BD2Decide

Big Data and models for personalized Head and Neck Cancer Decision support

TITLE	**Multilayer data acquisition and management services**		
Deliverable No.	D5.1		
EDITOR	V. Tountopoulos (ATC)		
Contributors	L. Kallipolitis (ATC), E. Mantziou (ATC), Alessio Fioravanti (UPM), Liss Hernández González (UPM), Peter van de Ven (VUMC), Florian Jung (Fraunhofer)		
WorkPackage No.	WP5	**WorkPackage Title**	Visualizing the HNC Virtual Patient
Status[1]	Final	**Version No.**	1.0
Dissemination level	PU		
DOCUMENT ID	D5.1 Multilayer data acquisition and management services		
FILE ID	BD2Decide D5.1		
Related documents	Technical Annex I DoA version 21/10/2015		

Distribution List

Organization	Name of recipients
AOP	T. Poli, E.M. Silini, S. Rossi, E. Martinelli, G. Chiari, C. Caminiti, G. Maglietta
VUMC	R. H. Brakenhoff, H. Berkhof, P. va de Ven
UDUS	K. Scheckenbach
INT	F. Favales, L. Licitra, E. Montini, G. Calareso, G. Gatta, A. Trama
UPM	G. Fico, A. Fioravanti, L. Hernández González
Fraunhofer	F. Jung, S. Wesarg
ATC	V. Tountopoulos, L. Kallipolitis, E. Mantziou

[1] Status values: TOC, DRAFT, FINAL

AII	A. Algom, R. Yahud
POLIMI	L. Mainardi
MAASTRO	P. Lambin, F. Hoebers
MME	S. Copelli, F. Mercalli
UNIPR	S. Rossi
European Commission	Project Officer: and all concerned E.C. appointed personnel and external experts

Revision History

Revision no.	Date of Issue	Author(s)	Brief Description of Change
0.1	18.05.2016	V. Tountopoulos	ToC
0.2	15.06.2016	V. Tountopoulos, L. Kallipolitis	Input to section 2 for data requirements
0.3	23.06.2016	V. Tountopoulos, E. Mantziou	Draft version of section 3
0.4	24.06.2016	A. Fioravanti, L.H. González	Input to sections 2.3 and 4
0.5	27.06.2016	V. Tountopoulos, E. Mantziou, F. Jung, P. van de Ven	Draft version of section 4
0.6	29.06.2016	G. Fico	Review of complete version
1.0	30.06.2016	V. Tountopoulos, E. Mantziou	Final version revised by coordinator

Addressees of this document

This document is addressed to the BD2Decide Consortium and describes the data management layer of the BD2Decide environment. More specifically, it builds upon the user needs and presents the data requirements with respect to the collection, analysis, management and visualisation of the BD2Decide information.

The deliverable elaborates on the architecture of the data warehouse, which integrates the envisaged data repositories for hosting information about the health record of the patients, as this is progressively built during the diagnosis, the treatment and the follow-up phases of an HNC situation, along with prognostic prediction model data and statistical population data, which is required to assess different HNC practices, such as to predict the effect of a treatment in the survival rate of a patient.

The result of this deliverable is an initial implementation of the data acquisition infrastructure, in order to enable the collection of the patients' health records for retrospective studies, and the realisation of the data warehouse environment, on which the BD2Decide clinical decision support system will be implemented.

This document will be delivered to the European Commission.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abbreviations and definitions

Abbreviation	Definition
CDSS	Clinical Decision Support System
CIS	Clinical Information System
e-CRF	Electronic Clinical Record Format
EHR	Electronic Health Record
HIS	Hospital Information System
HNC	Head and Neck Cancer
ID	Identification Number
IdM	Identity Management
LIS	Laboratory Information Systems
PACS	Picture Archiving and Communication Systems
PDS	Patients Documentation System
QoL	Quality of Life
RDBMSes	Relational Database Management Systems
XDS	Cross-Enterprise Document Sharing

Executive Summary

The BD2Decide aims to develop a Clinical Decision Support System (CDSS) that integrates big data and prognostic models analysis techniques to improve the decision making process in the treatment of Head and Neck cancer. This deliverable is part of the WP5 work on visualisation and emphasises on the architecture of the data warehouse environment in the project.

More specifically, this document is the first deliverable of WP5 and describes the services for the acquisition and management of the data information involved in the provision of personalised decision making in BD2Decide. The deliverable builds on the specification of the user needs and relevant use cases in D2.1 and describes the datasets involved across the implementation of the workflow processes from the diagnosis of a head and neck incident, through the decision on the personalised treatment method to the assessment of the follow-up period. The involved datasets are presented in four categories, depending on the layer of the BD2Decide platform environment that the specific data is produced. As such, we define: i) *collection datasets*, which are collected from the information systems of the clinical centres about the electronic health records of the patients or from Internet sources, which provide population-based data, ii) *analysis datasets*, which are produced within the BD2Decide environment, as a result of the processing algorithms built (or deployed) within the project scope, iii) *visualisation datasets*, which complement the previous categories in order to enhance the visualisation capabilities of the BD2Decide environment, and iv) *access control datasets*, which refer to the identity management processes of the BD2decide platform environment.

Furthermore, the deliverable presents the BD2Decide data warehouse service architecture, which aggregates the required data repositories for hosting the type of data identified in the above mentioned categories. Thus, this architecture implements the data requirements, considering the technical use cases that were introduced in D2.1. The architecture defines three main parts. The first one refers to the acquisition of data and the respective services from the side of the clinical centres. The second part of the architecture presents the BD2Decide storage layer, while the third part extends to meet the data hosting requirements in the envisaged BD2Decide big data infrastructure. For each part of the architecture and for each data category, we, also, define the expected functionalities, which result in the implementation of work flow activities for the collection, analysis, management and visualisation of the BD2Decide datasets.

Finally, the document provides an initial design of the structure of the repositories comprising the BD2Decide data warehouse, giving emphasis on the electronic health record for retrospective and prospective studies, the patients' documentation system and the BD2Decide Knowledge Base. For each repository, a description of the expected role in the BD2Decide environment is provided, along with details on the technologies to be used for the implementation of the databases and the respective management services.

1 ABOUT THIS DOCUMENT

1.1 Introduction and scope

This deliverable presents the results of Task 5.1 of the BD2Decide work plan for the specification of the patient's data warehouse. The document focuses on the requirements for the collection of multiscale and multivariable data for the patients of the clinical centres from different information sources and the analysis of this data to support the implementation of the personalised decision support system for head and neck cancer incidents. To this end, the document aims to deliver the baseline prototype work for the collection of the health related patients' information in an anonymised/encoded way and set the grounds for the specifications of the required data repositories, which harvest these health records and implement the expected BD2Decide processes.

1.2 Structure of the deliverable

In order to implement the set objectives, this document is structured as follows:

- Section 2 makes an overview of the data requirements in the BD2Decide project by analysing the findings of the Deliverable D2.1. It, then, presents the categories of the required datasets in the project and the sources from which is data is retrieved.
- Section 3 describes the architecture of the BD2Decide data warehouse. It presents the logical structure of the data environment and provides a realisation of the expected functionalities to be implemented in the project in order to execute the use cases, reported in D2.1.
- Section 4 elaborates on the details of the repositories defined in the previous section the BD2Decide data warehouse architecture, by providing an initial scheme of the repositories and the candidate technologies and tools, through which these repositories will be developed in the project.
- Finally, Section 5 concludes on the expected role of this document in the project work plan.

2 DATA REQUIREMENTS

2.1 Review of data needs in BD2Decide

BD2Decide provides a solution to introduce new knowledge and improve the decision making process in the treatment of Head and Neck Cancer (HNC) incidents, through supporting new tools and evidence. The steps that realise the approach developed in the project have been designed in the specification of the user needs and relevant use cases in the BD2Decide deliverable D2.1 [1]. As it can be seen there, the workflow is evolved around three main workflow phases, namely diagnosis, treatment and follow-up. Each workflow phase may be further divided to steps (see Figure 1), like the diagnosis phase, which involves the decision making process, such as the criteria to support decision making and the resulting treatment decision, and the engagement of patients in the treatment process.

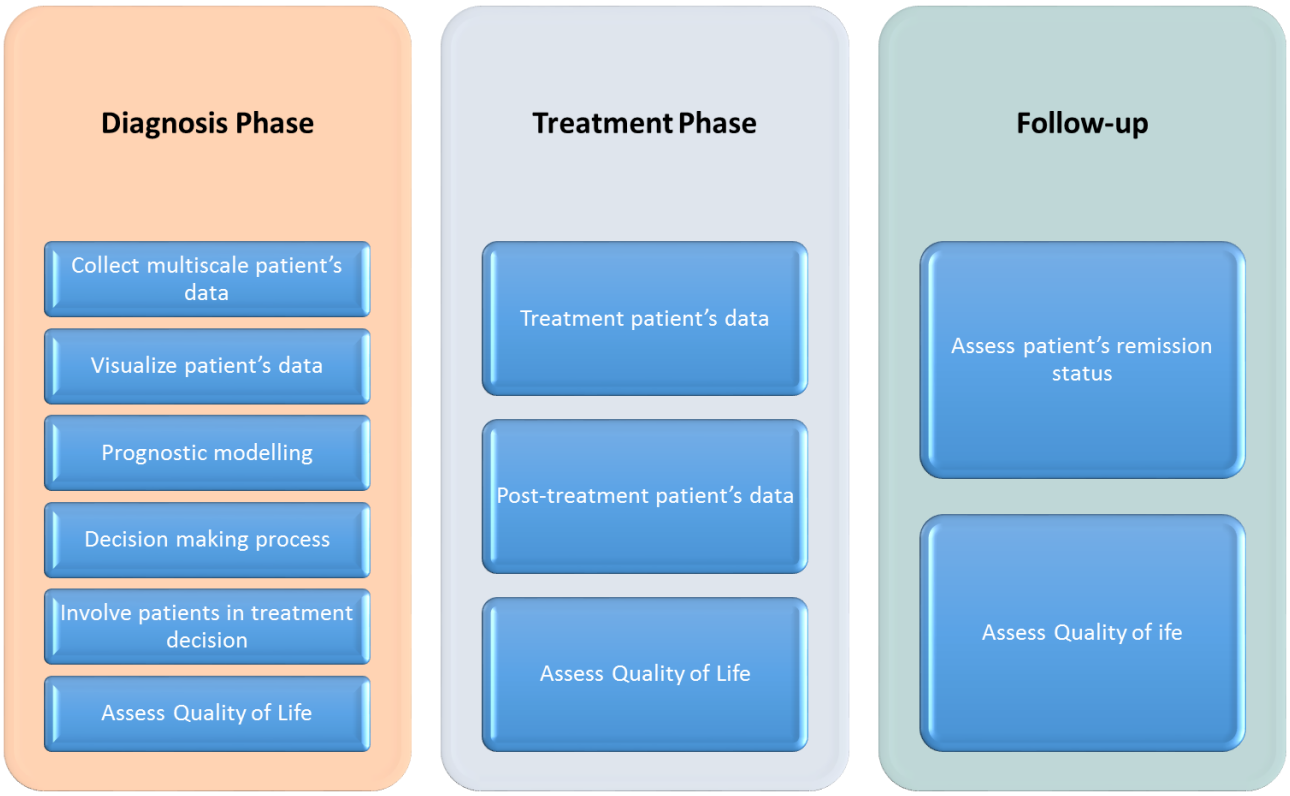


Figure 1: The workflow for the HNC process in the BD2Decide project.

In each phase and step, a set of data is involved, which facilitate the accomplishment of the machine functionalities envisaged in the BD2Decide platform. The involvement of data may be triggered either in a manual (user driven) or an automatic (machine driven) mode.

In the diagnosis phase, when a head and neck tumor is suspected, a set of clinical, radiological and genetic data should be collected and recorded. This data is necessary to provide the baseline information for compiling the electronic health record (EHR) of the patient and subsequently tracking their status along the treatment workflow process, shown in Figure 1.

As the health professionals run the steps in the diagnosis phase, they have to combine the clinical, radiological and genetic data with population data, which refer to the epidemiologic analysis of the

population in the specific region or geographical area, the behaviour analysis or the population, the use of medication, etc. Such population data requires a process within BD2Decide, which will enable collecting this data from existing Internet sources and associating it with the patients' data. The combination of all this kind of information results in the prognostic prediction indicators for each patient, based on the involved prognostic models. Thus, the information collected from Internet sources, extracted from the examination of a patient and calculated through the BD2Decide components comprises the decision making criteria, which will be the evidence to support the discussions within the tumor board and among the professionals and the patients on the most suitable treatment.

Moving across the steps of the diagnosis phase, the treatment decision process requires the retrieval of already defined datasets in the previous processes, like the patient's data and the potential treatment scenarios, to constantly monitor the patient's recovery curve. In this case, the BD2Decide platform requires a process to invoke the results of the previous steps and calculate the effect of the treatment process on the specific patient (through, for example, quality of life - QoL assessment) and the population (s)he represents.

As part of the diagnosis phase and in order for the patient to be prepared for making decisions on the treatment to be finally adopted, with the full support of their health physicians, BD2Decide introduces an intermediate patient personalised co-decision aid step, in which a single patient can be guided through information concerning the specific tumor suspected on them, the different treatment options and the advantages and disadvantages of each treatment option. This step requires a multimedia dataset that visually explains the expected treatment results and side effects to the patient. Since this process aims to allow patients in participating in co-decision process for the treatment option, this step may, also, require that the patient's preferences are recorded for future reference.

In the treatment phase, the main data process involved in the BD2Decide platform is the need for updating the values of the data collected or calculated in the previous phase. In that case, the steps involved in this phase introduce the need for synchronising datasets with their latest values (at the data collection point, i.e. the hospital information system) and updating datasets by executing the algorithmic processes (i.e. prognostic prediction analysis) being implemented in BD2Decide.

Finally, in the follow-up phase, BD2Decide aims to optimise the individual follow up schedule per patient and maintain a history of the quality of life assessment. Such data requires the implementation of the follow-up processes from BD2Decide, which allow the value calculation of such data (e.g. expected next visit time frame, overall QoL assessment by the end of the treatment iteration, etc.).

2.2 Analysis of data involved

Based on the requirements for the involvement of specific datasets in the execution of the BD2Decide workflow, as it was presented in the previous Section 2.1, this section analyses these datasets and provides a description of how they are involved in BD2Decide use cases. For the presentation of this data, we introduce four categories, which differentiate from one another, according to the layer of the BD2Decide platform environment that the specific data is produced.

The BD2Decide platform environment is defined as the set of system and platform components, which are developed in the course of the BD2Decide project, controlled by the BD2Decide Consortium partners and may reside in or outside the premises of the clinical centres.

These categories are the following:

- *Collection data*: in this category, we include all the datasets which are produced outside the BD2decide platform environment, like the Hospital Information System (HIS), Laboratory Information Systems (LIS), and Picture Archiving and Communication Systems (PACS) of the participating clinical centres. These datasets need to be collected inside the BD2Decide platform environment, subject to certain data protection and privacy restrictions.
- *Analysis data*: in this category, we classify all the datasets, which result from the execution of the algorithms built in the BD2Decide project (or adopted in it) and deployed in BD2Decide platform environment. The analysis data originate from the processing of the collection data for the certain purposes defined in the BD2Decide project.
- *Visualisation data*: in this category, the datasets required in the visualisation tasks of the BD2Decide platform environment are included. Such data has not been produced in the analysis phase, but they are necessary for visualisation purposes.
- *Access control data*: in this category, we include the datasets, which are required to uniquely identify a persona along the BD2Decide platform environment, including the set of rules that define the persona's privileges on accessing the data of the other categories.

The rest of this section is devoted to the presentation of the datasets involved in these categories. The analysis of the datasets in performed in the following chapters 3 and 4.

2.2.1 Collection data

The *collection data* category integrates all the datasets, which are produced outside the BD2Decide environment and are gathered within it to facilitate the implementation of specific use cases for the target stakeholders. Such datasets may be collected from the information systems that already exist out there in the clinical and research field or the Internet itself. Therefore, in this section, we present the type of the collection data involved in the BD2Decide environment.

The primary information in this category is the data involved in the patients' clinical studies for the retrospective and prospective cases. This data is the electronic health record of a patient and takes the form of an electronic Clinical Record Format (e-CRF), which is widely used in the clinical studies to describe the set of information that needs to be collected from the patients, in order to analyse their clinical status and investigate more on their prospective disease evolution. The fields comprising the patient health record may vary depending on the exact case that we investigate. For the BD2Decide purposes, a first approach for this record has been described in Deliverable D2.1 [1] and integrates the following set of data types:

- The eligibility criteria for the selection of a patient in a study, such as the HNC stage.
- The clinical and demographic data of the patient. For the demographic data, the health record shall track information, like the gender, the age, the time of diagnosis, the ethnicity, etc. For the clinical data, the record gathers the HB level, the PLT level, the value of lymphocytes, etc.

- The risk factor data, such as smoking and alcohol habits and the overall hygiene view of the patient.
- The Clinical T- and N-Characteristics, which describe the morphology of the tumor and the nodes.
- The available multimedia (i.e. images) visualising the problem, such as CT, MRI and DWI-MRI scans.
- The dataset describing the pathology status of the patient.
- The dataset describing information about chemotherapy and radiotherapy.
- The information of potential surgery treatment followed.
- The description of potential tissue samples available, which will be subject to genomic analysis.
- The follow-up dataset, describing the procedures in the follow-up period.
- The toxicity dataset, which describes the impact of radiotherapy or chemotherapy on the patients' clinical status.

Along with this data, the health record is composed of the quality of life questionnaires, which have to be collected along the execution of the retrospective and prospective studies. The structure of these questionnaires is standardised in the bibliography and will be collected from sources outside of the BD2Decide project (and subject to the needs of the clinical partners).

In the collection data category, we also include the population datasets, which are collected from the Web. Such statistical datasets include information about epidemiology data, lifestyle behaviour, such as nutrition habits, alcoholic consumption, smoking habits, statistical data on the health status of the population, and data regarding the use of medication per location.

2.2.2 Analysis data

In the *analysis data* category, we emphasise on the datasets, which are produced in the course of the BD2Decide project, either through thorough investigation of best practices and recommendations or as a result of the application of the BD2Decide technical components on the collection data to exact assessment, based in the research conducted within the project. Therefore, in this section, we present the type of the analysis data generated within the BD2Decide environment.

The patients' electronic health record introduced in the previous section 2.2.1 is further completed through the analysis data. More specifically, the e-CRF includes the following analysis data:

- The imaging analysis datasets, which are produced from the BD2Decide tools, responsible for the image anatomic features extraction process and from the radiomic feature extraction process applied on the CT, MRI, and DWI-MRI scans. These datasets include both the feature values and the imaging segmentation binary files.
- The genomic analysis datasets, which are produced from the BD2Decide tools, responsible for the genomic feature extraction process applied on the tissue samples.

Apart from the above, the following datasets are considered to be part of the analysis data category:

- The models developed in the BD2Decide project, which are used in the prognostic prediction use cases.
- The prospective data, which are part of the prospective clinical studies and are calculated after the application of the prognostic prediction models on a patient's health record.
- The BD2Decide Knowledge Base. This is the project ontology that describes the semantic relationships between the actors of the BD2Decide environment with the functions expected from the use of the respective components to implement the envisaged use cases.
- The metadata description of the patients' health record, according to the BD2Decide ontology.

2.2.3 Visualisation data

Another category of the data involved in the BD2Decide environment refers to the visualisation tasks, which allow the professionals in the HNC health care field and the patients to assess the analysis performed within the BD2Decide environment. Therefore, in this section, we present the type of the visualisation data, which integrate the collection and analysis data in order to present meaningful information to the end users of the BD2Decide environment.

As part of the *visualisation data*, we envision the structural information required to support intuitive visualisation schemes. In this category, we include the models that represent the patient's 3D digital avatar, consisting of 3D models from head and neck, which emphasise on the anatomy morphology and the visualisation of the levels for lymph-nodes and tumors.

It must be noted that the visualisation data category integrates the datasets of the other two categories in the sense that, subject to the technical requirements of the use cases from Deliverable D2.1 [1], an analysis dataset may require additional data representation structures on the presentation layer.

2.2.4 Access control data

This is another BD2Decide data category, which is not directly implied from the use case analysis in D2.1 [1], but it is a result of the technical requirements. Those requirements have been compiled to translate the business level functions reflected in the use cases into technical requirements for the development of the expected components and the BD2Decide platform. Furthermore, this data category is introduced, as a result of a legal analysis of the expected BD2Decide use cases. As such, the processing of patients' data and the use of big data infrastructures introduces the need for identity management, authentication and authorisation mechanisms, which are implemented through the exploitation of the following user profile data:

- The credentials for the authentication of end users in the BD2Decide platform.
- The list of roles and the corresponding access rights.
- The pseudonymization and anonymization model.

2.3 Data Sources

The BD2Decide platform manages a collection of multiscale and multivariable patients' data from different sources, including personal data from the participating hospital information, data manually inserted from clinical specialists, histology data, image data from PACS and from the BD2Decide diagnostic image processing tools, and data from research prognostic models.

