Report on ContraCancrum data clinical studies according to the scenarios

Project Number:	FP7--IST-*223979*
Deliverable id:	D 9.2
Deliverable name:	Report on ContraCancrum data clinical studies according to the scenarios
Submission Date:	02/10/2010

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	CONTRA CANCRUM
Project Full Name:	Clinically Oriented Translational Cancer Multilevel Modelling
Document id:	D 9.2
Document name:	Report on ContraCancrum data clinical studies according to the scenarios
Document type (PU, INT, RE)	RE
Version:	1.1
Submission date:	02/10/2010
Editor: Organisation: Email:	Norbert Graf USAAR graf@uks.eu

Document type PU = public, INT = internal, RE = restricted

ABSTRACT:

ContraCancrum aims to develop a multilevel platform for simulating malignant tumour development and response to treatment. It takes tumour and the interaction to normal tissue into consideration. The project is clinically oriented and focuses on lung cancer and gliomas. As a clinically driven project it is essential to understand the workflow of data within clinical trials by all participants. The time frame of data sharing for the clinical scenarios is explained. To have the ability of reusing data in different scenarios an extension of the ACGT Master Ontology for lung cancer and brain tumours is justified. This implies a good cooperation with ACGT, an EU-project funded in FP6. Some members of ContraCancrum are also members of ACGT. This cooperation allows also the use of tools developed in ACGT. Most important will be ObTiMA for data management. This deliverable serves as a guideline for the workflow of data types needed in the project. Such guidelines are important for all participants to know where and how to store data, where to find and how to use data. Data flow will always be done according to legal and ethical regulations. The integrated scenarios of ContraCancrum are described in this deliverable including the validation of the scenarios.

KEYWORD LIST: Data, clinical studies, scenarios

MODIFICATION CONTROL			
Version	Date	Status	Author
0.1	01/07/2009	Draft	Norbert Graf
0.2	01/08/2009	Draft	Norbert Graf
0.3	15/08/2009	Draft	Norbert Graf
0.4	27/08/2009	Draft	Norbert Graf
0.5	28/08/2009	Draft	Norbert Graf
0.6	07.09.2009	Draft	Norbert Graf
1.0	21.09.2009	Draft	Norbert Graf
1.1	01.10.2009	Final	Steven Manos

List of Contributors

- Norbert Graf, USAAR
- Eckart Meese, USAAR
- Rainer Bohle, USAAR
- Christian Veith, USAAR
- Yoo-Jin Kim, USAAR
- Holger Stenzhorn, USAAR
- Georgios Stamatakos, ICCS
- Manolis Skounakis, FORTH
- Christina Farmaki, FORTH
- Ioannis Karatzanis, FORTH
- Alexandros Roniotis, FORTH
- Kostas Marias, FORTH
- Steven Manos, UCL
- Amos Folarin, UCL

Contents

1 Executive Summary

ContraCancrum aims to develop a multilevel platform for simulating malignant tumour development and response to treatment. It takes tumour and the interaction to normal tissue into consideration. The project is clinically oriented and focuses on lung cancer and gliomas. As a clinically driven project it is essential to understand the workflow of data within clinical trials by all participants. The frame of data sharing for the clinical scenarios is explained. To have the ability of reusing data in different scenarios an extension of the ACGT Master Ontology for lung cancer and brain tumours is justified. This implies a good cooperation with ACGT, an EU-project funded in FP6. Some members of ContraCancrum are also members of ACGT. This cooperation allows also the use of tools developed in ACGT. Most important will be ObTiMA for data management. This deliverable serves as a guideline for the workflow of data types needed in the project. Such guidelines are important for all participants to know where and how to store data, where to find and how to use data. Data flow will always be done according to legal and ethical regulations. The integrated scenarios of ContraCancrum are described in this deliverable including the validation of the scenarios.

2 Introduction

The purpose of this document is to give all members of ContraCancrum a guideline of data workflows in the project. The various data of the project are described and how data transfer is done. The workflows do start from data flow within clinical trials and ends in the use of data for validation of the scenarios. Post-processing of data is described if this is necessary before entering the tumour models. The time frame of data sharing for clinical scenarios is essential to guarantee the usability of validated models in daily clinical practice as decision making tools. As no such decision making tools based on in silico models do exist today in clinical practice this is an essential task to be focused on. There are further obstacles that have to be overcome like dissemination and training before such decision making tools will be used routinely in the daily care of patients.

3 Data within clinical trials

3.1 Medical Data to be Used for the Adaptation, Optimization and Validation of the Models of Clinical Tumour Growth and Response to Treatment

This chapter lists the various types of (pseudo)anonymized multiscale medical data to be provided by the participating clinical partners and used for the adaptation, optimization and validation of the imageable tumour models which are described mainly in Chapters 6 and 7 of Deliverable D4.1. A description of data is also given in D2.2&D9.1

3.1.1 Glioma (glioblastoma) cases

A1. TREATMENT DATA

Detailed description of the treatment (chemotherapy and/or radiotherapy) scheme including

1. Drug(s) names and/or type of ionizing radiation
2. Actual treatment session dates
3. Actual dose(s) for each treatment session
4. Any other relevant data (eg. individualized pharmacokinetic data if available)

A2. NORMAL TISSUE COMPLICATION DATA

1. Side effects that will be measured are in case of patients with gliomas mainly due to irradiation. To describe these side effects the Common Terminology Criteria for Adverse Events (CTCAE) v3.0 will be used[1].
2. Most important are skin reactions and neurological side effects
3. Haematological toxicity will only be analysed in patients who have received chemotherapy.
4. Especially in patients with gliomas who develop new neurological symptoms it is sometime impossible to distinguish between tumour progression and side effects of treatment. In this respect the simulation of normal tissue reactions has limited potential.

