedures
    - Substance Administration
- *Observation view:*
    - Act
    - ActMethodCode
    - ActTargetSiteCode
    - ActObservationValues
    - ActObservationInterpretationCode
    - Observation

In Figure 26 the EURECA schema with the views is represented.

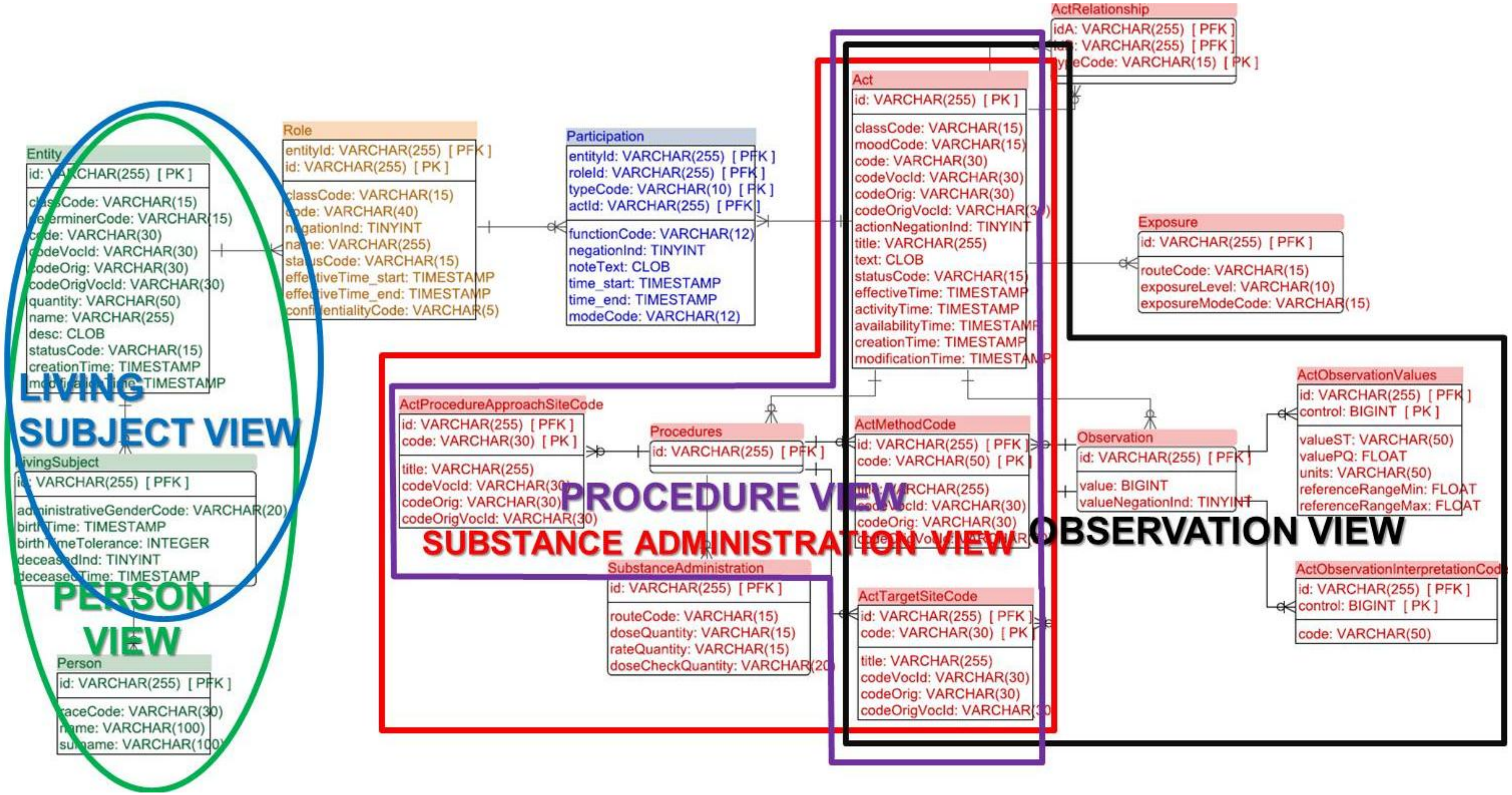


Figure 26 – EURECA schema with the views

4.3 EURECA Core dataset

“The semantic core dataset is an essential prerequisite to machine processable access to both EHR and Clinical trial data. Concepts in the dataset will have their unique identifiers, well understood meaning as well as a set of synonyms they can be referred as.”[20c]

Medical terms and vocabularies composing the Core Dataset in the EURECA project are going to be described in detail in WP4. The Core Dataset will be a compound from subsets from well-established bio-medical vocabularies and ontologies – such as NCI-thesaurus, HGNC, SNOMED-CT, etcetera – that could describe completely all concepts present in defined clinical scenarios and use cases. The Core Dataset will be used as common lingua to disambiguate those concepts that could be represented using different terms in different data sources with diverse vocabularies.

4.3.1 TermBinding

The main purpose of a data model is storing all available data of a given domain. Although, storing concepts from terminologies (developed outside the model) within a standard data model usually entails a set of constraints that depend on the limitations of the data model itself to represent all the information that these terminologies can express.

An expressive terminology as SNOMED CT, that represents similar concepts which are grammatically, linguistically and semantically distinct, requires a set of rules that simplify the decisions about which group of similar concepts are appropriate to a field of the data model.

An example of these constraints can be seen in the field negationInd of the Act table of HL7 RIM. This field specifies if the act represents a negation of the act itself. i. e. “Not” Currently pregnant, an Act will be created representing Currently pregnant and the negation information will be stored in the field negationInd of that Act. However, a constraint is indicated stating that this field shouldn’t be used in case the finding or procedure context already is encoded in domain vocabulary.