Specifically, data inputs come from (see Figure 2):

1. Data from clinical centres, consisting of:
    a. Retrospective studies
    b. Prospective studies
2. Population data, used for the "Knowledge Base". These data include:
    a. Environmental data
    b. Medication data
    c. Lifestyle Behavioral data
    d. Epidemiological data
    e. Health Data

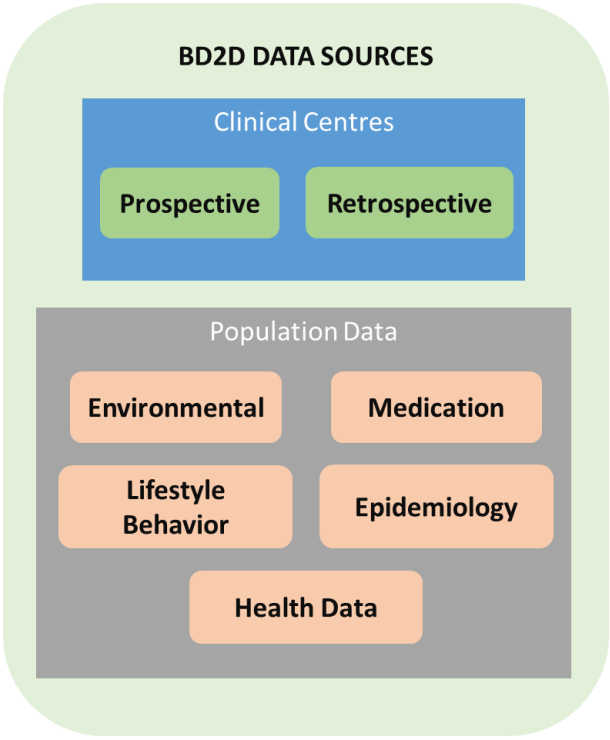


Figure 2: BD2Decide platform data sources

This set of data is subsequently used as inputs by an integrated CDSS that links together population-specific epidemiology, behavioral and environmental data, patient-specific multiscale data from genomics, pathology, clinical and imaging data (including radiomics). Moreover, data is used for a set of multiscale prognostic models, the Big Data Infrastructure and the graphical

visualization tools of the BD2Decide CDSS.

2.3.1 Clinical centres inputs

Five clinical centres retrieve prospective and retrospective patients inputs and provide the information to the BD2Decide cloud infrastructure. The inputs data coming from clinical centres refer to Hospital Information System (HIS), Laboratory Information Systems (LIS), and Picture Archiving and Communication Systems (PACS). The collection of data in the clinical centres is based on widely adopted clinical standards such as HL7, ISO/CEN 13606 and openEHR, adopting the IHE Cross-Enterprise Document Sharing (XDS) Integration Profile, in order to facilitate data collection.

A cloud exportation procedure that anonymizes and de-identifies the dataset according to the data policy of each clinical centre is executed. Internet Repositories refer to external database entities, which supply information on epidemiology and behavioural population.

The data collection in the clinical centre is used for two sets of clinical studies, namely retrospective and prospective, as following:

- ❑ Sources (existing systems providing data)
  - HIS/LIS
    - Patients' retrospective study data
    - Patients' prospective study data manual input
    - Patients' admin data (i.e. frequency of visits, timing, correlation to lab data)
  - PACS
    - Imaging data for retrospective study data

After the anonymization process, the clinical inputs are stored in an electronic Clinical Report Form (e-CRF). The e-CRF is the tool used by the physicians to collect the patients' clinical data from each participating clinical site. This tool allows clinicians for a faster and more efficient control of the patient data, while ensuring high security of the data and an easy and friendly usage.

The e-CRF is organized in 13 main Items, each one representing a database block responsible for a certain set of patient data, as shown in Table 1.

Table 1: Records to be stored for each patient in BD2Decide

One set of data per patient	Multiple sets of data per patient
Clinical data (first visit, pre treatment): Items 1,2 and 3 in e-CRF	Imaging data (CT and/or MRI data): Item 5 of e-CRF. We may have only one image (CT or MRI) or both
Tumor data: Item 4 of e-CRF	Radiomics data: multiple radiomics readings for the same image (Items 5b and 5a). Radiomics is linked to an image in Item 5 of e-CRF

One set of data per patient	Multiple sets of data per patient
Lymph nodes data: Item 4, N characteristics in e-CRF	Follow-up: Item 11 of e-CRF
Surgery data: Item 9 of e-CRF	Chemotherapy data: multiple sets if many chemotherapy treatments are applied to the tumor: Item 7 of e-CRF
Tissue sample: Item 10 of e-CRF	Radiotherapy data: multiple sets if many radiotherapy treatments are applied to the tumor: Item 8 of e-CRF
Pathology on tumor: Item 6 of e-CRF	Toxicity (one set of data for each treatment): Item 12 of e-CRF
	QoL questionnaires: Item 13 of e-CRF

2.3.2 Population data

Population-based cancer registry can provide data for the patients (date of birth, country of residence and gender), tumor characteristics (histology and topography), vital status, etc. Some cancer registries have information also on the hospital of diagnosis, on the main treatment and the hospital of treatment and on the extend of the disease.

The population-based registries host the following type of data:

- ***Data on Environmental indicators***: these inputs include the identification of Open Data sources containing environmental data, since genetic changes that are responsible for H&N cancer may be linked to the environmental exposures that damage DNA, such as air pollution in a specific region.
- ***Scientific data on medical knowledge***: these inputs refer to research and scientific publications and international guidelines.
- ***Lifestyle Behaviour data***: this information is retrieved through QoL questionnaires, in order to estimate the quality of life of the patients and compare it during the post-treatment and follow-up phases for population-based statistical purposes.
- ***Epidemiology data***: this information refers to tumor registries data at the level of populations or even at the level of the single individual, gathered from epidemiology datasets at the EU level, such as IARC [2].
- ***Health data***: this information refers to residents' demographic health variables, clinical variables (i.e. information on all treatments and procedures administered), administrative variables (i.e. type of admission, day hospital), which is aggregated on a per region or country level to provide information for population-based statistical purposes.

The information on the population data will be gathered from national public databases in Italy, the Netherlands and Germany [3][4].

3 THE ARCHITECTURE OF THE DATA WAREHOUSE

3.1 Logical Architecture

In the previous section 2, we focused on the data requirements in the BD2Decide project and the sources for gathering or producing the respective information. Based on this analysis, in this section we present the data warehouse service architecture, which emphasizes on the depiction of the required data repositories for hosting the type of data identified in the categories of section 2. This architecture is driven by the requirements defined in D2.1 [1] and the corresponding technical use cases described there.

Figure 3 shows the logical architecture of the data warehouse services for the BD2Decide environment, which depicts the repositories with the datasets involved in this environment. In this architecture, we define three main parts, each of which plays a different role in the overall data warehouse architecture. The first part is the clinical information systems, which refers to the data collection category of Section 2.2.1. The second part is the BD2Decide Storage Environment and integrates data from the analysis and visualization phase (see Sections 2.2.2 and 2.2.3) and the access control data of Section 2.2.4. These categories are also reflected in the third part of the logical architecture of the data warehouse services, which refers to the BD2Decide big data infrastructure.

The Clinical Information Systems (CIS) part of the architecture involves the information systems already being in operational mode in the clinical centres participating in the BD2Decide project. Such systems integrate the following sub-systems:

- The Hospital Information System (HIS), which manage the patients' presence at the hospital, along with information regarding the purpose of visit and the assigned professionals to handle the patients.
- The Lab Information System (LIS), which host the results from the clinical monitoring of the patient on a laboratory level, including the management of tissue samples from the patients and their genomic analysis.
- The Picture Archiving and Communication System (PACS), which manages the storage, processing and distribution of medical images.

Due to the implications of the national and European legislations about data security and privacy in the healthcare domain, and the internal security policies applied in the participating clinical centres, the project does not have direct access to the CIS. As depicted in Figure 3, a replication of the CIS in each centre is developed, which is an extract of the operational repositories in the CIS and hosts only the information that is absolutely required by the BD2Decide project to conduct the planned research and innovation activities.

Further to it, it must be highlighted that the replication of the CIS performs anonymization of the data shared with the BD2Decide environment, which resides outside the network of the clinical centres. This is necessary for the project to demonstrate compliance with the national and European laws and minimize the privacy risks from the exposure of the patients' medical information outside the network of the clinical centres.

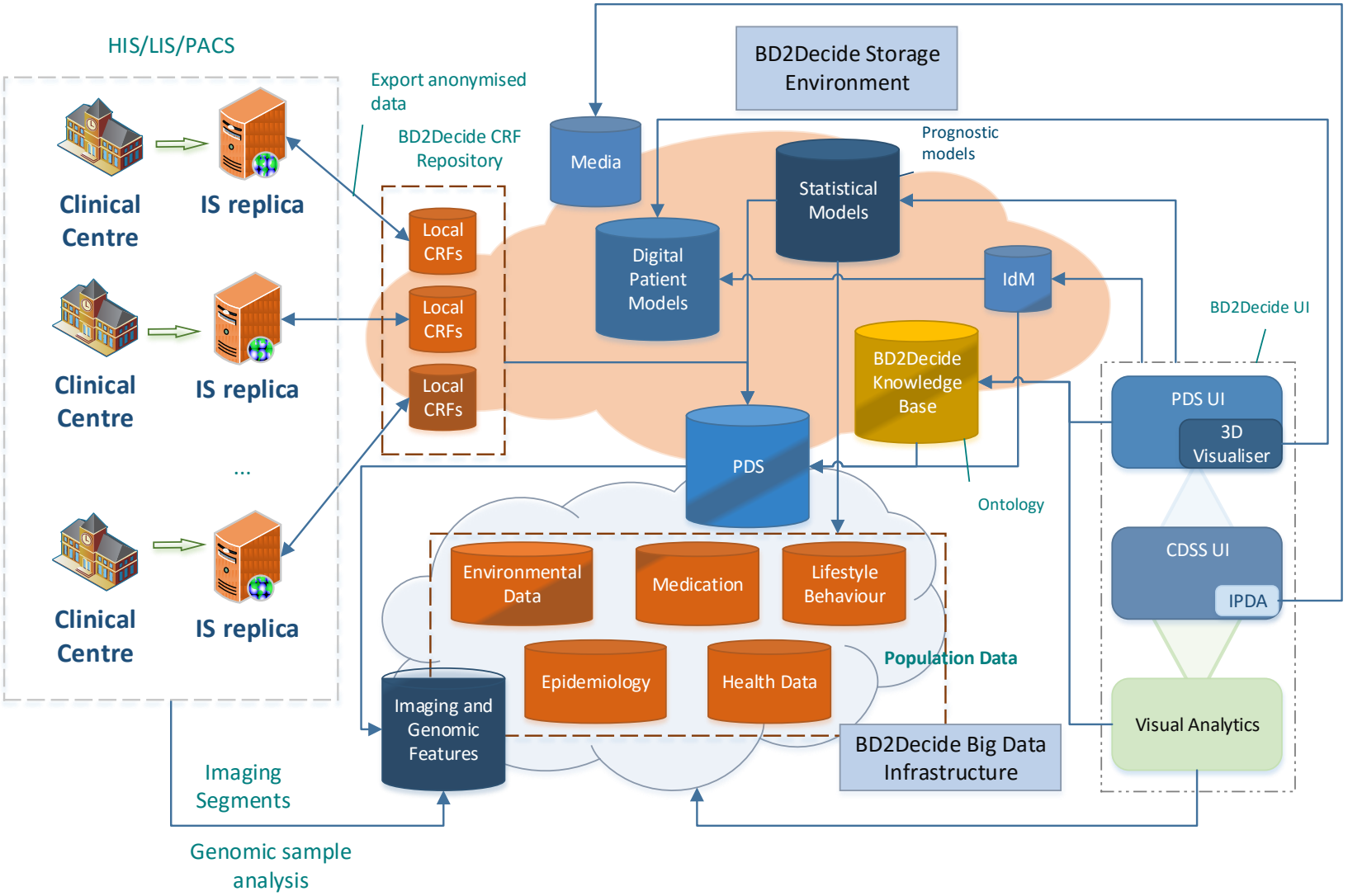


Figure 3: The logical architecture of the BD2Decide data warehouse.

We, also, emphasise here on the fact that the replication of the CIS files the patients' consent form, through which the patients participating in the clinical studies give their written consent to the relevant clinical organisations to exploit their medical information for the research purposes of the project. As mentioned above, no personal data is fused outside the CIS of the clinical partners, thus the signed consent forms must only be filled in the replication of the respective CIS and not within the BD2Decide environment.

On the BD2Decide Storage Environment side, the project defines a set of data repositories, which facilitate the analysis and visualisation tasks of the BD2Decide environment. These repositories are described below:

- The Local CRF repository; this is a repository of the retrospective studies data, which are first collected in the replication of the CIS and are pushed to these repositories in an anonymized way for further processing within the BD2Decide project.
- The Patients Documentation System (PDS); this is a central repository in the BD2Decide environment, which integrates all the patients' data required for the implementation of the project use cases. In that respect, the PDS interfaces with the local CRFs to retrieve data from retrospective studies, while it maintains the respective data from the prospective studies, after the analysis invoked within the BD2Decide environment.
- The Identity Management (IdM) repository; this is the data base for handling identity management. It maintains the identification details for the professionals experiencing with the BD2Decide environment. During the course of the project, we will further investigate whether this database is extended for patients as well.
- The Statistical Models repository; this database hosts the structure of the prognostic models, which are defined in the course of the project and are invoked in the prospective studies to project the patients' clinical status, according to the patient's personalised data and the population medical information.
- The Digital Patient Models (DPM) repository; this repository maintains the models required in the project to provide advanced visualisations of the patients' medical information in the form of a digital avatar. Through the contents of this repository, the relevant User Interface components will retrieve information about the appropriate composition of the avatar to visualize the clinical status of a certain patient.
- The Media repository; this is a multimedia database hosting the media items used in the development of the co-decision aid tool for the various disease cases and per the workflow envisioned in each clinical centre.
- The BD2Decide Knowledge Base; this is the ontology of the project that describes the structure of the conceptual entities involved in the research of the project and the relationship among them. The Knowledge Base may strongly depend to the PDS to describe the results analysis of the patients' data, after the application of the Knowledge Base on the structure of the clinical studies.

Moving to the BD2Decide Big Data infrastructure, this part of the logical architecture hosts information that requires big data processing and/or storage. The relevant repositories are:

- The imaging and genomic features database; this connects to the replication of the CIS of the clinical centres and maintains references to the patients' imaging data from scanning radiology devices and tissue sample data, as well as the resulting imaging radiomic and genomic features. Both the imaging radiomic and genomic features are also part of the electronic CRFs, which are maintained in the Local CRF repository, but, for the project purposes, we also maintain this database as a separate storage area for these features.
- The population repositories; this is a set of repositories, which integrate population data, such as environmental, epidemiology, health data, and information about statistical mediation and life style behaviour.

The data repositories that were described above constitute the basis of the BD2Decide data warehouse architecture and they are accessed, according to the data acquisition and management service functionalities, which are introduced in the next section 3.2.

3.2 Data acquisition and management service functionalities

This section describes the functionalities expected from the data acquisition and management services, which process the repositories depicted in the logical architecture of the BD2Decide data warehouse, in Figure 3. These functionalities comprise a sub set of the expected user level functionalities of the BD2Decide CDSS and they result in the implementation of work flow activities for the collection, analysis, management and visualisation of the BD2Decide datasets.

For the purposes of this document, the presentation of the data level functionalities is grouped into: i) functionalities of the data acquisition services: these functionalities, ii) functionalities of the data processing and management services, iii) functionalities of the data visualisation services.

These functionalities are presented in the following sections in the form of abstract workflow processes. The latter comprise an introduction to the use case flows that the BD2Decide tools will implement in the respective platform prototype. As such, these workflow diagrams aim to reflect the high level data flow within the BD2Decide environment and provide a first approach to the architecture specifications in D2.3, which is due M12 (December 2016).

3.2.1 Data Acquisition Service Functionalities

The data acquisition service functionalities refer to the collection of data from external sources. These functionalities mainly refer to the data collection category.

The most important functionality in the data acquisition services is the selection of the patients and their corresponding data to be involved in the retrospective and prospective studies. The patient selection process and the relevant workflow to collect the appropriate data of the patient's record are presented in Figure 4. As depicted there, the clinical centres define the criteria for the selection of patients to participate in the BD2Decide clinical studies. Following this selection process (which is out of scope of this deliverable), the respective patient's digital record must be extracted from the CIS of the centre and be stored in the replication system, which is dedicated to the BD2Decide project. For all the patients participating in the studies, we collect the appropriate information

present in the CRF from the CIS, including the images from CT or MRI scans and the sample reference numbers (with potentially a reference to the storage area of the sample).

The result of this workflow is the extraction of the patient's record from the CIS to be ported into the replication of the CIS (residing inside the network of the clinical centre). Before the record is ready for publication to the BD2Decide environment, the assigned data owners of the records (which are appointed from the clinical partners) perform anonymization on the record (through generating an encoded identification number - ID for the patient), so that any analysis occurred in the BD2Decide environment beyond this point cannot refer directly to the patient (as a physical entity) that this record belongs to. To this end, as part of the information stored in the replication of the CIS is the patient's signed consent form, through which the patient agrees on the use of their data for the research purposes of the project, and a map table associating the patient encoded ID with the identification number of the consent form.

As part of the data acquistion services, the BD2Decide project exploits open data from the Internet about population statistics, which refer to average data on the environmental conditions affecting head and neck diseases, epidemiology data per country and population group, genarl overview data of the health status per country or ethinicity, statistical data on the average use of medication on a per country basis and the life style customs and behaviour for general indicators affecting the survival rate per country, like smoking and alcohol consumption. The workflow which is engaged in the process is shown in Figure 5.

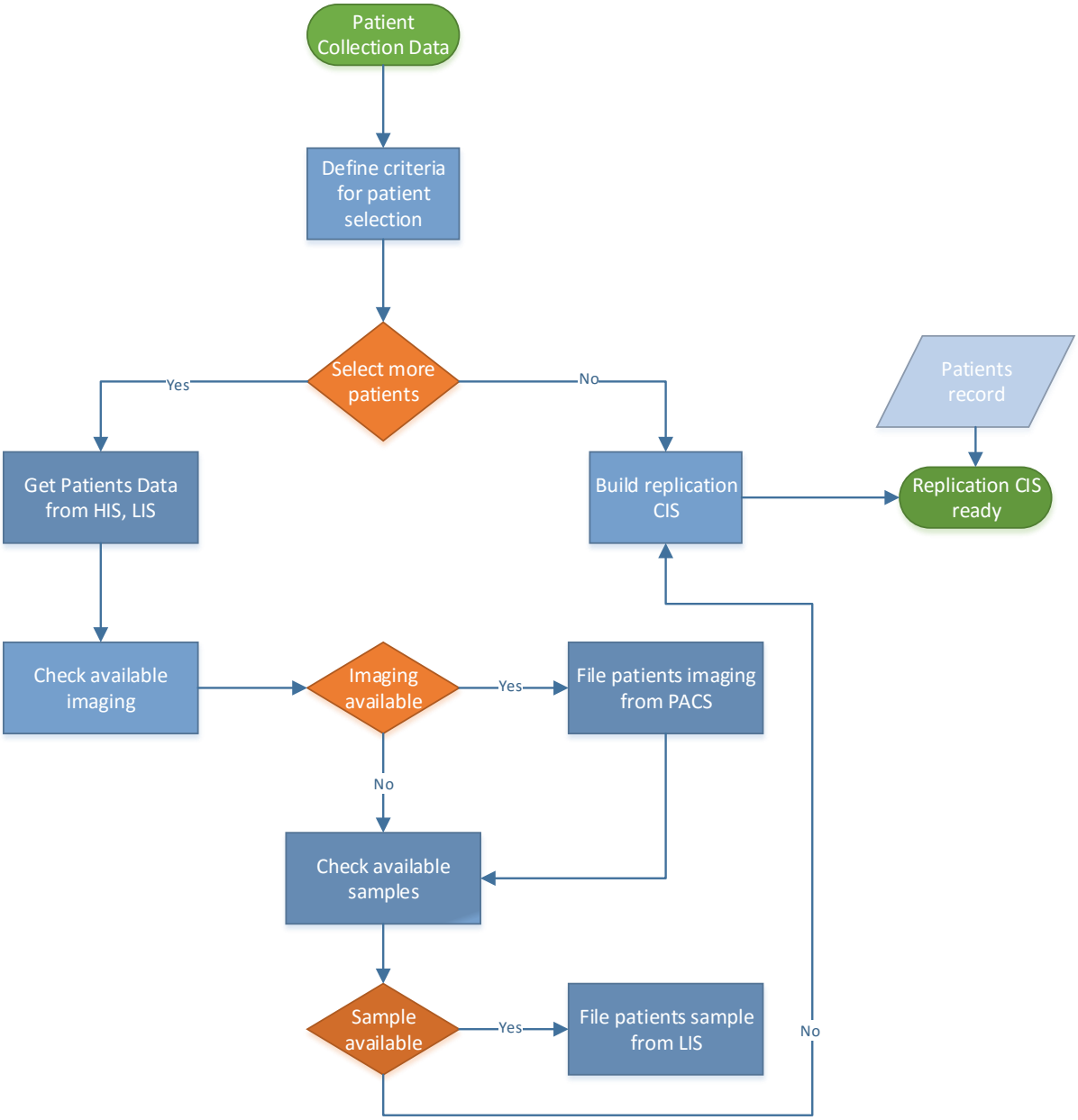


Figure 4: Selecting patients for clinical studies.