A3. CLINICAL DATA

1. Age
2. Sex
3. Description of eventual previous treatments
4. Any other pertinent clinical data

[1] https://webapps.ctep.nci.nih.gov/webobjs/ctc/webhelp/welcome_to_ctcae.htm

A4. IMAGING DATA

In the case of glioma the following imaging data will be made available:

1. T1 Gadolinium-enhanced MRI (dense coverage of the region of interest)
2. T2 MRI (dense coverage of the region of interest) and t2 flair
3. CT without contrast enhancement (Hounsfield scale as a quantitative measure of radiodensity as a reference)
4. Additional imaging data are desirable and will be used in the project if they are available for a patient. These imaging data are:
    i. Diffusion anisotropy MRI
    ii. Perfussion MRI
    iii. PET
    iv. SPECT
    v. fMRI
    vi. Flash
    vii. Spectroscopy (chemical shift imaging)

All imaging data sets correspond to a known time point before treatment initiation, eventually several time points during treatment (if this is possible) and at least one known time point after completion of treatment.

Information that is needed for the simulation model:

- Necrotic areas / cysts / vital tumour areas
- Ratio of vessels in the tumour (qualitative, quantitative?)
- Cell cycle distribution
- Anatomy of the brain is needed for the biomechanical point (distribution of grey and white matter, because of the different mechanical behavior) → normal brain (healthy brain) age adjusted ! is needed.
- Localization of neurons is important for biomechanical modeling
- Oxygen supply for radiotherapy response
- Perfusion studies (transport of nutrients) to the tumour
- Apoptosis and proliferation (mitotic rate)
- Glucose metabolism

Post-processing of images needs to be done to retrieve this information. Pre-existing tools will be used where possible otherwise the necessary tools will be developed in-house.

A5. SEGMENTATION DATA

1. External “conventional” boundaries of the tumour based on T2 MRI
2. External “conventional” boundaries of the tumour based on T1-gadolinium-enhanced MRI
3. Boundaries of the necrotic (“dark appearing”) region based on T1-gadolinium-enhanced MRI
4. Boundaries of any additional tumour structures lying within the previously mentioned “conventional” boundaries of the tumour (including their macroscopic characterization)
5. Desirably: boundaries of adjacent normal tissue structures with annotation.
6. Estimate of tumour volume (including calculation method)

In gliomas, especially in glioblastomas segmentation of the tumour is mainly impossible (see also D2.2 and D9.1). Therefore new methods for measuring the tumour boundaries, necrotic and cystic areas as well as the tumour volume have to be developed. For that purpose the histogram of signal intensities will be scientifically evaluated.

A6. HISTOPATHOLOGICAL DATA

In the case of glioma the following histological data will be made available:

1. Detailed description of the tumour subtype including grade and stage
2. Microphotographs of indicative histopathology slices. However, due to the surgical procedure in most cases it is impossible to correlate the histology to the region of the tumour from which the bioptic material has been extracted.
3. Estimates of the amount of dead cells (necrotic, apoptotic)
4. Estimates of the spatial tumour cell density for proliferative cells (Ki67 staining)
5. The degree of neovascularization for selected cases.

The microscopic imaging data are processed in such a way that the density of tumour cells as well as the percentages of proliferative cells (Ki67) and dead cells are calculated. In few cases, in which such data are available, a mapping of the bioptic areas and the imaging will be done. This will lead to a relatively refined consideration of the tumour spatial inhomogeneities in different areas of the tumour. In all other cases where such information is not possible to get the information from the biopsy will be extrapolated to the whole tumour, knowing that this is only a rough approximation.

A7. MOLECULAR DATA

In the case of glioma the following data will be made available:

1. An extended number of novel glioma expressed antigens reactive with patients' auto-antibodies will be identified.
2. The seroreactivity of antigens will be measured using a novel automated image analysis system.
3. Information of the reactivity of the immunogenic antigens against sera of healthy controls will be provided.
4. Each glioma antigen will be classified according to its reactivity against normal sera and sera of glioma patients. AUC values for the antigens will be determined to rank them according to their information content.
5. Cytogenetic data will be provided for those gliomas that can be used for short term tissue culture.
6. cDNA expression data will be provided for those gliomas that are available as frozen tissues.

3.1.2 Lung cancer cases

B1. TREATMENT DATA

Detailed description of the treatment (chemotherapy and/or radiotherapy) scheme including

1. Drug(s) names and/or type of ionizing radiation
2. Actual treatment session dates
3. Actual dose(s) for each treatment session
4. Any other relevant data (eg. individualized pharmacokinetic data if available)

B2. NORMAL TISSUE COMPLICATION DATA

1. Side effects that will be measured are in case of patients with lung cancer mainly due to irradiation. To describe these side effects the Common Terminology Criteria for Adverse Events (CTCAE) v3.0 [DA1] will be used[2].
2. Most important are skin reactions and acute pneumonitis in case of lung cancer.
3. Haematological toxicity will only be analysed in patients who have received chemotherapy.

B3. CLINICAL DATA

1. Age
2. Sex
3. Description of eventual previous treatments
4. Any other pertinent clinical data

B4. IMAGING DATA

In case of lung cancer the following imaging data will be made available:

1. CT (dense coverage of the region of interest)
2. Any modality providing metabolic information (eg. PET, SPECT, fMRI)

All imaging data sets correspond to a known time point before treatment initiation, eventually several time points during treatment (if this is possible) and to at least one known time point after completion of treatment.

B5. SEGMENTATION DATA

1. External “conventional” boundaries of the tumour based on CT
2. Boundaries of the highly metabolically active and necrotic regions based on any modality providing metabolic information
3. Boundaries of any additional tumour structures lying within the previously mentioned “conventional” boundaries of the tumour (including their macroscopic characterization)
4. Desirably: boundaries of adjacent normal tissue structures with annotation.
5. Estimate of tumour volume (including calculation method)

[2] https://webapps.ctep.nci.nih.gov/webobjs/ctc/webhelp/welcome_to_ctcae.htm

B6. HISTOPATHOLOGICAL DATA

In case of lung cancer the following histological data will be made available:

1. Detailed description of the tumour subtype including grade and stage.
2. Microphotographs of indicative histopathology slices from surgically resected specimens and from biopsy specimens. The former will include central and peripheral parts of the tumour, the latter will be representative for the peripheral/ endobronchial part of the tumour.
3. Estimates of the spatial tumour cell density will be given. The amount of dead (necrotic, apoptotic) cells will be estimated.
4. Estimates of the spatial tumour cell density for prolferative cells (Ki67 staining) will be given.
5. The degree of neovascularisation is given for selected cases.