Concretely, HL7 provides a set of recommendations specifying unambiguously where which concepts of a terminology should be stored in which field of the RIM model.. This in turn allows querying the different classes present in the model, knowing where each piece of data must have been stored.

[20] EURECA DoW.

5 CONCLUSIONS

This document has provided an overview of data schemas of the systems and terminologies used at the clinical sites. For a few scenarios, an example was provided of which data is needed to execute the scenario. For this data, it was indicated in which of the systems it can be found and what the structure of the data is. Next, a number of common information models in the clinical domain were described, leading up to a description of the common data model that was developed for the EURECA project.

Next steps include the construction of mappings from the clinical sites' systems data models to the EURECA common data model. This will enable execution of the scenarios on the common data model, populated with existing (deidentified) data from the clinical sites.

6 REFERENCES

1. Benson, T. *"Principles of Health Interoperability HL7 and SNOMED"*. 1st Edition, 2010, Health Informatics Series
2. HDOT Code Page. http://code.google.com/p/hdot/. February 2013.
3. Sanfilippo, Emilio M.; Schwarz, Ulf; Schneider, Luc; "The Health Data Ontology Trunk (HDOT). Towards an ontological representation of cancer-related knowledge," Advanced Research Workshop on In Silico Oncology and Cancer Investigation - The TUMOR Project Workshop (IARWISOCI), 2012 5th International , vol., no., pp.1-4, 22-23 Oct. 2012. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6397178&isnumber=6397169 .
4. P-Medicine Homepage. http://www.p-medicine.eu/. September 2012.
5. BRIDG Model Homepage. http://www.bridgmodel.org/. September 2012.
6. CDISC Homepage. http://www.cdisc.org/. February 2013.
7. BRIDG Model download page. http://bridgmodel.nci.nih.gov/files/BRIDG_Release_3.2_Package.zip . To be viewed with the SPARX System Enterprise Architect Viewer.
8. EURECA Deliverable: D9.1 Report on the development environment and on the available test data, January 2013.