As shown in Figure 5 and because the project spans along more than 3 years, we take into account the case that the population repositories (see Figure 3) are updated on a yearly basis. As a result, we must introduce a content update process, which allows the population repositories to be enriched with statistical data of the latest calendar year, as soon as they are made available.

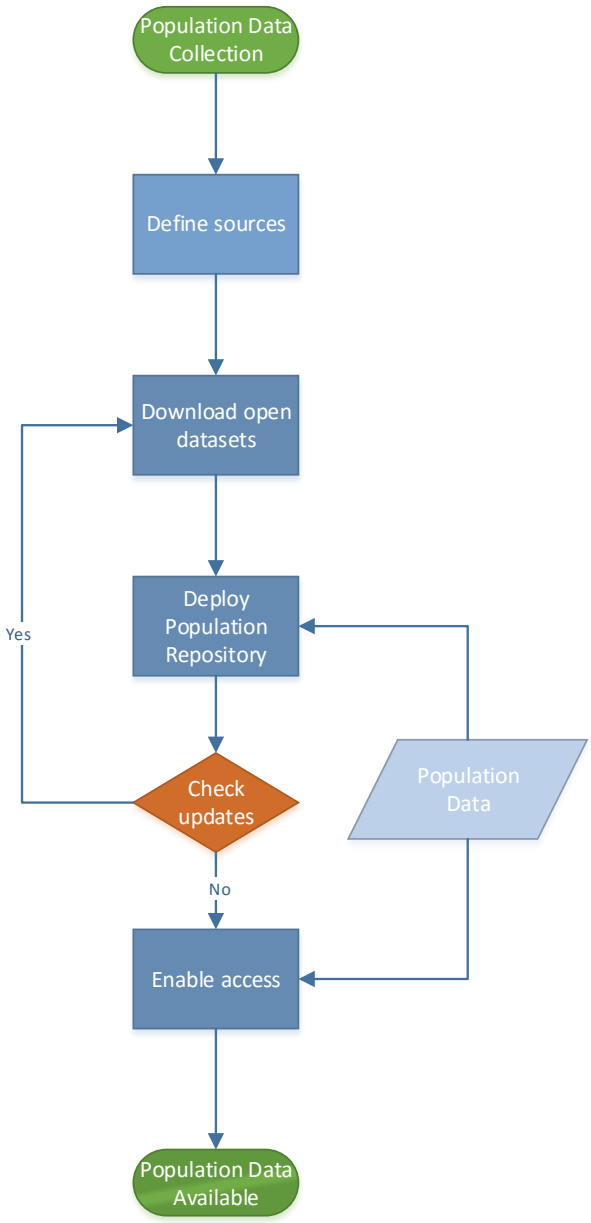


Figure 5: Collecting population data.

3.2.2 Data Processing and Management Service Functionalities

The data processing and management service functionalities refer to the use of the BD2Decide tools to analyse data collected from the replication of the CIS. These functionalities mainly refer to the data analysis category.

As part of the analysis functionalities in the BD2Decide project, the available media from the PACS of the CIS are processed to extract imaging features, customised to the radiomic analysis of these images. Although the exact process for extracting the radiomic features may involve detailed steps, a general overview is shown in Figure 6. This figure, also, presents the (similar) approach for the extraction of genomic features, which refer to the genomic analysis of (tissue) samples, performed in a LIS environment.

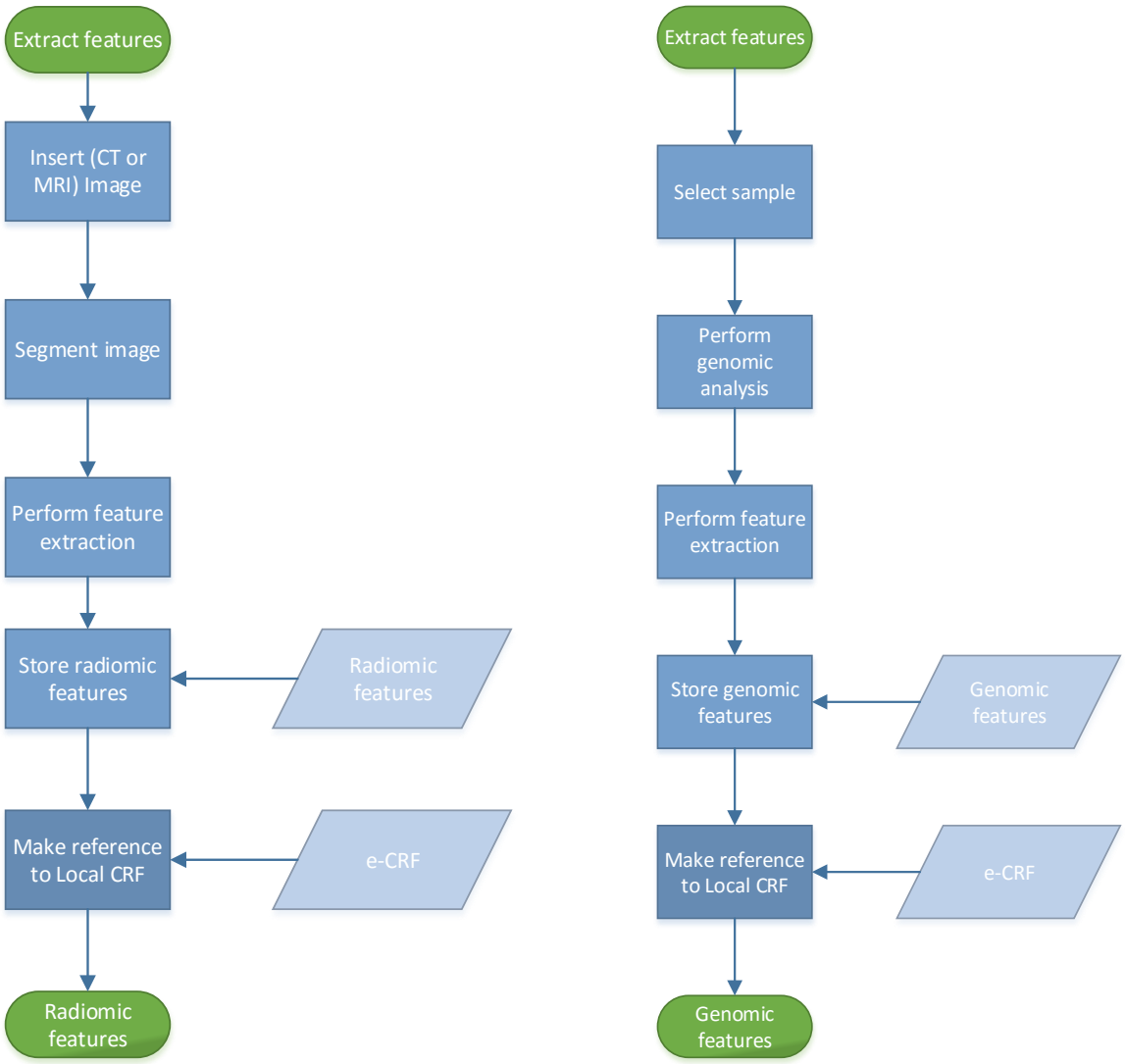


Figure 6: Extraction of features from: i) radiomic analysis, and ii) genomic analysis.

Further to the above analysis, BD2Decide builds PDS, which is a repository aggregating the anonymized patients' records for all the patients participating in the clinical studies. As presented in Section 3.1, this repository integrates all the information about the patients. The workflow for creating PDS is shown in Figure 7.

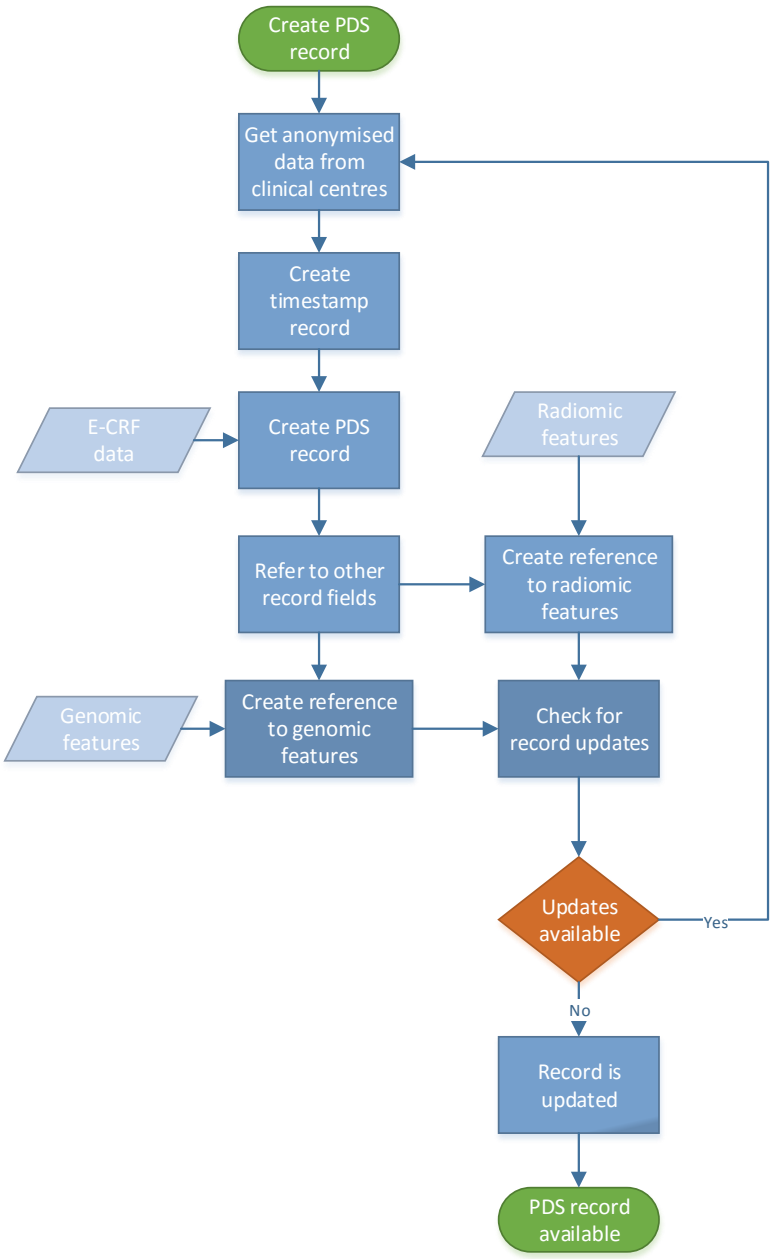


Figure 7: Developing the patient's record into the PDS.

3.2.3 Data Visualisation Service Functionalities

From a data perspective, the functionalities of the data visualisation services refer to the co-decision process, which allows the physicians and the patients to collaborate towards deciding the most appropriate treatment method to follow, the presentation of the patient's record, in order for the professionals to assess the clinical status of the patient, based also on the prognostic prediction of the patient's survival rate and other indicators, and the exploration of data for research purposes, in which researchers perform different types of queries to extract data from various patients and visually compare statistical cases, such as the application of a specific treatment in various patient records.

The high level overview workflow for the data process realising the co-decision approach is presented in Figure 8.

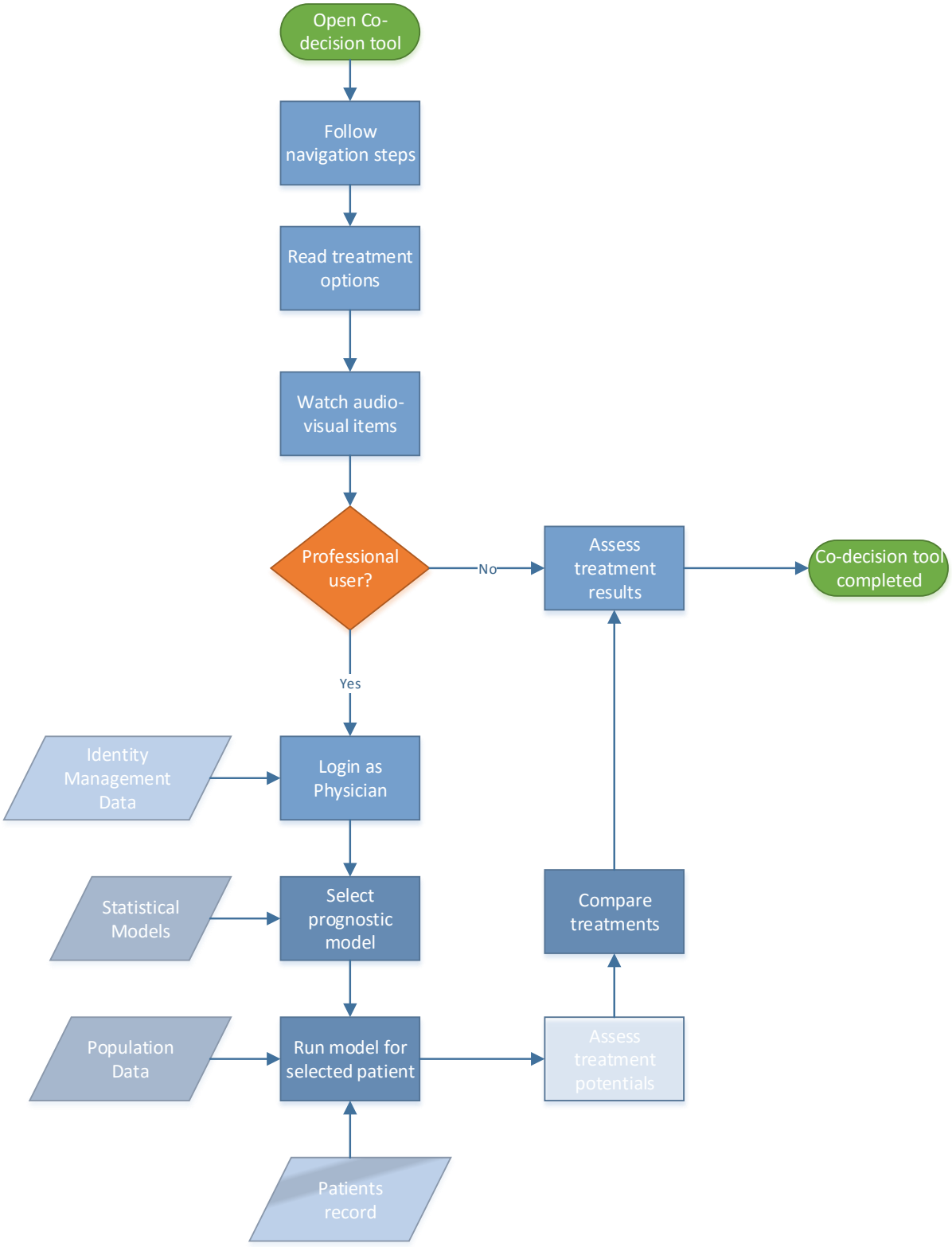


Figure 8: Realising the data flow of the co-decision approach.

As shown in Figure 8, the workflow involves a publicly available part, which is accessible without any authentication, while physicians can enhance the decision process by proposing their patients the most preferable treatment solutions, by comparing these solutions, based on the expected survival rate, which is calculated from the prognostic prediction analysis on the clinical status of the

patient. This personalised co-decision approach integrates the prognostic statistical models, publicly available population data and the selected patient's record.

In the functionalities of the data visualisation services, we, also, include the visualisation of the digital patient view process. This is a workflow that implements the data interaction flow for the presentation of the patient's clinical status in various forms. As shown in Figure 9, the visualisation of the patient record may involve various steps.

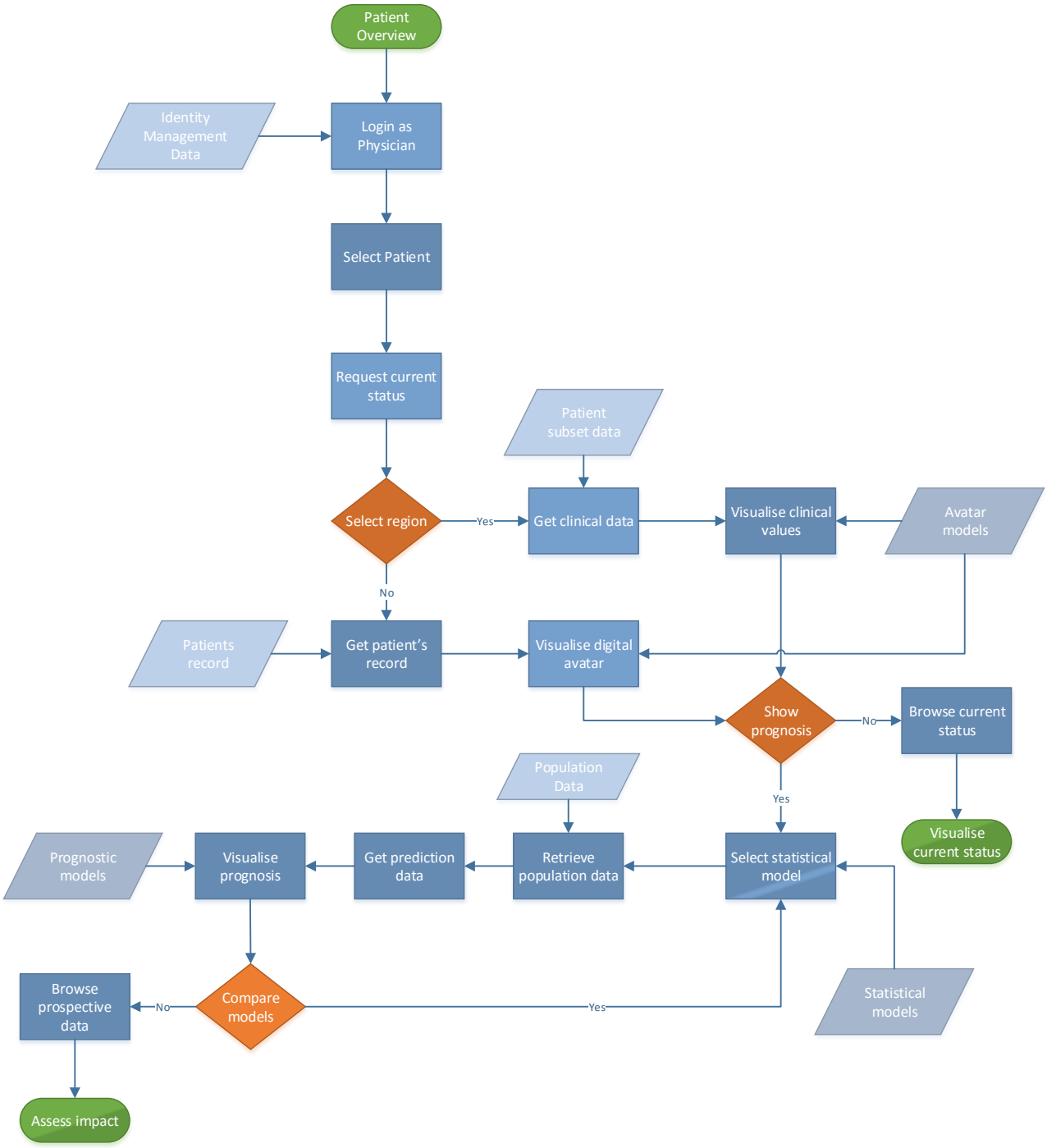


Figure 9: Visualising the virtual patient.

The visualisation data involved in this service functionality include almost all the data repositories defined in the BD2Decide environment. The tasks in this process result in the presentation of the patient's record, either as a whole (view the digital patient avatar) or parts of it (selected fields from the record and/or the avatar model segments). Furthermore, through invoking the prognostic prediction analysis, the workflow of Figure 9 results in the professionals being able to assess the impact of the followed treatment method on the patient's Quality of Life indicators, in order to decide on the follow-up strategy.

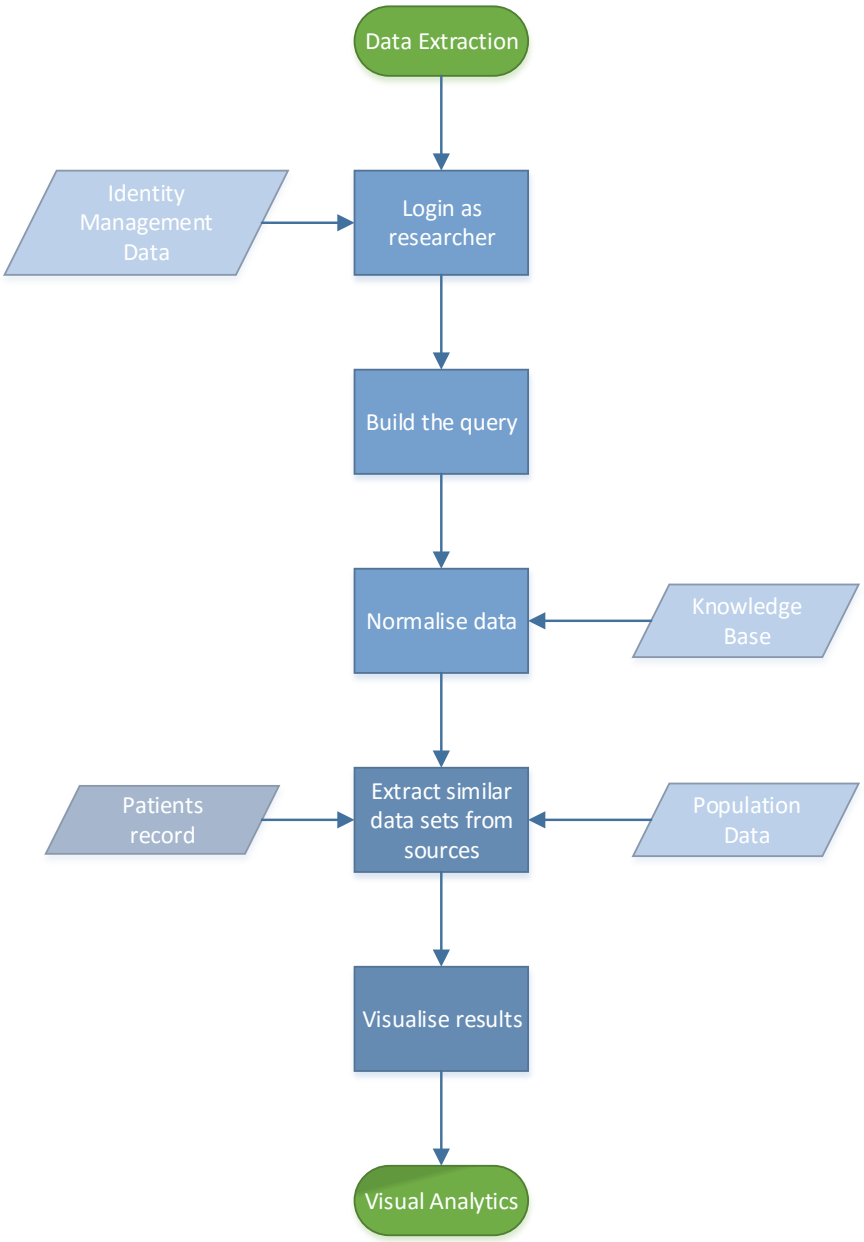


Figure 10: Realising the researcher cases in the BD2Decide environment.