The microscopic imaging data are processed in such a way that the density of tumour cells as well as the percentages of proliferative cells (Ki67) and dead cells are calculated. In few cases, in which such data are available, a mapping of the bioptic areas and the imaging will be done. This will lead to a relatively refined consideration of the tumour spatial inhomogeneities in different areas of the tumour. In all other cases where such information is not possible to get, the information from the biopsy will be extrapolated to the whole tumour, knowing that this is only a rough approximation.

B7. MOLECULAR DATA

In case of lung cancer the following data will be made available:

1. The mutational status of EGFR known to play a critical role in the response to targeted therapeutics will be provided. No other mutational analysis will be done.
2. Autologous antibodies against tumour specific antigens will be measured in all cases using a modified SEREX method.

ContraCancrum Lung Cancer Data Base

The above mentioned data are stored in excel data sheets. For each patient the whole data set (containing personal data, histopathologic data, data from the Saarland Cancer Registry, Ki-67 LI, MSA/MVD, EGFR mutation analysis results) is represented in a single raw of the data sheet. All of these data and the imaging data, non-DICOM imaging data (microphotographs, macrophotographs, Ki-67 images, and CD31 images), and molecular data (EGFR gene sequences) that will finally be stored in ContraCancrum will be anonymized and pseudonymized as described in “Data security”, D2.2 & D9.1 (see fig. 4.5.). Anonymization or pseudonymization is done at the University Hospital of the Saarland, before data are stored in the databases of ContraCancrum. Personal data that are anonymized are: name, birth date, place of residence and those given in 8.1.1 in D2.2 & D9.1.

Only the Institute of Pathology, USAAR has access to these personal data. Pseudonyms for each lung cancer case included in the ContraCancrum data base are structured as follows:

L-XXX (XXX = consecutive number)

3.2 Workflow of data within clinical trials

As soon as a patient is enrolled in a clinical trial data will be continuously generated during the whole run of the trial (fig. 3.1). They have to be documented on CRFs of the trial and transferred to the study centre. With the help of a RDE system this will be done automatically and the data are stored in the trial database.

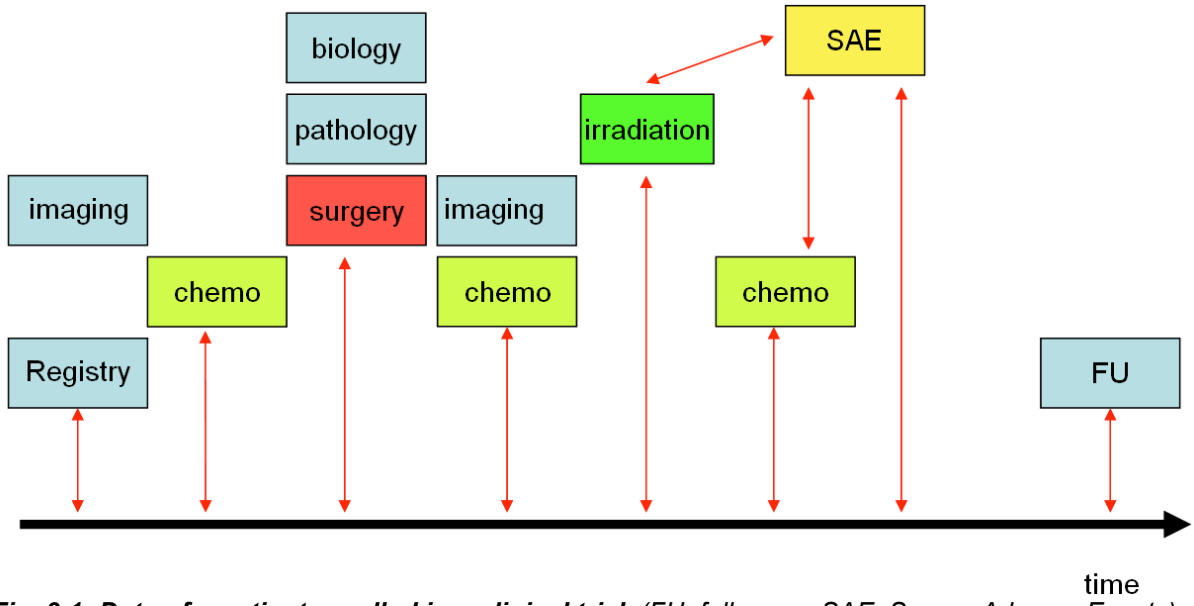

Fig. 3.1: Data of a patient enrolled in a clinical trial. *(FU: follow-up; SAE: Severe Adverse Events)*

The workflow of data in a clinical trial is shown in figure 3.2. The local hospital treating a patient has to register the patient in the trial centre and in a cancer registry after getting informed consent by the patient or parents or legal guardian. There needs to be an exchange of data between the trial centre and the cancer registry for especially follow-up data to synchronize them. For follow-up requests can go directly from the trial centre or the cancer registry to the patient or the local hospital depending on the rules of the trial and the informed consent the patient has signed. If an SAE or SUSAR occurs the local hospital has to report this event to the trial centre that is responsible for sending the information after approval to all participating centres and the legal authorities. In case of international trials sometimes a national data centre collects data for that country, in which case the national data centre has to transfer this data to the international data centre. For analysis data transfer has to be done to the statistical centre. If research projects are part of the trial exchange of data has to take place between the research centre and the trial centre. The same is necessary for reference centres, like reference pathology or reference radiology. All data transfer has to be done in a secure way. Personal data need to be pseudonymized or anonymized before such data is allowed to leave the local hospital to guarantee data safety. In case trial data are allowed for open access such anonymized data have to be transferred to a web database. An audit trail of the data flow is mandatory and GCP relevant. For safety reasons the official trial database needs back-up.