Another data visualisation service functionality engages the BD2Decide Ontology, which is structured to allow a common definition of entities and their relationships in the healthcare HNC field. The workflow presented in Figure 10 describes the involvement of the data repositories in the realisation of the researcher use cases in the project. As such, researchers may have access to the

anonymised patient's data for research purposes and, to this end, they need for a unified query formulation approach to be able to extract similar data from various sources. This data normalisation process is supported by the BD2Decide ontology (knowledge base), which allows clustering data and presenting them in a way that fits to the query request.

4 DESCRIPTION OF DATA REPOSITORIES

This section elaborates on the design and specification aspects of the BD2Decide data warehouse architecture that was presented in Section 3.1. It advances the introduction of the data requirements and the respective sources in Section 2 and analyses the structure of the databases envisioned in the BD2Decide environment, along with the candidate technologies and /or tools, which support the development of these data structures.

4.1 Local Electronic Clinical Record Format Document

The local electronic Clinical Record Format (e-CRF) document is a database locally deployed for each clinical centre to enable clinicians collect the patients' health record required for the retrospective and prospective studies. Thus, this database integrates the information collected from each CIS of the participating clinical partners.

4.1.1 The structure of the local e-CRF

Although the CRF protocol is widely used in the clinical research, each clinical centre may follow a different approach to define the required fields constituting this document. In the BD2Decide project, we have already defined a common CRF scheme, which has been introduced in the annex of Deliverable D2.1 [1]. As shown there, and as it has been reported in Section 2.3, the e-CRF database for the BD2Decide project should maintain the categories of records shown in Table 1.

Since the details of the CRF have been presented in [1], we only present here the structure of the local e-CRF database, highlighting the logical connections between the tables. More specifically, recalling from Deliverable D2.1 [1], the CRF structure is split into 13 items. The starting point for the development of the CRF is the definition of a study. In BD2Decide, we define two study categories, namely retrospective and prospective. Each category follows the same structure for the CRF and distinguishes from the other on the time that the study is created and tools, services or human resources being responsible for the completion of the various fields in each study.

As shown in Figure 11, a study is recruited by many patients, which are selected to participate in the study, following a set of criteria. Only if the patient signs a consent form, he/she can participate in the study by providing their health record. Once this consent form is available, the demographic and clinical data of the patient is collected, while further HNC related data is produced to fill in the CRF for a specific patient.

The complete view of the local e-CRF database is depicted in Figure 11.

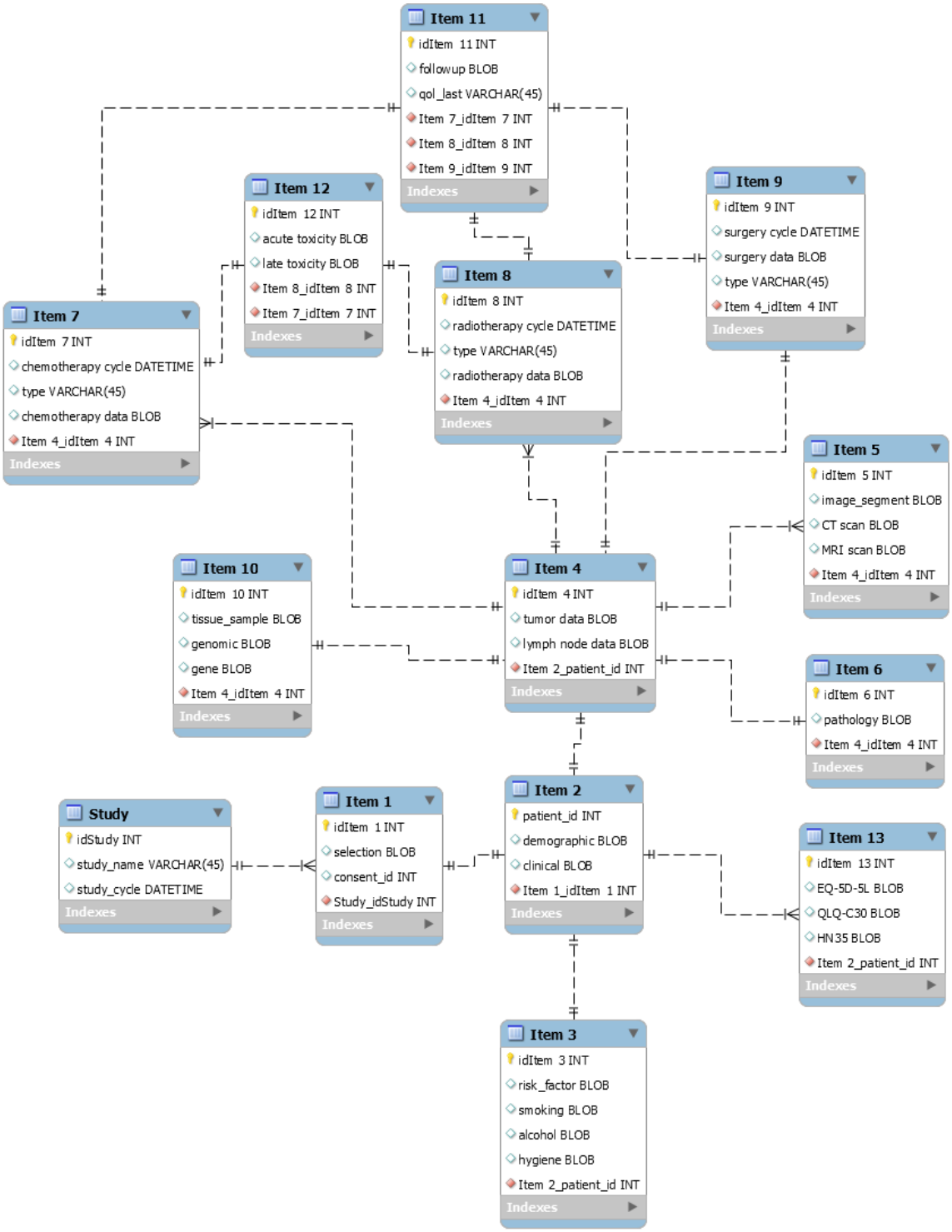


Figure 11: The structure of the local e-CRF database.

4.1.2 Tools and technologies for developing the e-CRF

The development of the e-CRF follows best practices from the clinicians' domain. As such, for the development of the local e-CRF we have considered OpenClinica[2]. OpenClinica is the world's most

[2] https://www.openclinica.com/

widely-used, open-source software for clinical research. First released in 2005, OpenClinica is designed to meet the diverse needs of modern research environments. It is built as a lightweight, extensible, and modular application[3].

OpenClinica provides the capture, clear and manage of the electronic clinical data and helps clinicians to organise and analyse their research. More specifically, OpenClinica supports the collection of electronic data through the e-CRF format and increase the clinical research quality and efficiency. Moreover, clinicians can manage their studies in real-time and in high scalability. Clinicians can create studies and import unlimited number of patients either manually or massively using the e-CRF format. Nevertheless, one can manage their records and be aware of any modification might happen to their records and they can automate visit schedules by creating events.

The administrator of OpenClinica can create different studies for every clinical centre and grant different permission to the involved clinicians (e.g. Data Manager, Data Entry person etc.). Clinicians can use and improve any clinical phase of their study (e.g. Phase I, Phase II, Academic etc.).

One of the greatest advantages of OpenClinca is that it offers monitoring functionalities to the clinicians in a dashboard like form. In the respective monitoring views, they can monitor the fulfilment percentage of their studies, actions such as patients' enrolments, events etc. With OpenClinica, clinicians can extract patient electronic outcomes, such as analytic reports. Finally, OpenClinica provides easy integration with external systems.

As aforementioned, OpenClinica is an open-source software that can support the dynamic import of e-CRF structures and build customisation with respect to the BD2Decide needs. During the installation, OpenClinica creates its own database either in Oracle or PostgreSQL (version 8.4). BD2Decide uses OpenClinica with PostgreSQL, which creates a database with 110 tables (see Annex7.1).

OpenClinica supports dynamic import of e-CRF data structures using the OpenClinica template. In BD2Decide, we have customised this template to meet the requirements of the CRF presented in Deliverable D2.1, so that the five clinical centres participating in the project be able to manage their patients' health records joining up the project in the execution of the pilots. The OpenClinica template is in Microsoft Excel format and, inside the OpenClinica tool, it is translated to a set of objects, such as text in string, date class, integers, etc.

Table 2 shows the mapping between the section items of the CRF, as it was presented in D2.1, and section label format of the e-CRF in OpenClinica, while a snapshot from the customised e-CRF excel format for the BD2Decide CRF structure is presented in Table 3.

[3] https://docs.openclinica.com/3.1/openclinica-user-guide/overview-openclinica

Table 2: Convert CRF items from D2.1 to the e-CRF scheme of OpenClinica.

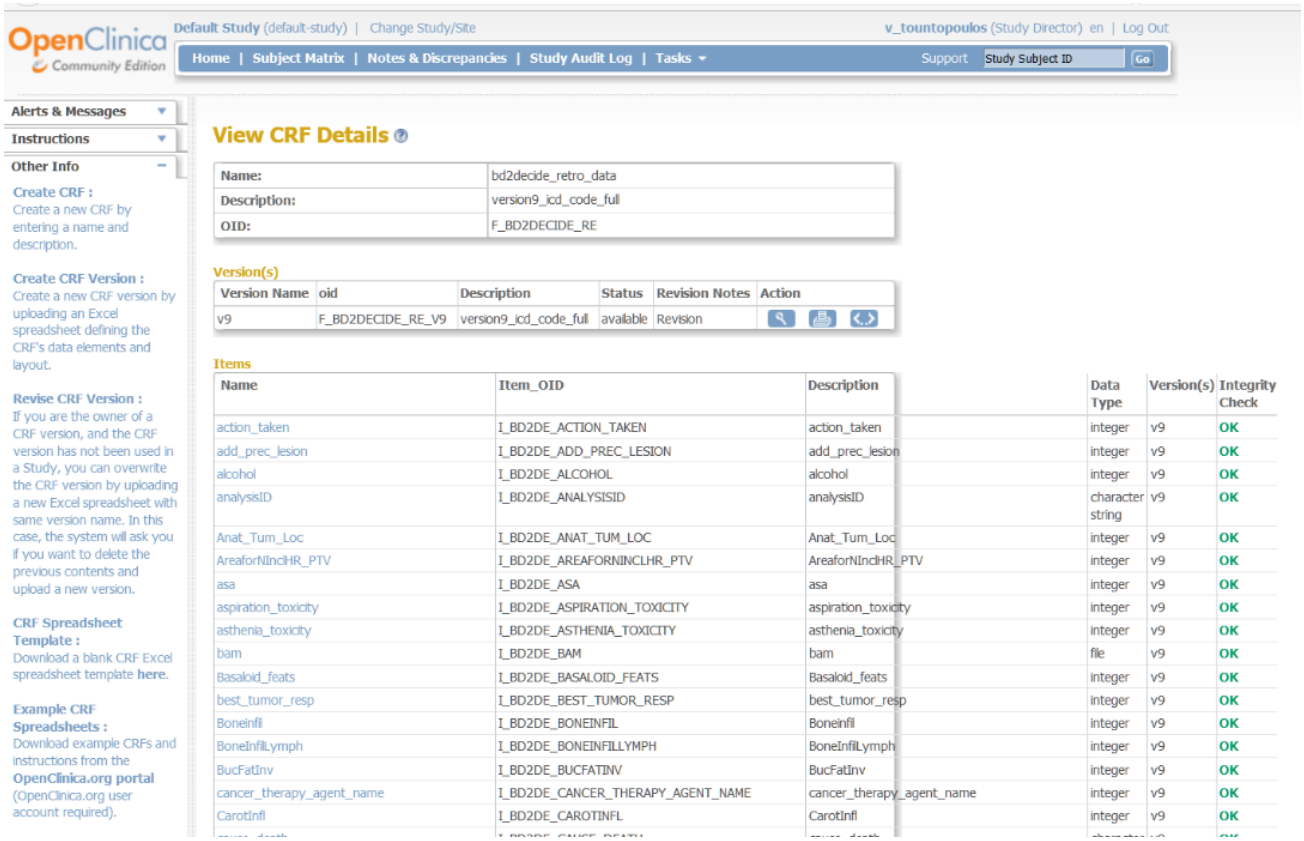
Section_label	Section_label
Patient selection	Patient selection
Demographic & Clinical data	Demographic & Clinical data
Risk factors	Risk factors
cTN	cTN
Imaging	Imaging
Radiomics data CT scans	Radiomics data CT scans
Radiomics data from MRI	Radiomics data from MRI
Pathology data	Pathology data
Chemotherapy	Chemotherapy
Radiotherapy	Radiotherapy
Surgery	Surgery
Tissue Sample	Tissue Sample
Genomic data	Genomic data
Genes relevant for prognosis	Genes relevant for prognosis
Follow up	Follow up
Toxicity	Toxicity
QoL	QoL

Table 3: An example for the OpenClinica e-CRF customised for the BD2Decide project.

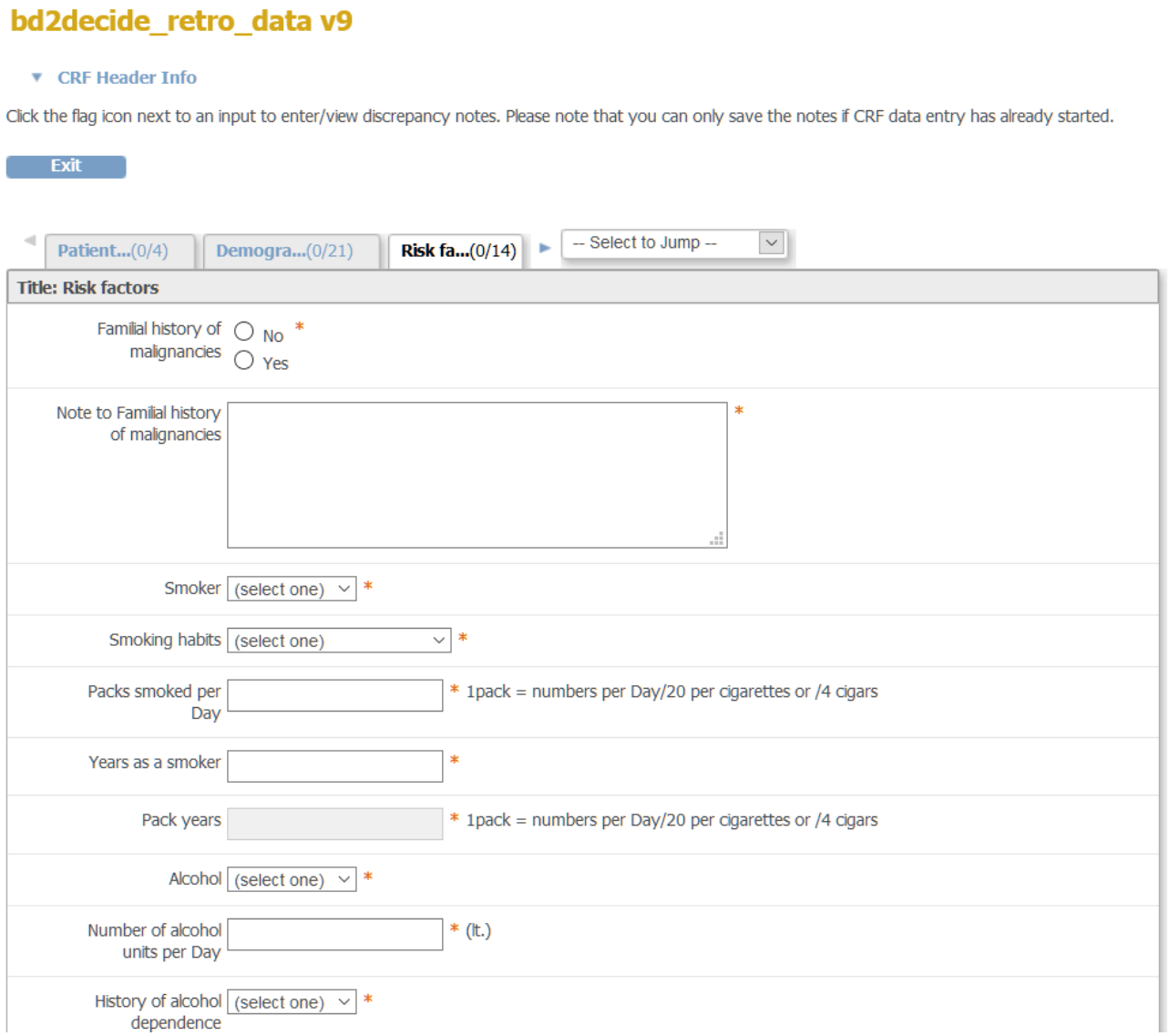
item_name	section_label	item_name	section_label	item_name	section_label
consent	Patient selection	single-select	consent_label	1,2,3	INT
DoICF	Patient selection	text			DATE
elC	Patient selection	radio	eligibility_criteria	1,2	INT
noElC	Patient selection	text			ST
PatientID	Demographic & Clinical data	text			ST

item_name	section_label	item_name	section_label	item_name	section_label
hospital	Demographic & Clinical data	single-select	hosp_label	1,2,3,4,5,6,7,8,9	INT
uId	Demographic & Clinical data	text			ST
lpID	Demographic & Clinical data	text			ST
DoB	Demographic & Clinical data	text			DATE
DoD	Demographic & Clinical data	text			DATE
diag_age	Demographic & Clinical data	calculation	diag_age_label	func:(DoD-DoB)	INT
sex	Demographic & Clinical data	radio	gender_label	1,2	INT
ethnicity	Demographic & Clinical data	single-select	ethinicity_label	1,2,3,4,5	INT
ethNote	Demographic & Clinical data	text			ST
asa	Demographic & Clinical data	single-select	asa_label	1,2,3,4,5,6	INT
comor_score	Demographic & Clinical data	single-select	asa_score_label	0,1,2,3,9	INT
hb_level	Demographic & Clinical data	text			REAL
plt_level	Demographic & Clinical data	text			REAL
lymph	Demographic & Clinical data	text			REAL
hpv_status	Demographic & Clinical data	single-select	hpv_status_label	1,2,3	INT

(a) Home Page



(b) The details of the e-CRF



(c) An example of e-CRF view for the risk factors table.

Figure 12: Screenshots from the BD2Decide instance of OpenClinca.

Further to it, we mention that OpenClinica allows massive import of e-CRF data through XML.

A set of screenshots from the implementation of the local e-CRF through OpenClinica is presented in Figure 12.

While most of the tables in Figure 11 are implemented through the OpenClinica tool, for Item 13 of the CRF (the QoL questionnaires) we use Limesurvey[4]. This is a Web-based open source tool for creating and managing surveys. The reason for doing so was to provide us with flexibility on the analysis of the questionnaires where needed. Further to it, Limesurvey offers multilingual survey setup, which is a desirable feature for us, since the QoL questionnaires have to be accessed from the

[4] https://www.limesurvey.org/

physicians and their patients, thus being in the native language of the patient is an added value for them.

Using Limesurvey, we have deployed the multilingual versions (in English, Italian, Dutch and German) of the QoL questionnaires from:

- The EQ-5D-5L questionnaire developed by the EuroQol Group[5] for the measurement of health outcome;
- The QLQ-C30 from EORTC[6], which is a questionnaire developed to assess the quality of life of cancer patients;
- The QLQ - H&N35 from EORTC, which is a questionnaire developed to assess the symptoms or problems arisen from a treatment followed by cancer patients.

A screenshot from the implementation of the QoL part of the e-CRF through Limesurvey is presented in Figure 13.

Figure 13: Screenshot from the Quality of Life questionnaires developed using Limesurvey.

4.2 Patients' Documentation System

The Patients' Documentation System (PDS) integrates the health records of the patients from the various clinical centres and hosts the information required for the visualisation of the expected tasks within the BD2Decide environment. To this end, PDS aggregates the information stored in the local e-CRF database from section 4.1, along with data directly received from other BD2Decide databases (i.e. the prognostic models database), or repositories from the clinical centres (e.g. imaging databases, etc.).

[5] http://www.euroqol.org/eq-5d-products.html

[6] http://groups.eortc.be/qol/

4.2.1 The structure of PDS

The core of the PDS structure, which is shown in Figure 14, is the virtual patient class, which represents the basic profile of a patient in the BD2Decide environment. This class links the BD2Decide Clinical DSS environment with all the information that might make reference to the information on a specific (anonymised) patient in BD2Decide. The virtual patient is connected to:

- Hospitalisation information: this table integrates any data retrieved from the HIS of the clinical centres, listing information about hospitalisation of the patient mapped to the virtual patient (the mapping is done inside the clinical centre and only the patient ID is populated to the BD2Decide environment), and the physician being responsible for the patient.
- The patient's health record, which is represented as an extension of the scheme presented for the local e-CRF in Section 4.1.1).
- The tumor board class, which manages the virtual / physical meetings occurred between the involved specialties in a potential tumor board being held to assess the decision required following a diagnosis on a certain patient.
- The patient prognosis class, which links to references for the patient radiomics and genomics features involved in the prognosis of the life expectancy.
- A table maintaining information about the avatar visualisation of the virtual patient.
- The tracking of the patient's selections when running the co-decision aid tool with their physicians to evaluate the impact of a treatment in their life.

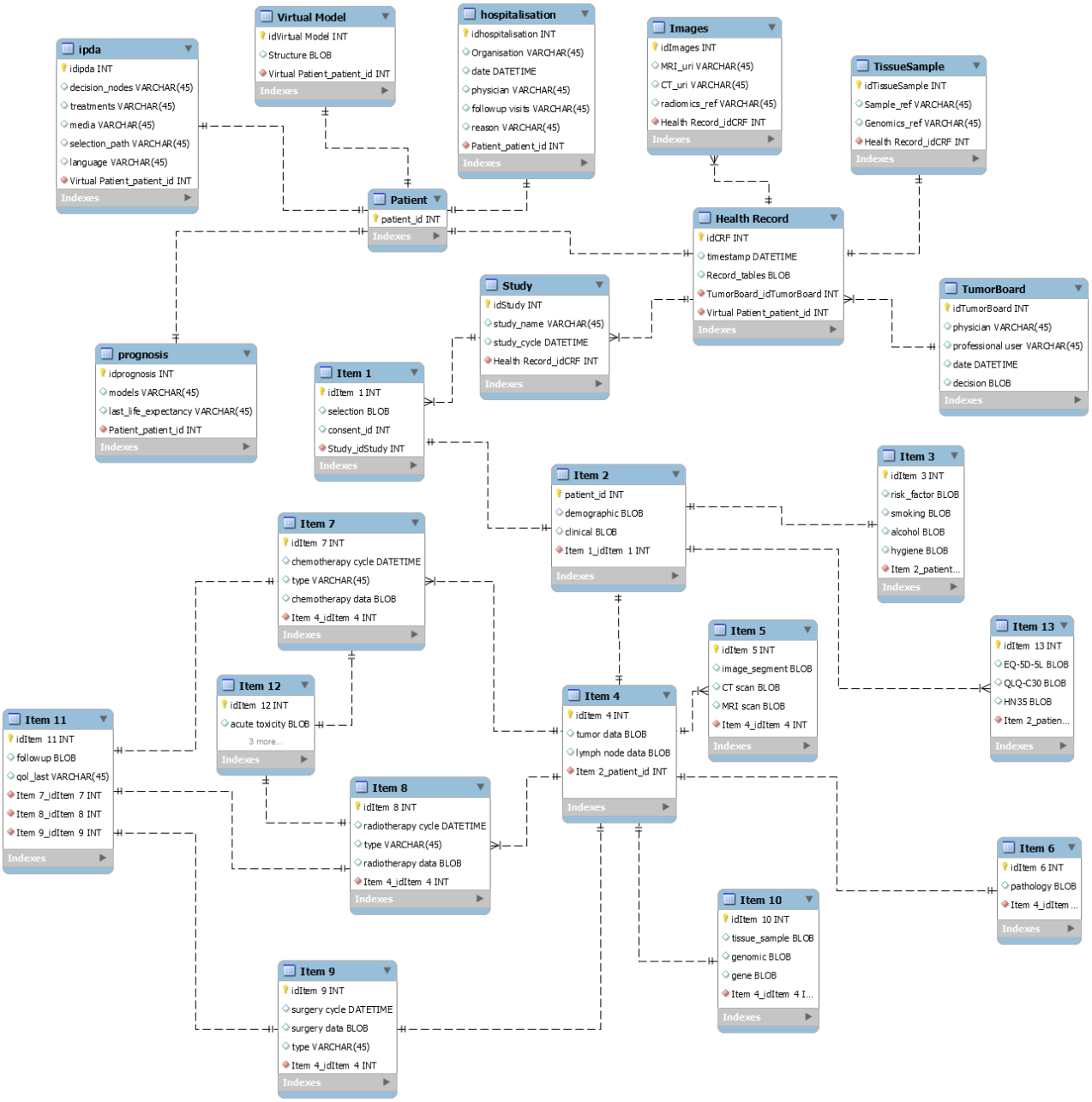


Figure 14: The structure of the Patients' Documentation System (PDS).

The description of the PDS structure in Figure 14 is an initial approach to the PDS development, which is expected to be revised in the forthcoming architecture deliverable D2.3 by end of December 2016.

4.2.2 Tools and Technologies for the implementation of PDS

The implementation of PDS is based on widely known technologies for data access and storage in distributed systems.

MySQL[7] is considered to be the most popular object relational database system. The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation.

Free-software-open source projects that require a full-featured database management system often use MySQL. MySQL is the M in the LAMP acronym, which stands for Linux-Apache-MySQL-PHP and refers to the open source technologies used by numerous frameworks and applications. MySQL works on many different system platforms, including AIX, BSDi, FreeBSD, HP-UX, eComStation, i5/OS, IRIX, Linux, Mac OS X, Microsoft Windows, NetBSD, Novell NetWare, OpenBSD, OpenSolaris and others. It is a very mature system, very good documented and with a wealth of resources available. Many graphical interface applications for the management and administration of a MySQL database are available, both free and paid.

Although MySQL was the database of choice for almost all open source projects, the acquisition of Sun by Oracle, which brought MySQL under Oracle's control, made many developers sceptical about using it in new projects. Although Oracle promised to continue supporting the database system and keep offering the community edition, it is generally thought that a sudden policy change from Oracle's part is not to be excluded. For this reason PostgreSQL is considered as an alternative for open source projects.

PostgreSQL[8] is a powerful, open source object-relational database system (also used as the storage layer of OpenClinica tool in Section 4.1.2). It has more than 15 years of active development and a proven architecture that has earned it a strong reputation for reliability, data integrity, and correctness. It runs on all major operating systems, including Linux, UNIX (AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64), and Windows. It is fully ACID compliant, has full support for foreign keys, joins, views, triggers, and stored procedures (in multiple languages). It includes most SQL:2008 data types. It also supports storage of binary large objects, including pictures, sounds, or video. It has native programming interfaces for C/C++, Java, .Net, Perl, Python, Ruby, Tcl, ODBC, among others, and exceptional documentation.

PostgreSQL SQL implementation strongly conforms to the ANSI-SQL:2008 standard. It has full support for subqueries (including subselects in the FROM clause), read-committed and serializable transaction isolation levels. And while PostgreSQL has a fully relational system catalogue, which itself supports multiple schemas per database, its catalogue is also accessible through the Information Schema as defined in the SQL standard.

It is released under a liberal open source license: the PostgreSQL License (MIT-style license) and is thus free and open source software. As with many other open-source programs, PostgreSQL is not controlled by any single company — a global community of developers and companies develops the system.

[7] http://www.mysql.com/

[8] http://www.postgresql.org

MongoDB[9] is a free and open-source cross-platform document-oriented database. It is a popular NoSQL database solution. NoSQL refers to a broad class of database management systems that differ from classic relational database management systems (RDBMSes) in some significant way. These data stores may not require fixed table schemas, usually avoid join operations and typically scale horizontally. NoSQL databases provide no SQL interface and usually rely on much simpler interfaces that use associate arrays or key-value pairs. Besides their simplicity, a big advantage of many NoSQL technologies is that they use a distributed architecture, which allows them to easily be deployed in a cloud system. A NoSQL system can be deployed in many servers and a failure of a server can be tolerated.

MongoDB shuns the relational database's table-based structure to adapt JSON[10]-like documents that have dynamic schemas which it calls BSON. This makes data integration for certain types of applications faster and easier. MongoDB is built for scalability, high availability and performance from a single server deployment to large and complex multi-site infrastructures[11]. Also MongoDB through automatic sharding enables data in a collection to be distributed across multiple systems for horizontal scalability as data volumes increase[12].

MongoDB can run in almost all operating systems and binaries are available for Windows, Linux, OS X and Solaris. MongoDB uses JavaScript for making queries. It offers official drivers for all popular programming languages (C, C++, C#, Haskell, Java, Javascript, Perl, PHP, Python, Ruby, Scala).

Object Relational Mapping (ORM) is not a storage component rather than a mechanism to map persistent objects stored in the Database to Java objects used directly by Java code. The ORM is a technique that can be approached through plenty of implementations. Hibernate[13] is the most popular solution, which is an ORM library implemented for the Java programming language, providing the framework for mapping an object oriented domain model to a traditional relational database.

4.3 Imaging analysis and storage

The images used in the assessment of a cancer diagnosis incident and the follow up process are maintained by the clinical centres and are processed subject to the procedures defined by each centre and the professionals' community as well. In BD2Decide, we develop software, which analyses the images and extracts features, which can then be used in the Big Data Infrastructure for making prognosis. Therefore, in this section, we describe the format of the feature data extracted from the imaging software.

[9] https://www.mongodb.com/

[10] http://www.json.org/

[11] https://www.techopedia.com/definition/30340/mongodb

[12] http://searchdatamanagement.techtarget.com/definition/MongoDB

[13] JBoss Community – Hibernate: http://www.hibernate.org/

4.3.1 Imaging software data formats

This section is divided into three parts, each describing the data output produced from the three different software solutions developed in the BD2Decide project and in the context of WP3, namely the Fraunhofer image analysis software, the CT radiomics software from MAASTRO and the MRI radiomics software from POLIMI.

Fraunhofer image analysis tool

The Fraunhofer image analysis tool will be used inside the clinical centres to support the clinicians during the image feature extraction process and the segmentation of the tumor and suspicious lymph nodes. Furthermore, it automatically extracts features from the image data and allows the clinician to manually access features. The extracted data can be stored in different formats, as explained in the following.

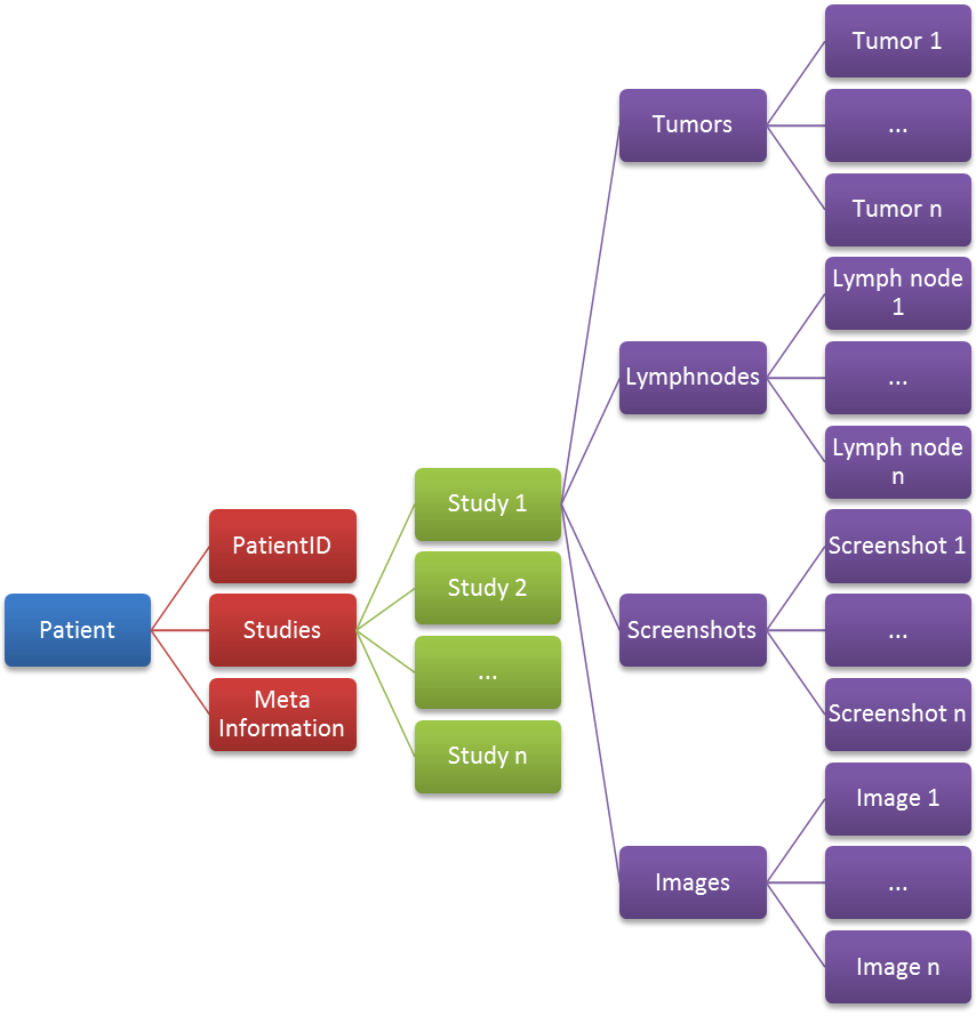


Figure 15: The structure of the image JSON-based file for the extratction of features.

The extracted features are stored in JSON file format. The JSON standard is among the most used formats for data exchange and has the huge advantage, that it is human readable. The document contains all information available about the patient from the medical image data. It is tree-based therefore a single document can hold the information for multiple studies at once. Each file can hold different image modalities, all of them with their unique features (see Figure 15). The document can

be either uploaded as a single file to the BD2Decide environment or it can be transferred using a REST (Representational State Transfer) API[14], which is a again common web standard.

The segmentations of the images are stored in NRRD (nearly raw raster data) file format[15], which is widely used file format in medical image data. The data sets contain some meta information in the file header and are followed by binary data, which represents the actual voxel data of the segmentation and can be stored in any database which accepts binary data. Additionally, the segmentations are stored in DICOM-RT[16], which is the clinical standard for annotations in radiation therapy. It is compatible to the DICOM standard and therefore can be stored on a clinical PACS server alongside the original image data.

These two different file formats will be supported, because the radiomics software from MAASTRO and the radiomics software from POLIMI expect different input data for the segmentations.

POLIMI MRI radiomic software

The radiomics tool provided by POLIMI uses the medical image data in DICOM format and the NRRD segmentations created by the Fraunhofer image analysis tool as input to determine the radiomic features. The extracted features consist of tag (feature name) and value (calculated feature value) and are stored in a comma separated values CSV file, which is a widely used standard format. Thereby, it can be easily transferred and inserted into any database.

The radiomic features produced by the POLIMI MRI radiomic software have been presented in Annex II of D2.1 about the patient's data and the expected BD2Decide HER (Item 5).

MAASTRO CT radiomic software

The workflow of the MAASTRO software is similar to the POLIMI solution. The main difference is, that the segmentations exported from the Fraunhofer software are in the DICOM-RT format. The extracted features are exported as a CSV file as well. This would enable a common interface for sending the radiomic features of the two different software solutions to the BD2Decide environment and parsing them in order to be inserted into the big data infrastructure.

4.3.2 Technologies and Tools for storing the imaging features

The databases required to maintain the data produced by the different imaging software tools will be developed through known and widely used database technologies, like the ones described for the PDS in section 4.2.2.

4.4 Prognostic Models

4.4.1 High level description of the database structure

The database for the prognostic models is mainly comprised of:

[14] http://docs.oracle.com/javaee/6/tutorial/doc/gijqy.html

[15] http://teem.sourceforge.net/nrrd/

[16] http://www.dclunie.com/dicom-status/status.html

- A library of existing prognostic models for (treatment-specific) prognosis of patients with different types of H&N cancer.
- A collection of tools that can be used to update/calibrate the models, combine prognoses from different models, use prognostic models in a stepwise manner, including data from more expensive modalities only if necessary.

Prognostic models are functions that use patient-specific data (clinical, pathological, imaging and/or genomics) as their input. The output is a prognosis, which is specific to the patient.

The tools will make it possible to tailor the models to the country's, hospital or subgroup of the population. The structure of such a country-specific, hospital-specific, subgroup-specific model will be similar to the prognostic models in the library. These new models need to be stored in separate library, which is structured in as shown in Figure 16.

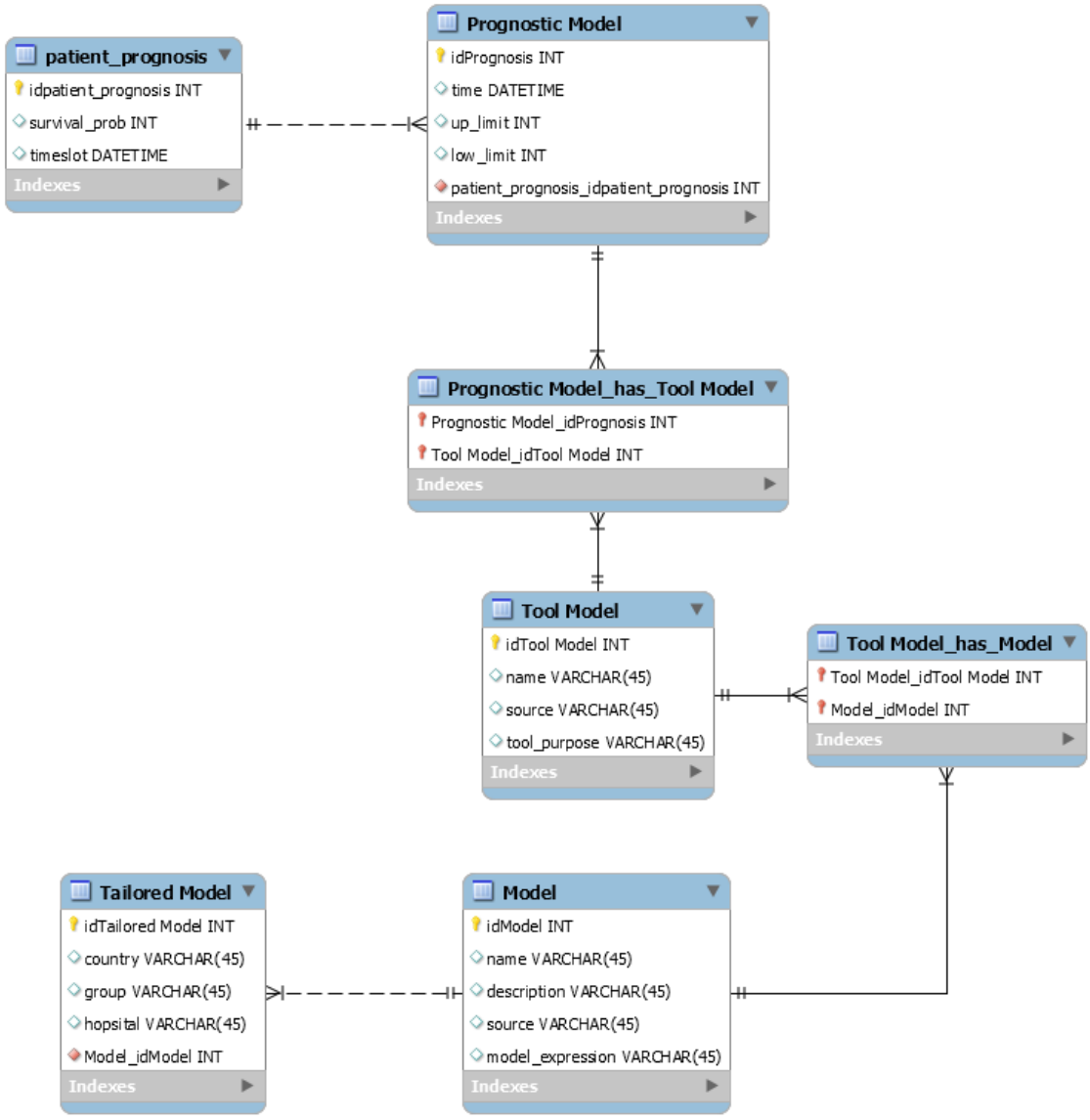


Figure 16: The high level structure of the prognostic models database.

Each patient-specific prognosis can be represented as a function of time and/or a vector with predicted survival probabilities for the patient for different time-points at follow up. Additional vectors are needed to store the lower- and upper limits of the prediction interval for the prognosis (to quantify the uncertainty in the prognoses). A prognosis can be presented to the user by means of a plot of the survival probability over time (with separate bands denoting the limits of the prediction interval). In addition, a table can be presented with survival probabilities at some pre-specified time-points for the follow-up process (e.g. 2 and 5 years) together with the prediction interval.