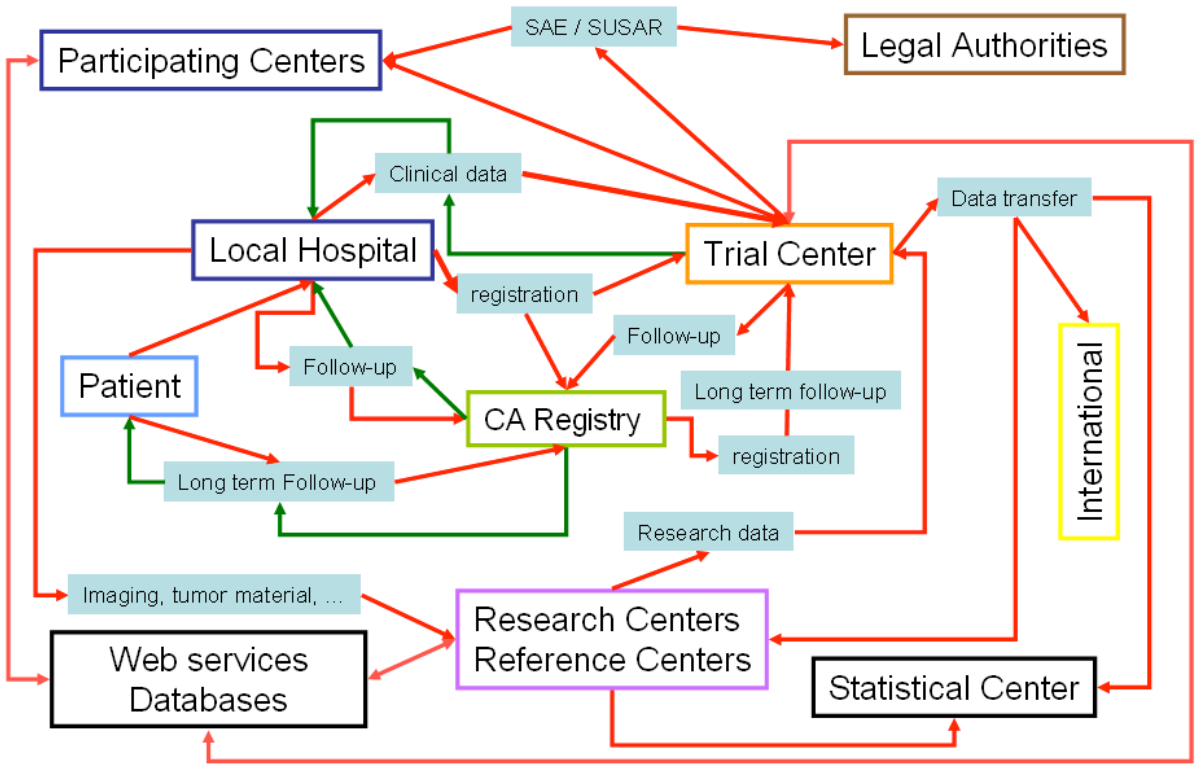

Fig. 3.2: Workflow of data within a clinical trial. *(Red arrows illustrate data that are send, green arrows illustrate request for data; CA Registry: Cancer Registry)*

4 Time frame for in Silico Models

4.1 Introduction

ContraCancrum aims to contribute to the understanding of complex biological phenomena through the multilevel modelling of cancer in the clinical setting. This will take research on cancer modelling a step further by integrating molecular, cellular, tissue and higher level concepts into a single technological entity that will simulate therapy outcome based on the individual patient information.

In order to construct multi-level simulation models of tumour growth and tumour and normal tissue response to treatment schemes and schedules ContraCancrum will:

- Develop medical image analysis algorithms and software for extracting pathophysiological information from different levels of diagnostic information (e.g. MRI, CT, PET, and ultrasound. Several bio-medical parameters/markers will be tested in order to optimize information extraction from treatment monitoring medical images.
- Develop macroscopic biomechanical finite element models of the brain and lung to calculate the stress/strain in these tissues. Deformations of both tumour and neighbouring normal tissues will be simulated based on tissue biomechanical properties and detailed anatomic atlases.
- Deploy two important clinical studies for validating the models, one on lung cancer and one on gliomas. The crucial validation work will be based on comparing the multi-level therapy simulation predictions with the actual medical data (including medical images), acquired before and after therapy. ContraCancrum aims to pave the way for translating clinically validated multilevel cancer models into clinical practice.
- Create a workflow environment that will allow remote access to clinical data and will assist the end clinical user to validate the cancer models by using its web services. The project will use open-source software that will allow for future extensions of models as well as the extension to large scale clinical trials. Data pseudonymization will ensure adoption of the European legal and ethical data handling guidelines.

4.2 Clinical requirements

If ContraCancrum brings multilevel cancer models into clinical practice these models have to be validated in daily practice as decision making tools. From a clinical standpoint models in ‘InSilico Oncology’ should answer one of the following questions to be used as a decision making tool:

- What is the natural course of a tumor?
    - growth rate, shape and volume
    - interaction with the surrounding tissue
- When and whereto is a tumor metastasizing?
- Can the response to a given treatment be predicted?
- What is the best treatment schedule?
- Is it possible to predict SAEs?
- Is it possible to predict the development of cancer in advance?

All of such questions have to be asked within clinical scenarios or trials to get clinical approval. If such a model will be used as a decision making tool the answer of the model is critical in terms of time. Having a model asking the question 'What is the best treatment for a specific patient?' the answer is needed before the treatment for the specific patient will start, otherwise the model is useless as a decision making tool.

Such time constraints need to be considered in developing models. As the execution of a model in the computer is depending on different and heterogeneous data the availability of specific data for the model will limit the usability of the model as a decision making tool. The data that are needed as input in the model are mainly:

- Clinical data
- Imaging data and other DICOM data
- Molecular genetic data
- Pharmacological / pharmacokinetic data
- Open source data (Web databases, etc.)