4.4.2 Technologies to support implementation and use of the library

The prognostic models library will be based on algorithms that will be implemented in R[17]. More specifically:

- The update/recalibration tools use as their input databases with EHR data (i.e. clinical, pathological, imaging and/or genomics data) for different patients and one or more prognostic models. The output is a new prognostic model.
- The synthesis tool uses different prognoses from the patient as its input. Its output will again be a new (pooled) prognosis for the patient.
- The cost-utility tool will use as its input the patient-specific data from EHR, a set of prognostic models and (user-specified) cost for obtaining additional data for the patient with modalities, as imaging and genomics. The output will be a set of separate net-benefits (single numbers: utility minus costs) for each of the additional modalities.

4.5 Knowledge Base

A large set of data are processed within the BD2Decide platform, consisting of different types of inputs, sources and formats. Such data is based on patient inputs, population data and other external data sources. The data collected will be used within the BD2Decide infrastructure and by the different tools to be implemented.

As known, ontologies are very useful tools for modelling, mapping and standardizing data [5]. Nowadays, there are several ontologies involved in the domain of health [6][7], particularly in the subdomain of cancer [8][9] and in population data[18].

An added value of the ontologies is the use of logic and semantics, very useful in projects that handle large amounts of data, like BD2Decide. These features are essential for a rigorous definition and development of a knowledge management system (Knowledge Base Manage). The principal aim of creating an ontology in the BD2Decide project is based on the need for semantic interoperability, standardization and integration of data.

To build the BD2Decide Ontology, the first step consists of the result for the definition of the data that is involved in the project. One should also take into account other existing ontologies related to the areas of health, cancer and the environmental domain. As a general guideline, it has been

[17] https://www.r-project.org/

[18] https://bioportal.bioontology.org/ontologies/PCO

decided to include in the core ontology those concepts that are represented by the data that is defined in the e-CRF and the BD2Decide Deliverable on the user needs (see [1]), and would be recollected from the participating medical and clinical centres involved in the project.

To create a comprehensive model of data, existing ontologies must be used. Currently, there are many ontologies that define and model concepts related to the domain of health and cancer disease specifically, context, epidemiological and behavioural data, as well as issues related to literature references and publications. Examples of these ontologies are: SNOMED CT[19], ENVO[20], pMedicine[21], NEOMARK[22], DOID[23], ICD10[24], BIBO[25] and CTO[26]. For this reason, one of the most efficient and recommended solutions, in order not replicate work and previous research, is based on the mapping of ontologies and the reuse of existing concepts and terms already defined.

To achieve this mapping, the main objective is to identify reuse concepts that are already defined in other ontologies, which will be included in the BD2Decide ontology. These concepts will be mapped or imported into the ontology. In addition, some concepts and terms with specific interest to the project, such as head and neck cancer denominations, models and tools developed during the project lifetime, will be, also, included in the BD2Decide ontology.

4.5.1 The structure of the BD2Decide ontology

The BD2Decide Ontology will be distributed into four main sub-domains, as illustrated in Figure 17Figure 1:

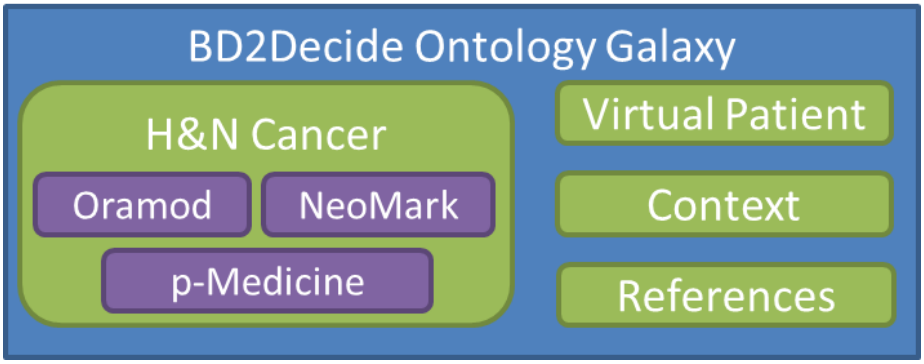


Figure 17: The structure of the BD2Ddecide Ontology.

[19] www.ihtsdo.org/snomed-ct

[20] https://bioportal.bioontology.org/ontologies/ENVO

[21] http://www.p-medicine.eu/

[22] https://datahub.io/tr/dataset/bioportal-neomark

[23] disease-ontology.org/

[24] http://apps.who.int/classifications/icd10/browse/2016/en

[25] http://bibliontology.com/

[26] https://bioportal.bioontology.org/ontologies

4.5.1.1 H&N Cancer

This section includes the concepts and definitions related to the clinical aspect of cancer disease. This sub-domain will be created through importing and mapping existing classes from the NEOMARK and p-Medicine ontologies.

NEOMARK and p-Medicine ontologies are used to define Head and Neck section into the BD2Decide Ontology. These ontologies involve terms and concepts more related to the data and characterization of the cancer diseases, and specifically with tumors and lymph nodes. The data include clinical, imaging, localization and pathological data, as well as the information related to prognosis, risk factors and treatment.

As an example, a set of classes that we are linked from ontologies are shown below:

- ***Tumor finding***: this class collects the concepts related to clinical information of tumor (Figure 18Figure 18).

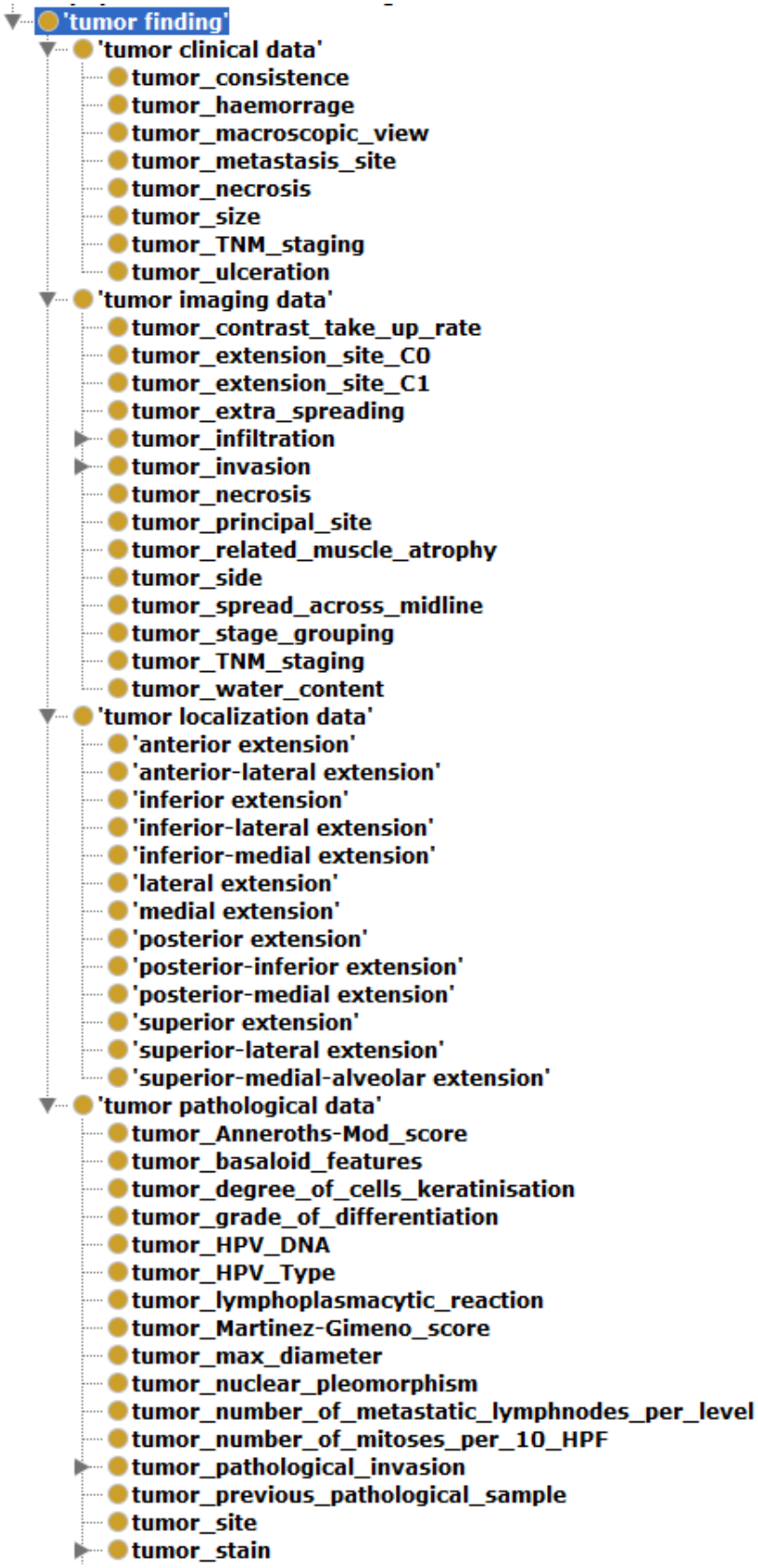


Figure 18: Tumor finding Class of Neomark Ontology.

- ***Lymph node finding***: this class collects the concepts related to clinical information of lymph node and specifically about clinical and imaging data (Figure 19).

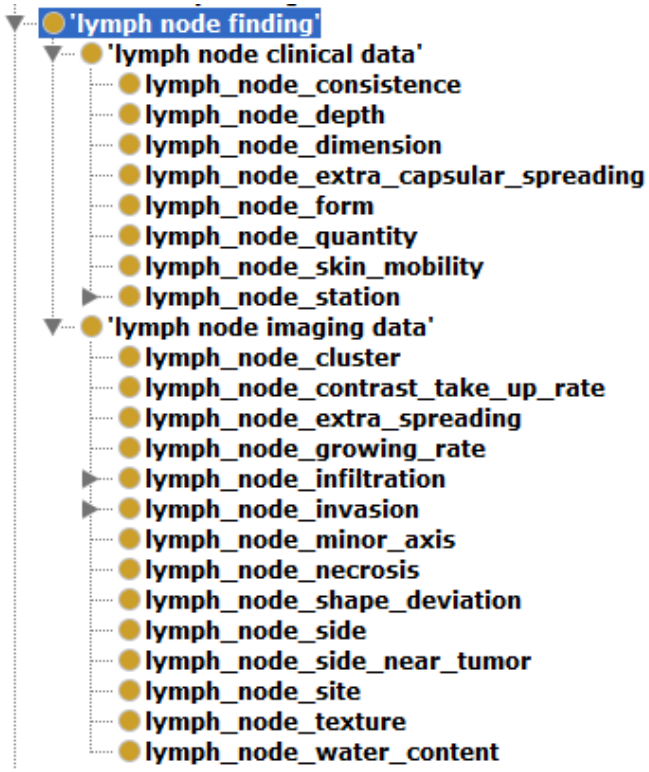


Figure 19: Lymph node finding Class of Neomark Ontology.

- ***Treatment***: this class represents specific concepts related to the surgical or non-surgical treatments for the patients (Figure 20).

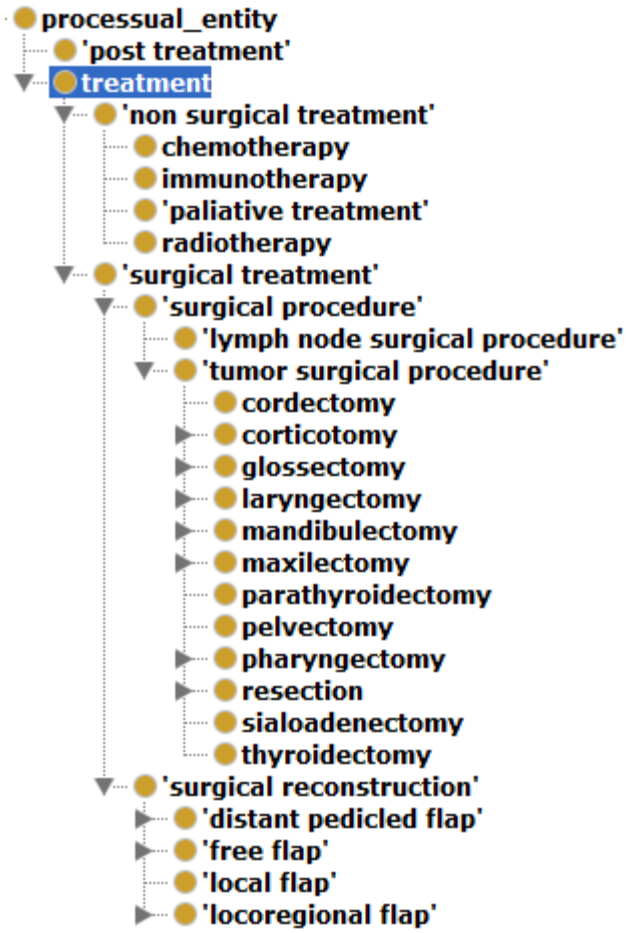


Figure 20: Treatment Class of Neomark Ontology.

- ***Prognosis***: this class collects the concepts related to the hypothesis of some future parts of a disease course (Figure 21).

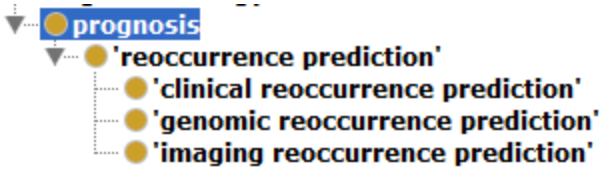


Figure 21: Prognosis Class of Neomark Ontology.

Additions to the H&NCancer Class:

Further concepts that can be added to this node are the following:

- ***Post treatment***: new concepts related to the post-treatment recommendations for the patient will be included. Recommendations can be related to rehabilitation, especially for speaking and/or swallowing, speech therapy, nutritional needs, addressing addiction and complementary therapies[27].

[27] http://www.headandneckcancerguide.org/adults/caring-for-someone-with-cancer/life-after-treatment/post-treatment-care/

- ***Follow-up***: new concepts related to data and planning of follow-up visits and patient recovery information will be included.

Inside the prognosis node, a set of data related to the results of the prediction models for the disease prognosis will be added.

In addition, new classes related to the imaging data of lymph nodes and tumor nodes will be upgraded, with the aim to include specific data related to MRI and CT images separately. The parameters defined in the CRF will be taken into account and included in this section (H&N cancer).

4.5.1.2 Virtual Patient

This section collects the concepts related to a patient, which is not depending on the follow-up (sort of "static" data). This sub-domain includes demographic and clinical data and risk factors presented for the patient. The information is collected once, usually during the first encounter or during the diagnosis.

Most classes defined in the Virtual Patient node come from patient data and risk factors that are defined in the CRF.

In addition, a set of Comorbidity classes from existing ontologies is also used:

- DOID Ontology – Human Disease Ontology: ontology that implements a comprehensive hierarchical controlled vocabulary for human disease representation.
- ICD10 – International Statistical Classification of Diseases and Related Health Problems - 10th revision.

Figure 22 depicts a set of variables defined within the Comorbidity classes.

BD2D_Virtual Patient

- ❑ Demographic Data
  - ♦ Gender (M/F)
  - ♦ Date of Birth
  - ♦ Ethnicity (White, Black or African, Asian, others)
  - ♦ Age at diagnosis
- ❑ Clinical Data
  - ♦ Date of first diagnosis
  - ♦ Diagnosis **(from addendum of CRF document and ICD10)**
  - ♦ Physical Examination
    - Weight, Height, Blood Pressure
  - ♦ ASA
    - I, II, III, IV, V, Not Available
  - ♦ Comorbidity **(from ICD10 and/or DOID Ontology)**
  - ♦ Overall Comorbidity Score according to ACE27
    - None, Mild, Moderate, Severe, Unknown
  - ♦ Laboratory Exam
    - HB level

- PLT level
- Lymphocytes

❑ Risk Factor
- ♦ Family history of malignancies
- ♦ Smoker
  - Smoker (Current, former, never, unknown)
  - Smoker habits (Cigarettes, cigar, pipes, betel quid, Smokeless (spit) Tobacco)
  - Passive smoker (Yes/No)
  - Packs per day
  - Years as a smoker
- ♦ Alcohol
  - Current, former, never, unknown
  - Alcohol habits
    - Units per day (liters)
  - History of alcohol dependence (Y/N/Unknown)
- ♦ Oral hygiene
  - Good, poor, bad
- ♦ HPV, Human papillomavirus
  - Status
  - Method of HPV testing
  - Date of HPV testing
- ♦ HIV, Human Immunodeficiency Virus
  - Status
  - Date of diagnosis
- ♦ Additional precancerous lesión
  - No, Leukoplakia, Lichen ruben planus, Erythroplakia, Oral submucous fibrosis
  - Location **(from ICD10)**
- ♦ Eating habits
  - salty foods, spicy foods, mate, betel quid, unknown
- ♦ Physical Agent
  - Ionizing radiation exposure
    - Type, hours of exposure
  - Substance
    - Wood dust, nickel powder, asbestos
- ♦ Drugs
  - Consume of marijuana

Figure 22: Virtual Patient Class of the BD2Decide Ontology.

4.5.1.3 Context

This section summarizes all the concepts related to the patient' context. It includes a set of conditions about the different environments and geographical or political regions where patients interact. These concepts define the environment where a patient lives and specifically the

environment where a patient works and spends his/her leisure time. In order to define this sub-domain, concepts defined previously in the SNOMEDCT and ENVO ontologies must be linked.

Figure 23Figure 8 shows the Environment or geographical location class.

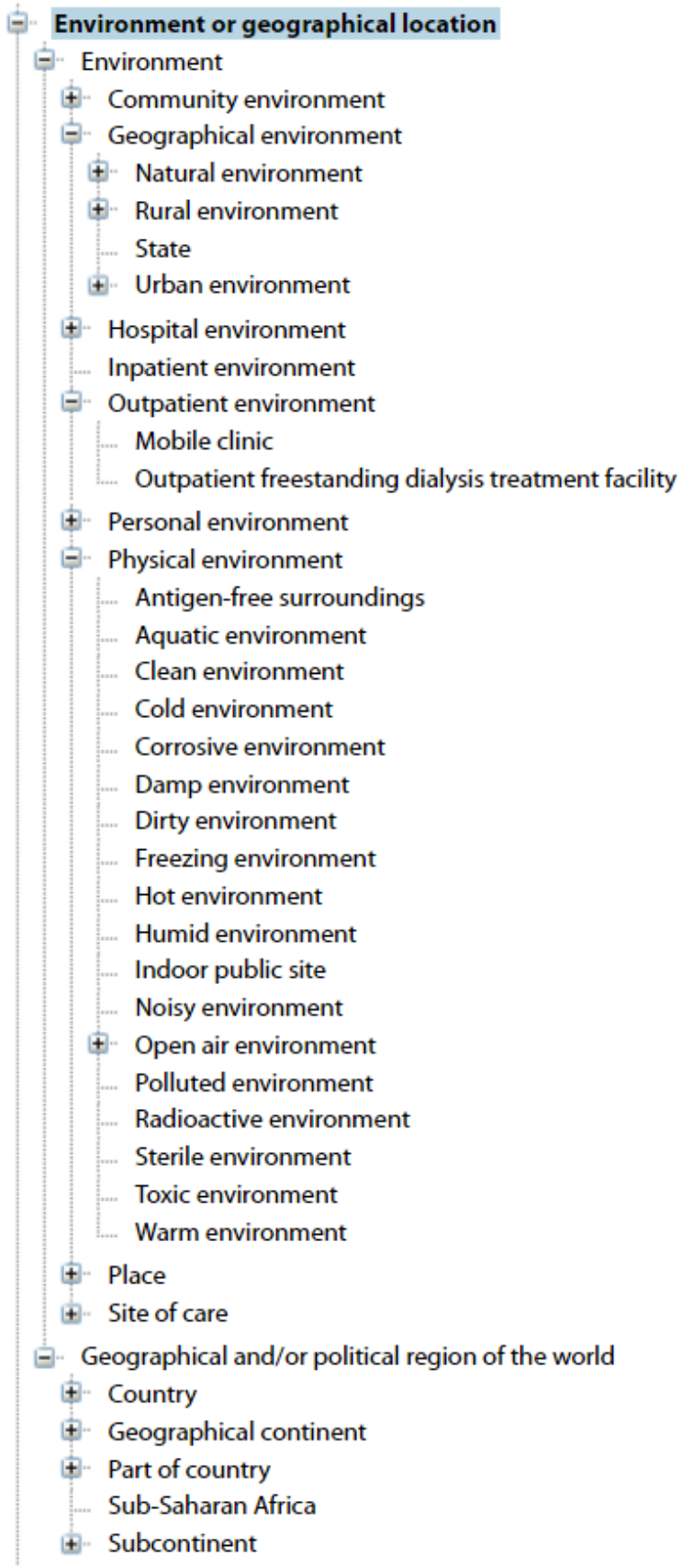


Figure 23: Environment or geographical location Class of SNOMEDCT Ontology.

4.5.1.4 References

This section collects all the data related to the clinical guidelines, treatments, scientific publications, technical publications with reference to head&neck cancer, results of clinical studies and all the literature for health and specifically for head&neck cancer disease.

Figure 24, Figure 25, Figure 26Figure 9 and Figure 27 show respectively the class References, the BIBO Ontology classes, the CTO Ontology classes and the InformationObject Class of ACGT-MO Ontology.