Often these data do need processing before they can enter the model. Imaging data need segmentation of the tumour, molecular genetic data are the result of time consuming laboratory experiments, clinical data have to be collected, etc. If the data are generated they have to be validated and then transferred to run the simulation. The simulation itself needs time. After simulation the result needs to be validated before it will be send to the clinician to help him in applying an individual treatment to a patient. If all these steps are done within a limited time period such a model can be used as a decision making tool in the clinical setting. Only if a physician gets an answer before he starts treatment such tools are useful for clinicians. Therefore it is of utmost importance to calculate the time needed for the following steps:

- Generating of data
- Processing of data
- Validation of data
- Transfer of data
- Execution of the model
- Validation of the result
- Sending the result to the treating clinician

In this time frame issues of data security and pseudonymisation / anonymization need to be regarded according to the legal and ethical requirements and may take further expenditure of time. Such a calculation of the time frame for a model has to be an integral part of description of the model. Developers of a good model for clinical practice need to shorten the time for every step as much as possible without lowering the quality of the data needed for the model. Such models are validated by the question if the results can be delivered in due time for the use in a specific patient. Only in then such a model should be called a decision making tool.

5 Extension of the ACGT Master Ontology for lung cancer and gliomas

5.1 Introduction

Tom Gruber's often cited definition for the term "ontology" says that an "ontology is a formal, explicit specification of a shared conceptualization"[3]. Taking this definition as basis, the advantage of applying ontologies for the purpose of data mediation and integration becomes obvious: By introducing both well-defined and well-structured ontological concepts – instead of just using idiosyncratic terms – to describe some given domain, an ontology provides for the fundamental building blocks to enable full and adequate semantic interoperability between heterogenous data sources.

Within the ACGT project the aim is to create a common ontology which covers the whole spectrum of cancer, ranging from basic molecular research up to clinical patient management. This so-called ACGT Master Ontology (ACGT-MO) has been implemented in the standard representation language OWL-DL (Web Ontology Language – Description Logics) enabling the application of many existing tools, e.g. to perform editing, syntactical integration or logical reasoning, and the direct development of own specific tools.

At its upper level, the ACGT-MO incorporates the Basic Formal Ontology (BFO), a coherent and logically consistent top level ontology which is already used successfully in several biomedical ontology projects. The inclusion of BFO is thus a highly important step in supporting the "top down" approach in ontology development that is inevitable in order to arrive at common and sound conceptual definitions and a consistent taxonomy. In light of this, Lassila et al.[4] show that a number of criteria exist for classifying ontologies: In order to achieve the best possible results some of those high-level criteria are integrated in the ACGT-MO, such as the formal definition of subsumption ("is-a") and the application of general logical constraints to refine and restrict concepts. Furthermore, the developers of the ACGT-MO aim to meet the various principles of the OBO (Open Biomedical Ontologies) Foundry[5] which serve as an established facility for assessing the quality of an ontology.

Due to the multiplicity of entities and processes present in the clinical trial domain, the ACGT-MO contains elements that cover a wide range of topics, from genetics, the medical field to the administrative field (e.g. participation in a study) as well as the legal domain (e.g. consent). Concerning the medical content covered, the ontology presently focuses on the definition of the concepts needed for Wilm's tumours, Rhabdoid Tumour and breast cancer, in respect to anatomy, specific medical processes and procedures, etc.

In order to apply the ACGT-MO within ContraCancrum, it will therefore be necessary to broaden the ontology content with lung cancer and brain tumour concepts. Here the fact comes into play that such extensions have been envisaged from the outset of the ontology development: Both its structure and concepts (along with their definitions) make the ACGT-MO directly adaptable to novel scenarios without significant changes. The extension will therefore also adhere to the current well-established development procedures and follow the same design principles of the ACGT-MO. Thus, available material will be studied first in order

[3] R. Studer., V. R. Benjamins and D. Fensel, Knowledge Engineering: Principles and Methods., IEEE Transactions on Data & Knowledge Engineering, vol. 25 no. 1-2, pp. 161-197, 1998

[4] O. Lassila, D. McGuinness, The role of frame-based representation on the Semantic Web. Technical Report KSL-01-02. Knowledge System Laborartory. Stanford University, Stanford

[5] www.obofoundry.org

to determine the actual description needs (e.g. medical textbooks and more specific material from project partners) and create an initial "design" which will then be refined in an iterative close collaboration with the domain experts within the project.

5.2 Concepts for lung cancer description

Specific requirements for the description of lung cancer are:

- Pathological entities
  - anatomical site
  - histopathology
- Co-morbidity
  - other diseases
  - toxicity
- Procedures
  - diagnostic procedures
    - radio-diagnostics
    - laboratory
  - surgical procedures
    - biopsy
    - resection
  - therapeutic procedures
    - medical procedures
    - radio-therapeutic procedures
- Prognostic entities
  - Late effects
  - Outcome

5.3 Concepts for brain tumour description

Specific requirements for the description of brain tumours are:

- Pathological entities
  - anatomical site
  - histopathology
- Co-morbidity
  - other diseases
  - toxicity
- Procedures
  - diagnostic procedures
    - radio-diagnostics
    - laboratory
  - surgical procedures
    - biopsy
    - resection
  - therapeutic procedures
    - medical procedures
    - radio-therapeutic procedures
- Prognostic entities
  - Late effects
  - Outcome

6 ObTiMA - a tool for data management

6.1 Introduction

Post-genomics research and clinico-genomic trials on cancer have resulted in an enormous amount of information and multilevel biomedical data that are stored in many disparate data sources. The analysis of such integrated data provides a great opportunity to improve treatment in cancer patients and to meet the demanding individualization of care needs. But up to now, the heterogeneity of data sources and the lack of a common infrastructure have mainly prevented clinical research institutions from mining and analyzing disparate data sources. This problem is addressed by ContraCancrum and ACGT (Advancing Clinico Genomic Trials on Cancer)[6]. One of the main objectives of the ACGT project is the semantic integration of heterogeneous biomedical databases using a mediator. The mediator exploits the ACGT Master Ontology on Cancer (ACGT-MO), which comprises the knowledge needed to design and carry out specific clinical trials. To integrate a data source into the mediator it has to be mapped to the ontology, an error prone and tedious task. It would be desirable that databases of newly developed data management systems could be set up during creation in an ontology compliant way to allow a seamless integration of the data collected in these systems. This approach is done in ObTiMA (Ontology based Trial Management Application) and enables a chairman of a clinical trial to set up a clinical data management system with comprehensive metadata in terms of the ACGT-MO.