BD2D_References

- Title
- Authors
  - GivenName, FamilyName, Institution
- Date
- Rating
- Publication date
- Publisher
- Keywords
- Type of reference
  - Clinical Guidelines
  - Scientific papers
  - Technological papers
  - Clinical study
    - Type (Retrospective, Prospective)
    - Number of patients
- Structure
- Results
- Conclusions
- Bibliography

Figure 24: References Class of BD2D Ontology.

In order to define this sub-domain, a set of concepts defined previously in existing ontologies and related to bibliography and clinical trials will be linked. The ontologies include: BIBO Ontology (Bibliographic Ontology), CTO Ontology (Clinical Trials Ontology) and ACGT-MO Ontology (Cancer Research and Management ACGT Master Ontology).

Figure 25: BIBO Ontology Classes.

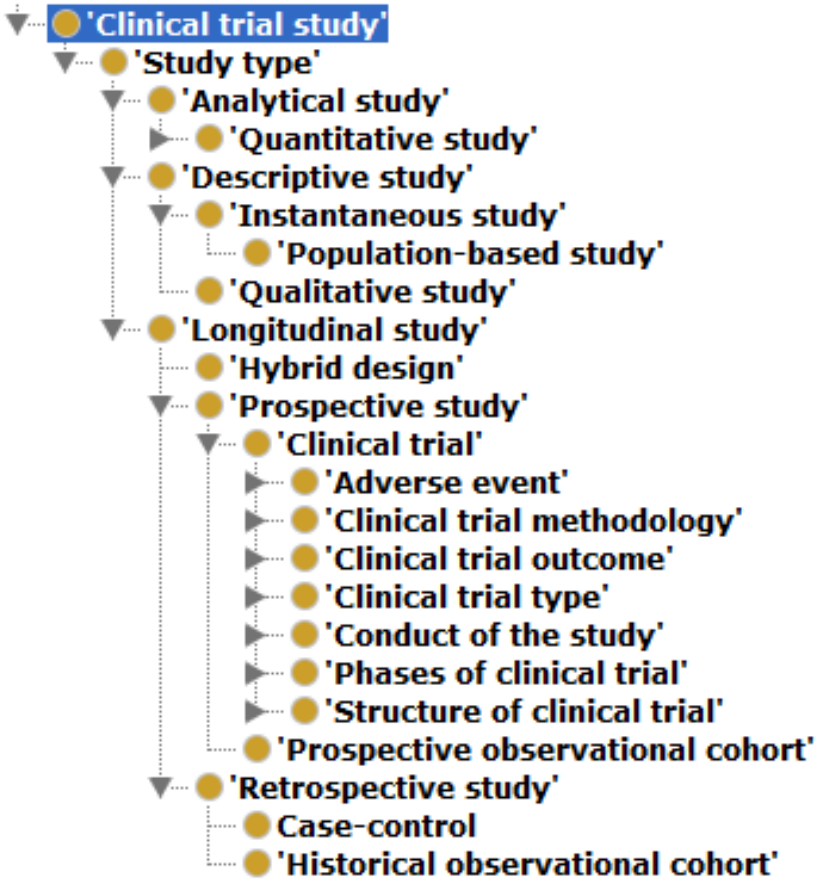


Figure 26: CTO Ontology Classes.

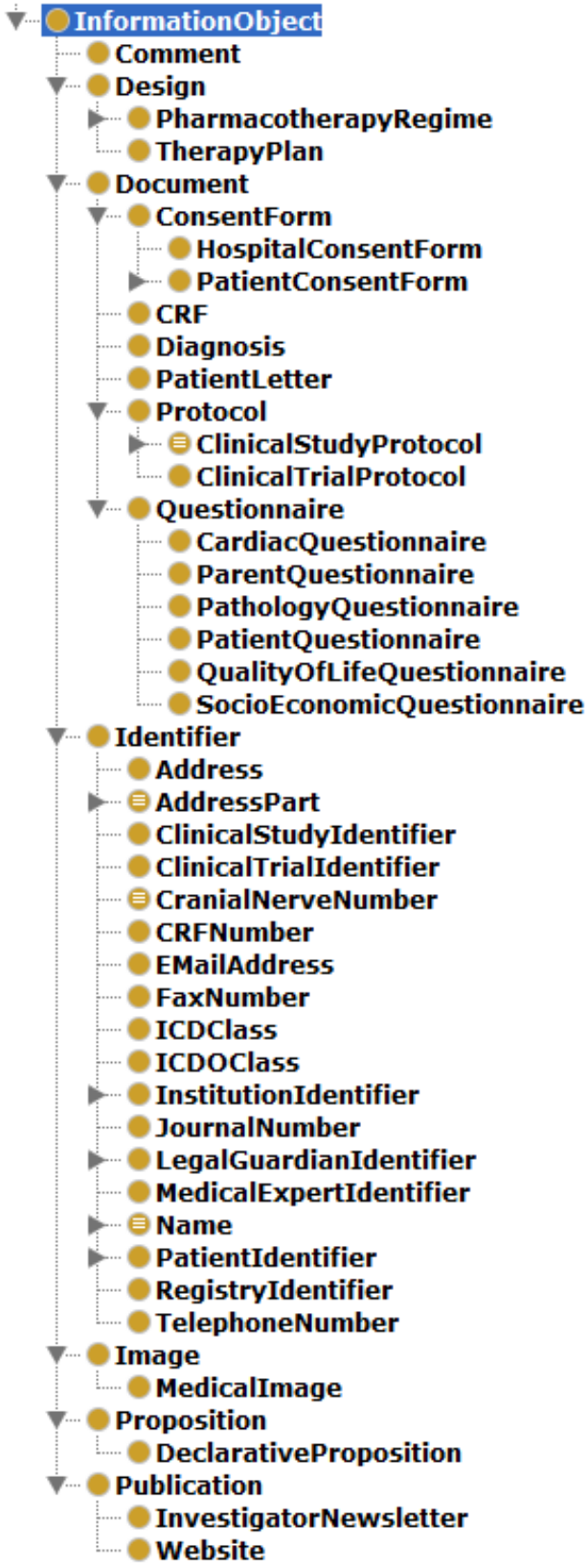


Figure 27: InformationObject Class of ACGT-MO Ontology.

4.5.2 Tools and Technologies

Taking into account the needs and characteristics of the BD2Decide project, it is mandatory to use semantic databases in order to avoid problems related to scalability, data abstraction, connection between different terms and concepts (semantic logic) and semantic overload. Such databases allow interoperability between the different tools of the platform.

Ontologies and semantic databases require a logical and formal language to be expressed. Thanks to the use of this language, a high degree of expressiveness and use will be intended.

One of the most important ontologies languages is the Web Ontology Language (OWL)[28]. This represents a semantic tagging language for publishing and sharing ontologies on the Web. It can be used to represent ontologies explicitly, that is, to define the meaning of terms in vocabularies and the relationships between those terms (ontologies). OWL is structured in layers that differ in complexity and can be adapted to the needs of each user, the level of expression that is required and to different types of existing applications.

OWL is designed to be used by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates machine interoperability of Web content supported by XML, RDF, and RDF Schema (RDF-S), by providing additional vocabulary along with a formal semantics[29].

In order to design the BD2Decide ontology and mapping concepts extracted from others ontologies, the framework PROTÉGÉ[30] will be used. Protégé is a free, open-source platform that provides a growing user community with a suite of tools to construct domain models and knowledge-based applications with ontologies. Protégé uses the languages RDFS, OWL and XML Schema to develop ontologies and build knowledge-based solutions in areas as diverse as biomedicine, e-commerce, and organizational modeling. It is one of the most popular ontology editors and frameworks for building intelligent systems.

With respect to the BD2Decide ontology, two different solutions are presented:

1. Solution based on Semantic Web framework:

The use of semantic databases, supporting the features and benefits generated by an ontology, will be created in the BD2Decide project. The following technologies must be taken into account:

- JENA[31]: an open source of Semantic Web framework for Java. It provides an API to extract data and write to RDF graphs. The graphs are represented as an abstract "model". A model can be sourced with data from files, databases, URLs or from a combination of these. Jena provides support for OWL. The framework has various internal reasoners based on inference rules.

[28] W3C Web Ontology Language (OWL): https://www.w3.org/OWL/

[29] http://www.w3.org/TR/owl-features/

[30] http://protege.stanford.edu/

[31] http://jena.sourceforge.net/

- Sesame[32]: an open-source framework for querying and analysing RDF data. Supports SPARQL language and have an API that allows the mapping of Java classes into the ontologies and generating Java source files from ontologies.

2. Solution based on No-Relational Database:

This approach is based on using non-relational databases, by exporting the concepts defined in the ontology to a database of non-relational model data. Examples of these databases are:

- OrientDB33: a 2nd Generation Distributed Graph Database with the flexibility of documents in one product. It is a unique, true multi-model DBMS equipped to tackle today's big data challenges and offers multi-master replication, sharding as well as more flexibility for modern, complex use cases. OrientDB includes the concept of "classes", so from the defined ontology it is possible to generate NoSQL classes. OrientDB is not a semantic database so it is impossible the use of reasoning which is defined in the ontology, but being a database based on graphs is possible to use 90% of the logic defined in the BD2D ontology. Also OrientDB includes SQL among its query languages along with a custom SQL based language which reduces the learning curve for those new to OrientDB.
- The MongoDB solution, which was described in Section 4.2.

4.6 The databases for population-based data

The population-based databases are extracted from various Internet sources with the aim to advance the prediction capabilities of the prognostic models and enhance their precision for the cancer cases under inspection. As such, this type of data is retrieved and is exploited in the BD2Decide project to optimise the personalised decision making, according to statistical information.

As presented in Section 2.3 and Figure 2, the population-based data can refer to epidemiology, lifestyle behaviour, health, medication and environmental factors. Thus, in this section we describe the logical structure of the respective datasets, which are involved in the accomplishment of the workflow shown in Figure 5, giving emphasis on the fact that the scope of accessing such kind of data is to investigate the occurrence of an incident for a particular population, which is defined mostly on demographic criteria.

4.6.1 Description of the data structure

The population-based data collected from the Internet are maintained within the BD2Decide environment, in order to be used in the prognostic prediction process.

For the epidemiology data, the primary concept lies in the diagnosis of a cancer case. This drives the collection of the remaining statistical data, which is presented in Figure 28. In more details, a diagnosis case can be categorised per tumor case, geographical area, age group and smoking habits to allow the reception of statistical information on the following treatment process and the resulting

[32] http://www.openrdf.org/

[33] http://orientdb.com/orientdb/

vital status. The database allows the professionals to track the trend for each diagnosis case and request for comparison with the case of a patient under investigation.

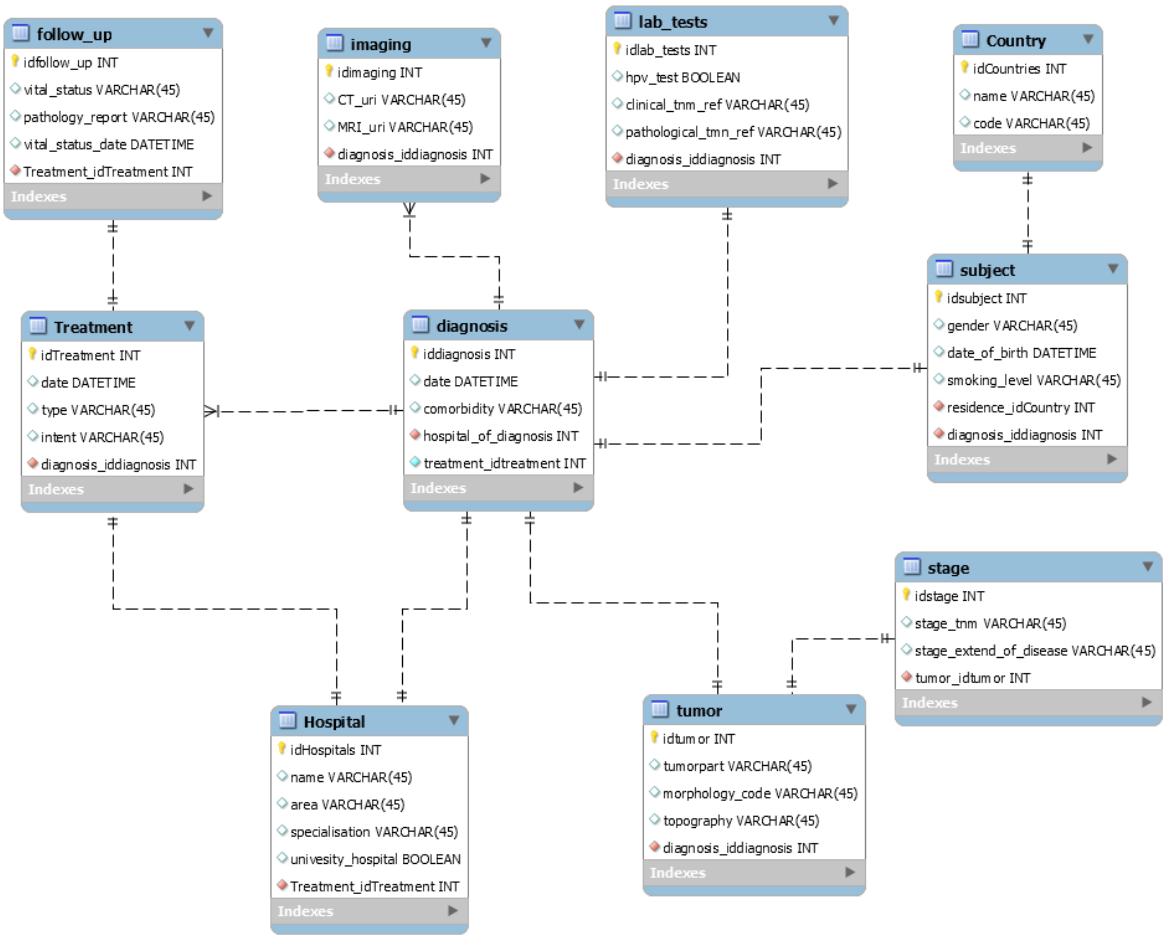
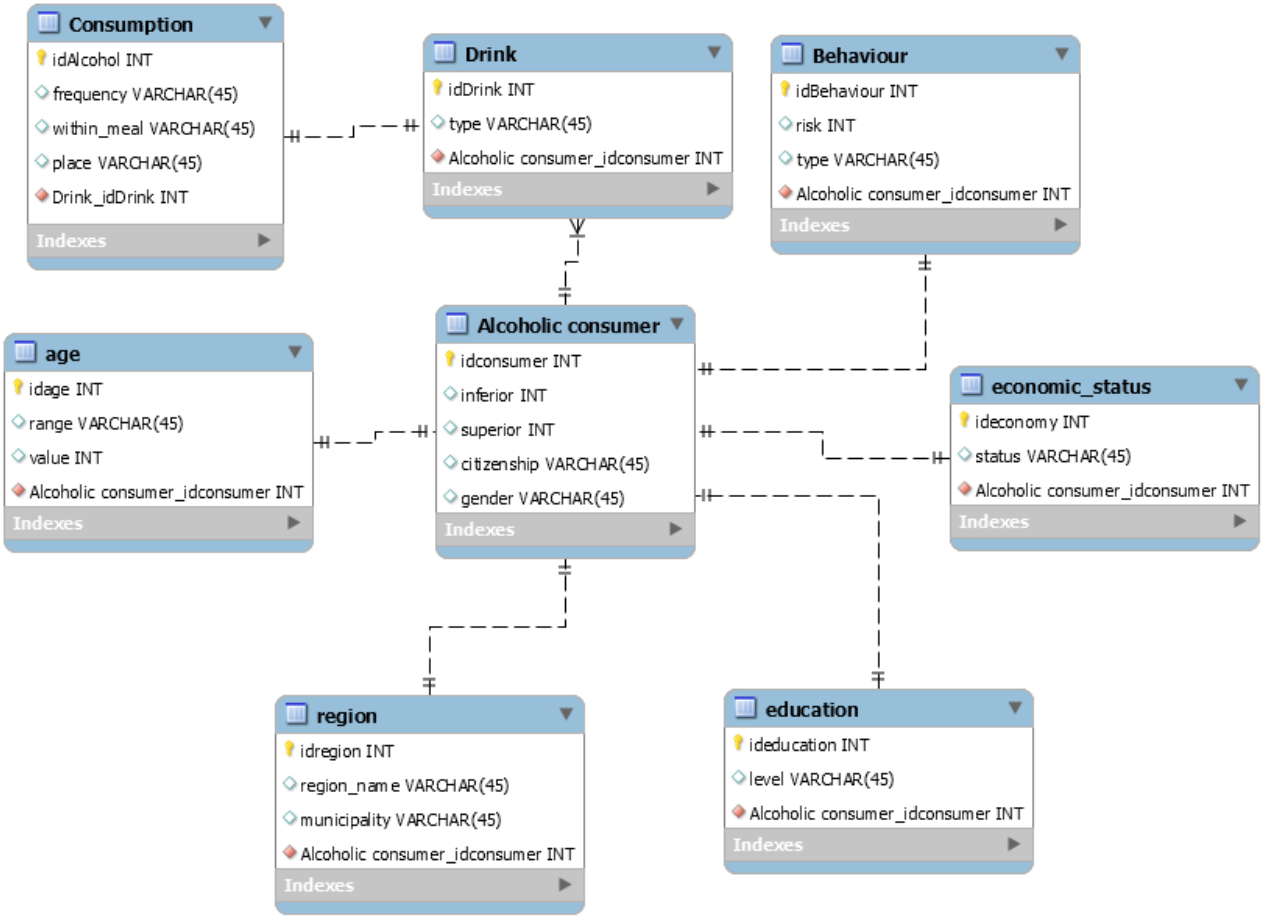
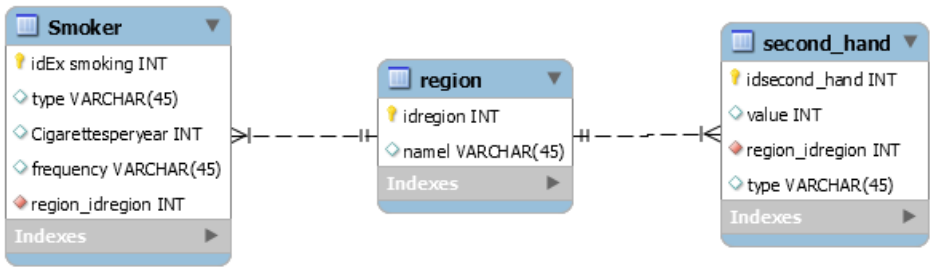


Figure 28: The structure of the epidemiology data in the BD2Decide project.

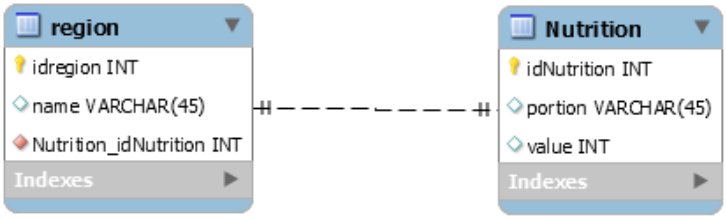
In the same sense, Figure 29 presents the structure for the lifestyle behaviour data streams, which is split into: a) datasets for alcoholic consumption per region, age range, gender, level of education, economic status and citizenship (see Figure 29 a), b) datasets for smoking habits (see Figure 29 b), and c) datasets for nutrition habits for the consumption of fruits and vegetables (see Figure 29 c).



(a) Alcohol consumption



(b) Smoking habits



(c) Nutrition habits

Figure 29: The structure of the lifestyle behaviour data in the BD2Decide project.

Then, Figure 30 presents the structure for the population-based health information data streams, which analyse data for diabetes and cardiovascular population and per region, age range, gender, level of education, economic status and citizenship and the overall perceived health status for the per region population.

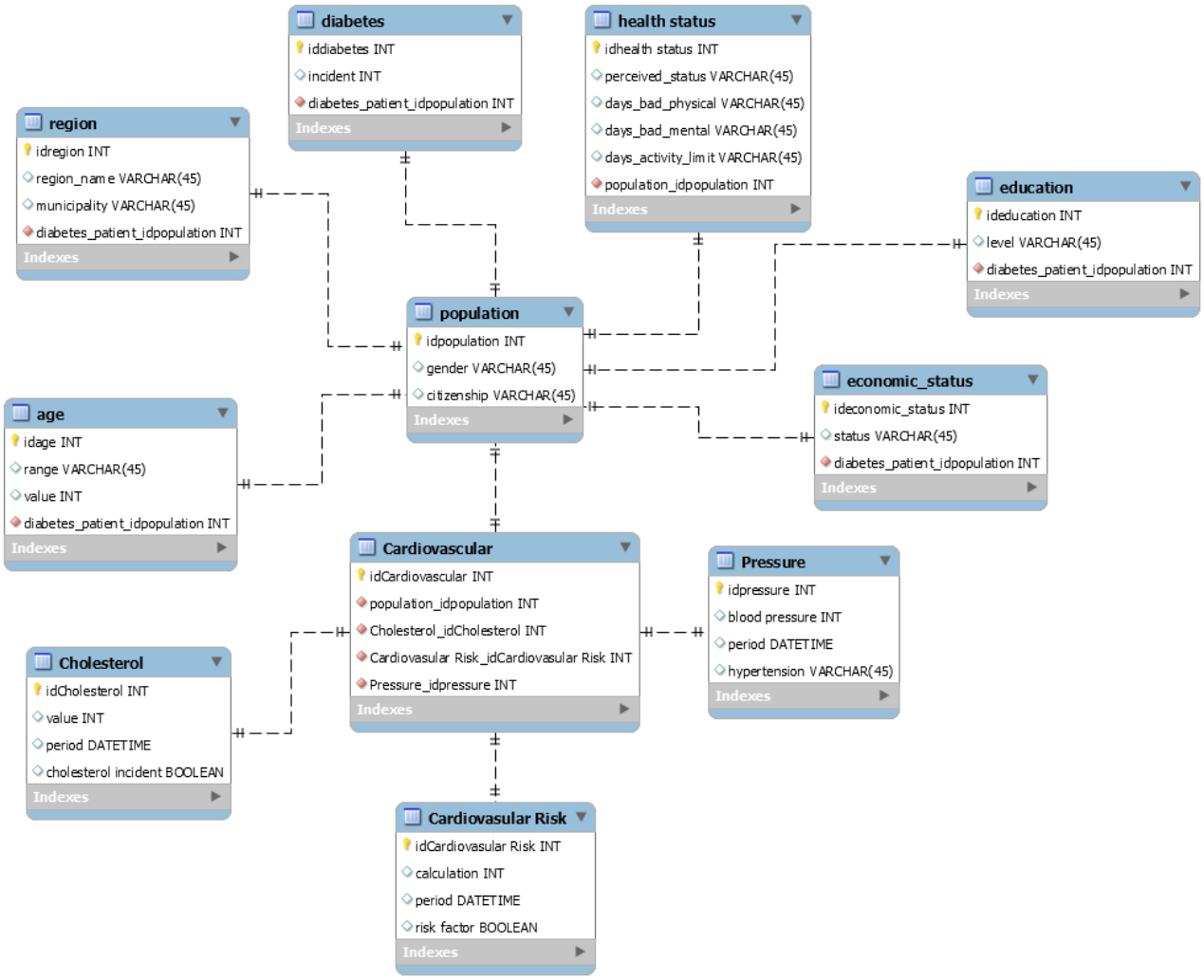


Figure 30: The structure of the health population database in the BD2Decide project.

Finally, Figure 31 describes the structre of the database for the environmental indicators used to assess the pollution per year in a certain country. This database will also be used in the prediction analysis and prognosis performed in the project.

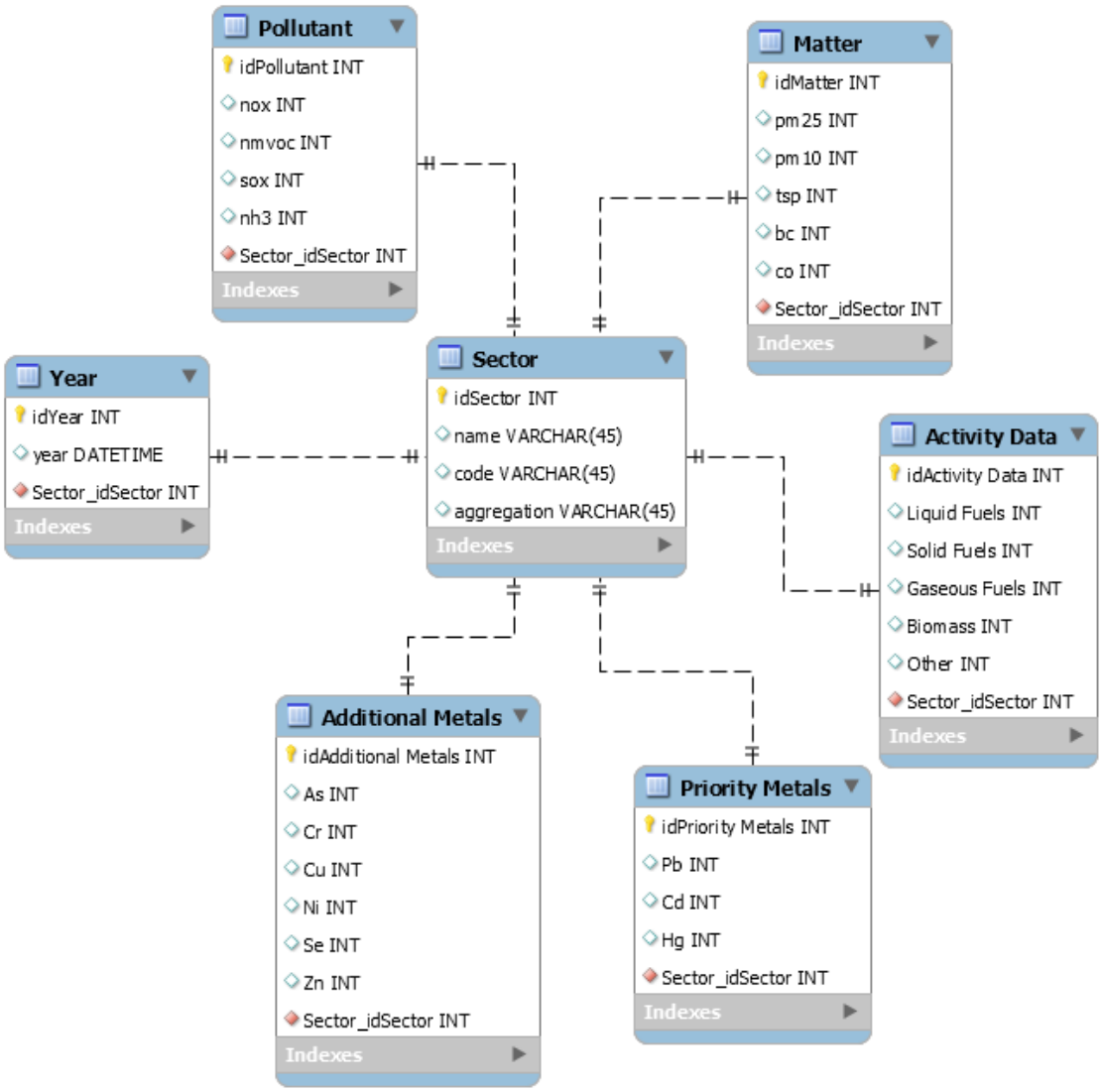


Figure 31: The structure of the environmental population database in the BD2Decide project.

4.6.2 Tools and technologies to present and exchange population-based data

The datasets presented in the previous section comprise open data made available to various official sources per country or on an EU level. In most of the cases, there is no unified and/or standardised approach to crawl or collect such data. Thus, in the BD2Decide project, we consider different approaches, depending on the data provider. In all cases, we focus on the use of open file formats[34], like JSON, XML, Spreadsheets, Comma Separated Files (CSV) or txt.

These datasets can be accessed and processed through a database technology (either an SQL like and a NoSQL one) or Web services (like REST or SOAP). In BD2Decide, we will consider both technologies to process the population-based datasets collected from Internet sources. Thus, the format of the dataset will be processed to create intermediate databases that can be accessed by the prognostic tools mainly, while a Web service technology will be used for the communication of the BD2Decide components with these databases.

[34] http://opendatahandbook.org/guide/en/appendices/file-formats/

4.7 Identity Management Database

Identity Management is a broad concept for the implementation of basic security attributes in software systems. As a concept, it integrates a lot of aspects which refer to the secure interaction of resources, being either human or computational resources, with one another, such as access to networks, services or applications. Identity management mainly refers to the identification of physical entities, but it also covers the distinction of computational resources as well. The functions involved in an Identity Management process include secure and private authentication, authorisation, trust management and user profile management. Identity management refers to the relationship of users to devices, networks, applications and services, but it may be realised as a single sign-on (SSO) process to service domains or a federated identity management process to similar application domains.

In BD2Decide, identity management is implemented mainly for the following purposes:

- Identify professionals as users in the BD2Decide environment;
- Identify patients as users in the BD2Decide environment;
- Define the policies under which a user has access to specific resources, being either BD2Decide services or data resources processed within the BD2Decide environment;

For each of these purposes, we exploit a set of data, which are maintained in the Identity Management database (IdM DB), as explained in the following section.

4.7.1 The structure of the IdM DB

The IdM DB is represented through the structure that is defined in Figure 32. As shown there, identity management refers to professionals, but the database connects them with patients in the PDS.

More specifically, within the BD2Decide environment, a professional has a specific profile and can be assigned one or many roles depending on their involvement in the accomplishment of the HNC process of Figure 1. The role assigned to a professional user may be attributed specific permissions to the functionalities of the CDSS, according to the policy implemented for this purpose. The latter defines the access rights of the role to the resources of the CDSS.

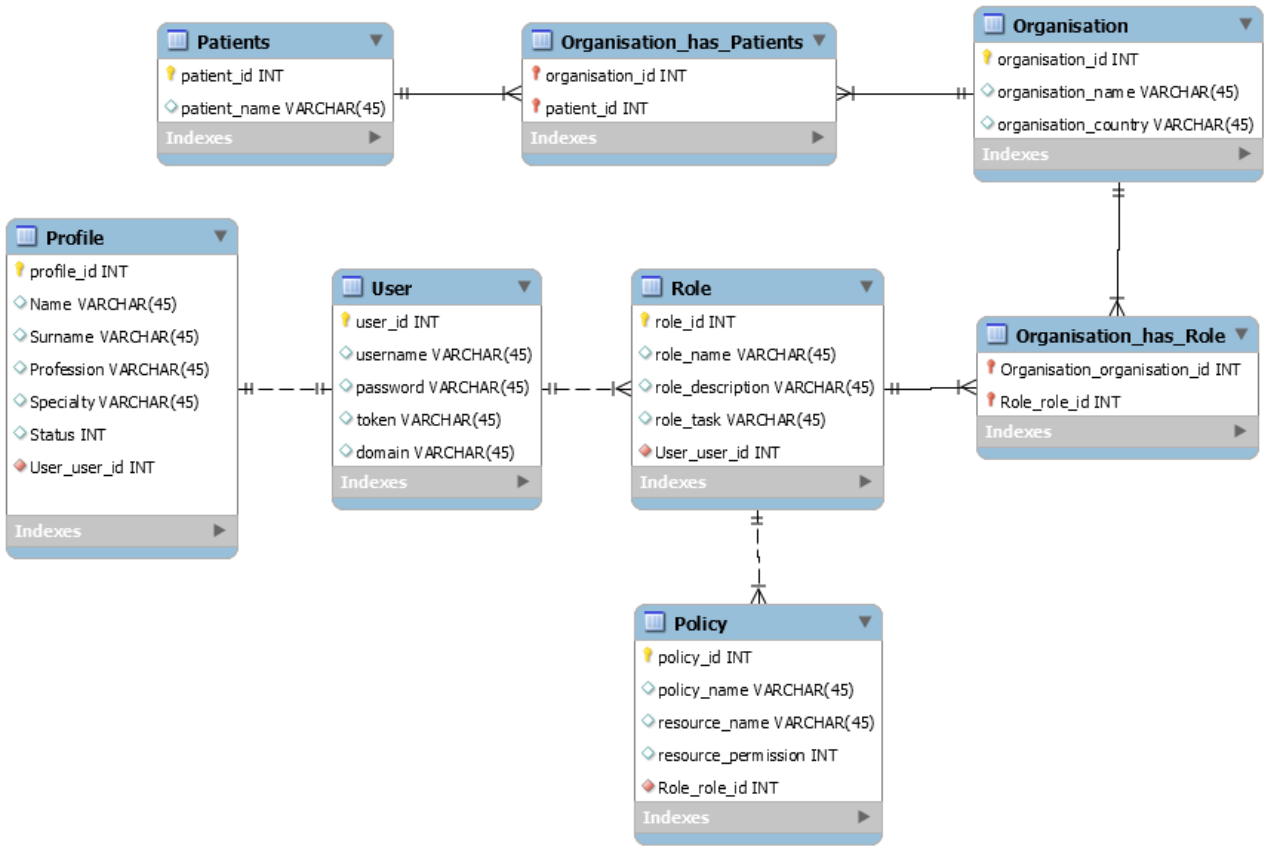


Figure 32: The structure of the Identity management database.

A professional user may belong to one or more clinical centres (organisations). Within each organisation, the user is assigned a role and based on this role, the user may be responsible for certain patients. The latter offers the connection of the professional to the PDS with the records of the patients participating in the BD2Decide environment.

4.7.2 Tools and technologies for the Identity Management record

Identity Management is developed through a set of open source technologies or proprietary solutions, depending on the level that this function is implemented within an ICT system. Thus, identity management can be implemented as an embedded mechanism or a remote service.

OpenID[35] facilitates for single-sign-on by providing mechanisms for both authentication and authorization processes. Both request and response message formats are defined to facilitate the transmission of necessary credentials within a Web service activity. It is decentralized standard, meaning that it is not controlled by any website or service provider.

OpenID allows use of an existing account to sign in to multiple websites, without needing to create new passwords. It requires a Web Application to be OpenID-enabled, meaning that the authentication process can be achieved through an external service provider.

[35] OpenID, http://openid.net/

Open Authorisation (OAuth)[36] is an open protocol to allow secure authorization in a simple and standard method from web, mobile and desktop applications. OAuth allows users to share their private resources stored on one site with another site without having to hand out their credentials, typically username and password. This is achieved through tokens, exchanged between service providers. Each token grants access to a specific site for specific resources and for a defined duration. This allows a user grant a third party site access to their information stored with another service provider, without sharing their access permissions or the full extent of their data.

Currently, OAuth is exposed on the Web through three versions, namely OAuth 1.0, 1.0a and 2.0. OAuth 2.0 is the next evolution of the OAuth protocol and is not backward compatible with OAuth 1.0. The OAuth 2.0 authorization framework is an IETF standard[37] for authorization, which enables a third-party application to obtain limited access to an HTTP service, either on behalf of a resource owner by orchestrating an approval interaction between the resource owner and the HTTP service, or by allowing the third-party application to obtain access on its own behalf.

It should be noted that OAuth is distinct to OpenID and, thus, differentiates from it by complementing the process of authentication and authorization in software and service distributed environments.

Recently, the FIWARE community[38] has released the open specification for identity management in the FIWARE wiki[39]. This specification describes the main contents of the identity management process as a generic enabling technology for the implementation of security in Future Internet applications. A reference implementation is, also, provided in the FIWARE Catalogue[40] under the Apache 2.0 license.

[36] Open authorisation: http://oauth.net

[37] https://tools.ietf.org/html/rfc6749

[38] https://www.fiware.org/

[39] https://forge.fiware.org/plugins/mediawiki/wiki/fiware/index.php/FIWARE.ArchitectureDescription.Security.Identity_Management_Generic_Enabler

[40] http://catalogue.fiware.org/enablers/identity-management-keyrock

5 CONCLUSIONS

This deliverable presented the results of Task 5.1 of the BD2Decide work plan for the specification of the patient's data warehouse. In that respect, the deliverable presented the BD2Decide data warehouse service architecture and the details of the respective data repositories, which host the data required in the BD2Decide to develop the clinical decision support system and improve the decision making process in the treatment of Head and Neck cancer incidents.

More specifically, the document provided the connection to the specification of the user needs and relevant use cases in D2.1and reported on datasets involved across the implementation of the workflow processes from the diagnosis of a head and neck incident, through the decision on the personalised treatment method to the assessment of the follow-up period. It, then, presented the three parts of the logical architecture of the data warehouse environment. These parts refer to: i) the acquisition of data and the respective services from the side of the clinical centres, ii) the analysis, visualisation and access control data maintained within the BD2Decide storage layer, and iii) the analysis and visualisation data processed in the BD2Decide big data infrastructure.

Through this deliverable, we presented a holistic view of the data acquisition and management services to implement the personalised decision making process in BD2Decide. The already developed prototype for data collection from the clinical centres will be further exploited in WP5 to develop the environment for the aggregation of the patients' health records in the PDS, subject to data privacy and security issues. The resulting database schema in this document will be further exploited in WP2 for the specifications of the BD2Decide architecture.

6 REFERENCES

[1] BD2decide Consortium, D2.1: User needs and use cases, May 2016.

[2] International Agency for Research on Cancer (IARC), World Helath Organisation, https://www.iarc.fr/

[3] ARPA (Agenzia Regionale per la Protezione dell'Ambiente) regions of Lombardía and Emilia Romagna

[4] Databank of air pollutants from traffic- Italy. Source ISPRA (Instituto Superiore per la Protezione e la Ricerca Ambientale)

[5] Bodenreider, O. (2008). Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. *Yearbook of Medical Informatics*, 67–79.

[6] Madurai N Meenachi and Sai M Baba. Article: A Survey on Usage of Ontology in Different Domain. *International Journal of Applied Information Systems* 4(2):46-55, September 2012. Published by Foundation of Computer Science, New York, USA.

[7] Smith Barry et al, The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature biotechnology*, pp. 25.11: 1251-1255., 2007.

[8] Isac, C., Viterbo, J., Conci, A. (2016). A Survey on Ontology-Based Systems to Support the Prospection, Diagnosis and Treatment of Breast Cancer. *XII Brazilian Symposium on Information Systems*, Florianópolis.

[9] Hua Min, Frank J. Manion, Elizabeth Goralczyk, Yu-Ning Wong, Eric Ross, J. Robert Beck. Integration of prostate cancer clinical data using an ontology, *Journal of Biomedical Informatics* Volume 42, Issue 6, December 2009, Pages 1035-1045, ISSN 1532-0464.

7 ANNEX

7.1 Openclinica

The Entity Relationship (ER) diagram of the OpenClinica database is presented in Figure 33. More details about it can be found in https://dev.openclinica.com/tools/db/relationships.html.

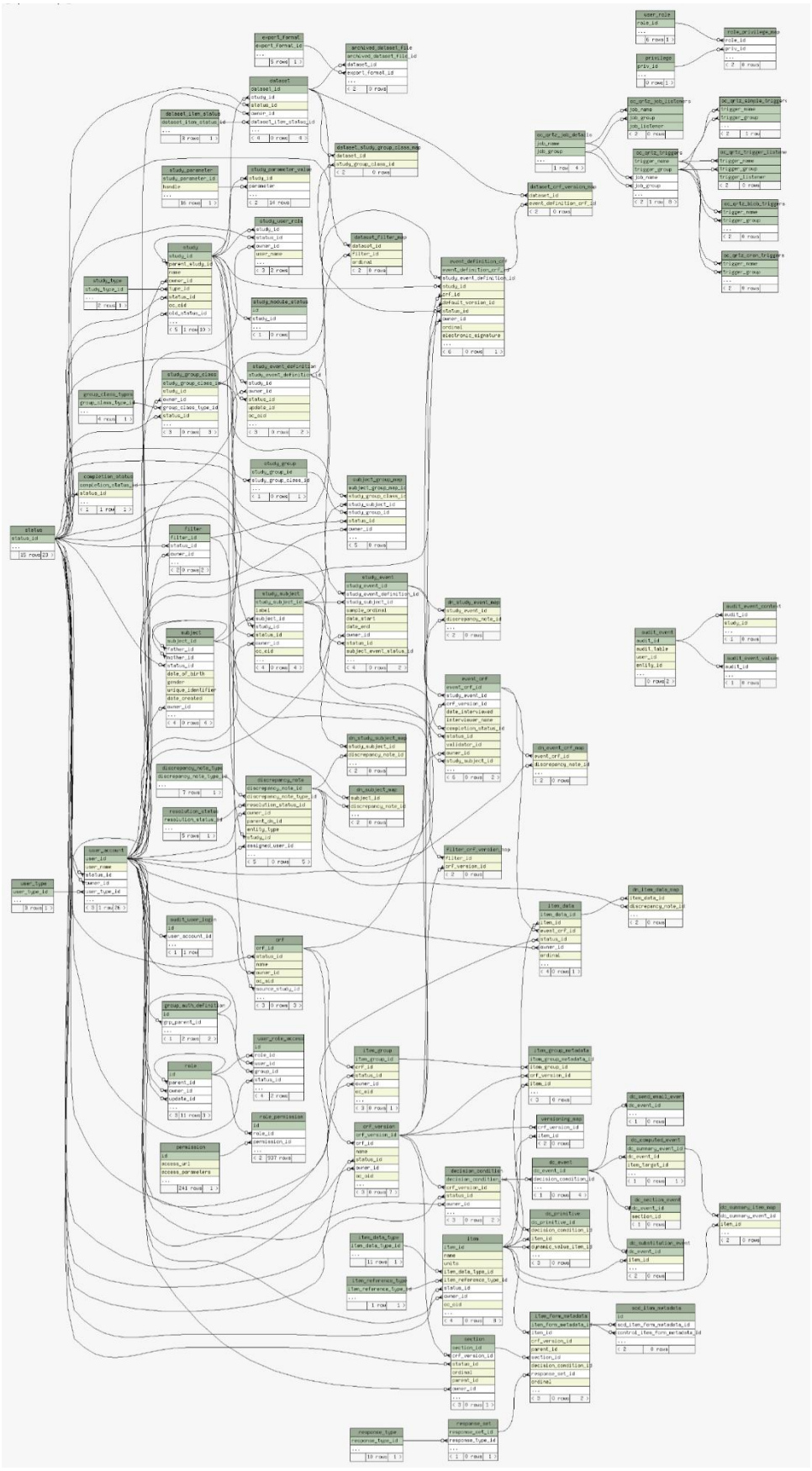


Figure 33: OpenClinica database schema.