6.2 Requirements for Clinical Trial Management Systems

A **Clinical Trial Management System**, also known as *CTMS*, is a customizable software system used by clinical researchers, biotechnology and pharmaceutical industries to manage the large amounts of data involved with the operation of a clinical trial. There are different requirements that such a system must satisfy. Some popular requirements include: budgeting, patient management, compliance with government regulations, and compatibility with other data management systems. It has to maintain and manage the planning, preparation, performance, and reporting of clinical trials, with emphasis on keeping up-to-date contact information for participants and tracking deadlines and milestones such as those for regulatory approval or the issue of progress reports. It has also to include the database for the clinico-genomic data of the trial.

IN ACGT a CTMS is created. It is called the '**O**ntology **b**ased **Tri**al **M**anagement **Application**' (ObTiMA). ObTiMA consists of three parts:

- Trial Builder
  - Trial Outline Builder (TOB)
    - Including a graphical schema of the trial
  - CRF Creator (CC)
- Repository
- Patient Data Management System (PDMS)

[6] Tsiknakis M, Rueping S, Martin L, Sfakianakis S, Bucur A, Sengstag T, Brochhausen M, Pucaski J, Graf N: Developing a European Grid infrastructure for cancer research: vision, architecture, and services. Ecancermedicalscience 1: DOI: 10.3332/eCMS.2007.56, 2007

6.3 ObTiMA

The main components of ObTiMA are the Trial Builder and the patient data management system. The Trial Builder allows a trial chairman to define the master protocol, the Case Report Forms (CRFs) and the treatment plan for the trial, in a way that is both semantically compliant with the MO and user-friendly. From these definitions, the patient data management system that allows the collection of relevant data for individual patients can be set up automatically. This collected data is stored in trial databases whose comprehensive metadata has been rendered in terms of the MO.

ObTiMA is based on reusable components. There are basic components eg. for patient specific data (age, gender, date of diagnosis), or for histology, or for SAE and SUSAR reporting etc., there are administrative components and there are trial specific components. According to the reusability of components there is the need of a repository for trials and CRFs. ObTiMA will include multinational language support and a reporting and querying system.

The Patient Data Management System (PDMS) will be automatically developed during the process of building a new trial. The PDMS is ontology based. All administrative and scientific aspects of clinico-genomic trials will be integrated, that are needed to run the clinico-genomic trial. The Ontology and the security are integrated parts of the system without being noticed by the end-user. The ACGT Master Ontology (MO) will be used.

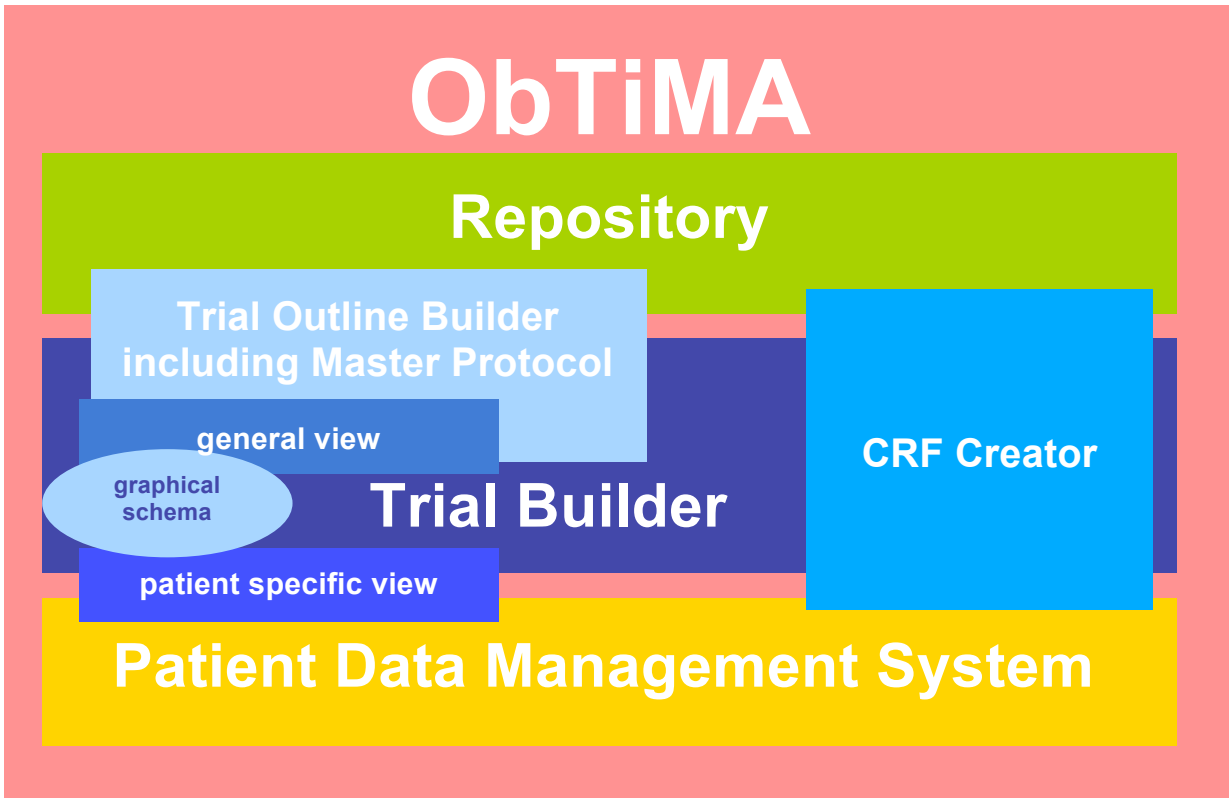

Fig. 6.1: Schematic diagram of ObTiMA

ObTiMA can be used in two different ways. With the help of ObTiMA a new trial can be build up with all its components and ObTiMA can be used as a Remote Data Entry (RDE) system for managing a patient in a clinical trial including data entry and regarding all administrative

aspects for patients enrolled in a trial. These two different functionalities of ObTiMA are strictly separated according to the roles and rights of an end-user.

The Trial Builder is primarily used to build a new trial. The user will be guided by a Master Protocol for clinical trials (Appendix 2) to write the Trial Protocol, to build a graphical schema of the trial and to create all CRFs that are needed for the trial. All legal and ethical requirements will be considered during this process and appropriate solutions provided. By creating new CRFs the database for the trial will be automatically generated. This database is always ontology based. This primary functionality of the Trial Builder is available by the chairman of the trial, or other end-users having the right to build a new trial. No other user is allowed to edit, change, delete or save anything using the trial builder. But every user having enrolled patients or provided molecular data into the trial can use the Trial Builder for reading and printing the Trial protocol including all CRFs and the graphical schema of the trial. Such an end-user will be able to use the graphical schema of the trial not only as a general overview of the trial but also as a tool for getting a personalized view of a patient he has enrolled in the trial. This functionality is still under construction.

The repository is used for storing of trials and CRFs that are created in ObTiMA.

The PDMS is the data management system of the trial used by participants of a trial via remote data entry (RDE).

Access to ObTiMA is regulated by the roles and rights management of ACGT.

	Roles	
	User building a new trial (e.g. chairman of a trial)	**User taking part in a trial and enrolling Patients' or providing molecular data**
	Rights	**Rights**
Trial Builder		
Trial Outline Builder	enter, edit, store, delete	read, print
Graphical schema of the trial		
general view	enter, edit, store, delete	read, print
patient specific view	not available	read, print
CRF Creator	enter, edit, store, delete	enter, edit, read, print
Repository	enter, edit, store, delete	read, print, upload molecular data
PDMS	will be automatically build	enter, edit, store, delete

Tab. 6.1: Roles and rights regarding the two different ways ObTiMA can be used.

6.4 Case Report Forms (CRFs)

The most important parts of data management systems in clinical trials are the Case Report Forms (CRFs) which are designed to collect the required research and administrative data as well as the trial database for storing these data. In many multicentre trials paper based CRFs are still used today. From participating hospitals thousands of CRFs are sent to a central data facility, where the data are entered into a trial database. This is very time consuming and error prone. Often the clinical trial databases are in-house developments that have to be implemented from scratch for each new trial[7]. Today preferable systems are Web based remote data entry systems where the data are captured at the participating site and transferred electronically to the trial centre. Most of these management systems allow trial chairmen to design a trial and to create eCRFs (electronic CRFs) without informatics skills (e.g. in the OpenClinica system[8]). The user is free in defining database tables, items on CRFs and attributes before the database is automatically generated from these definitions. Although that is a desirable feature, such databases lack standardization and do not comprise comprehensive metadata. Queries across trials are not possible to run.

In setting up a trial using ObTiMA, clinicians should not be concerned with theoretical aspects and design principles of databases or ontological metadata. Therefore, in ObTiMA, the trial chairman defines both, by creating the CRFs for his trials. He is assisted in defining the questions on the CRFs, the order in which the questions will be queried, and constraints on the answer possibilities. Creating a question on the CRF is supported by simply selecting appropriate classes from the MO. We will give the creation of the question field “Patient’s Gender” as an example. The clinician is given a list of relations or properties relating the class *Patient* to other classes in the MO. He observes that a relation between the classes *Patient* and *Gender* does exist. In creating the corresponding question, he simply has to choose the class *Gender*. The attributes required in order to create the possible answers on the CRF are then determined automatically. The values allowed are set automatically to *Male*, *Female*, and *AmbiguousGender* since the class *Gender* is defined as an enumeration in the ontology containing these values and a multiple choice question is subsequently automatically created on the CRF.

This procedure implements the semantics of the ontology in the CRFs in an automatic fashion. We expect that the description alluded to above be a path from the ontology starting at the class *Patient*, as this is normally the focal point of CRFs. From there the clinician grasps the relevant MO classes and relations that connect with *Patient*, e.g that the patient can have diseases, a blood pressure, etc. Now, he has the opportunity to either select one of the classes in order to create a question on the CRF, or to explore them further and show their subclasses or relations to assemble more complex questions.

With the aim of setting up the appropriate database for storing the data, the following attributes are needed for each question: the question itself, data type of the answer and optionally possible data values, range constraints and measurement units. These attributes will, as much as possible, be determined automatically from the path the trial chairman has selected, but can later be changed by the trial chairman to a certain extent. Through the integration of the MO into ObTiMA, data sharing between clinical trials becomes possible. This is necessary to leverage the collected data for further research like cross-trial meta-analysis.

---

[7] Weiler G, Brochhausen M, **Graf N**, Hoppe A, Schera F, Kiefer S: Ontology Based Data Management Systems for post-genomic clinical Trials within an European Grid Infrastructure for Cancer Research. Proceedings of the 29th Annual International Conference of the IEEE EMBS, Cité Internationale, Lyon, France, August 23-26, 2007, SuA11.4; Conf Proc IEEE Eng Med Biol Soc. 2007;1:6434-6437

[8] www.openclinica.org

6.5 Planned Extensions

A “hidden” asset of the ObTiMA system is the underlying architecture of that system: Its modular design is based on coupling and integrating purpose-specific modules through well-defined, open interface definitions, which enables the creation of functionality extensions in a straightforward way. These modules can be activated or deactivated according to the specific needs and rights of the particular ObTiMA user.

At the moment, one main extension idea is to include the possibility to view and transfer images in the DICOM standard within ObTiMA. Even though several (and also some free) DICOM viewers exist for various computer platforms, they are still available as separate applications only. Thus, in order to create a single continuous workflow, it would be advantageous to integrate DICOM facilities within ObTiMA itself. To do so, existing open-source DICOM viewers are currently investigated and evaluated whether they offer the necessary programming interfaces. In this context, one major challenge lays in the fact that the found DICOM viewers are desktop applications which then need to be ported into ObTiMA’s web application base.

The same aspects apply for adding (video) conferencing possibilities to the system. Here several systems already exist on different levels of sophistication (i.e., from simple chat to complete conferencing systems). It needs to be evaluated which level of sophistication is truly necessary and can be feasibly integrated into ObTiMA.

To better blend ObTiMA in into current medical working environments, the system can be extended with modules to import existing data into ObTiMA and export newly created data into existing (legacy) systems. It is therefore crucial that ObTiMA understands established data formats the relative importance of which will be assessed together with the expert partners of the projects. Currently, the ObTiMA system “understands” CDISC's Operational Data Modeling (ODM) format for representing the study metadata, study data and administrative data associated with a clinical trial.

As last point to mention is the ability to customize the user interface according to the specific needs at the level of some institution and at the level of a particular user. This means that, for example, certain interface areas can be adapted to (graphically) mimic existing workflows at an institution or to hide complexities unnecessary for a user.

7 Workflow of DICOM data

7.1 Introduction

Patients enrolled in the lung cancer scenario or in the glioma scenario undergo different imaging processes over time as described in 3.1.1 A4/A5 and 3.1.2 B4/B5. These data are stored as DICOM files and needed for the 'in silico' model. Part of this workflow is described in deliverable 4.1. in section 6 and 7 explaining the modelling of imageable tumour growth and response to treatment and the modelling of normal tissue response to treatment.

The ContraCancrum Integrated Data Environment is the platform used for storing and retrieving DICOM files as well as other data within the project. An initial version of this platform is available at http://test.contracancrum.eu/ as shown in figure 7.1. Data are only allowed to upload after pseudo-anonymization. In most cases further processing of data is necessary before they can be used in the modelling scenarios. This processing includes mainly the segmentation of the tumour in case of imaging studies.

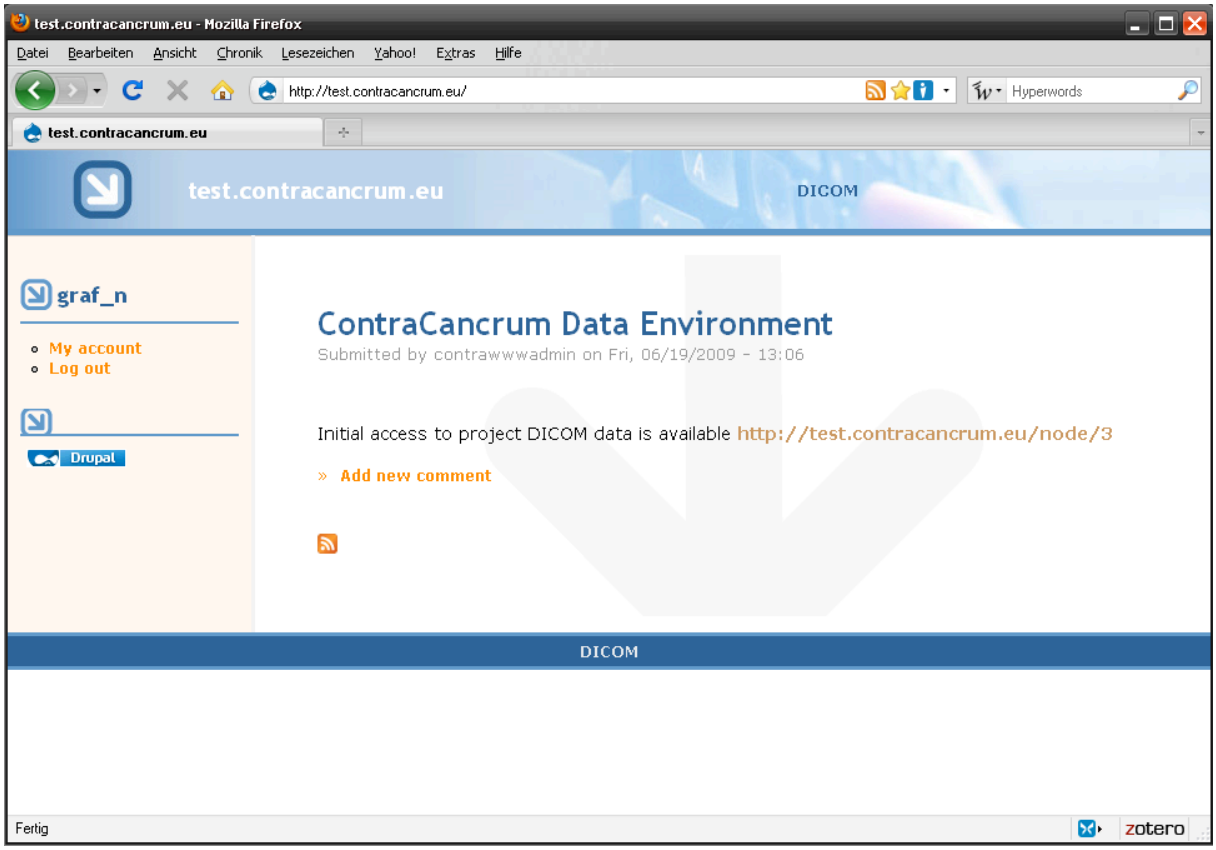

Fig. 7.1.: ContraCancrum Data Environment platform.

MRI images are saved in dcm files (DICOM format). Such DICOM files for lung cancer and glioma cases will be provided by Universitätsklinikum des Saarlandes. After having acquired these files according to the protocol between ContraCancrum and Universitätsklinikum des Saarlandes, they will be uploaded to the dedicated server that UCL manages.

Thereafter, the clinicians could use the ContraCancrum tool, DoctorEye that is a general segmentation/annotation/3D representation tool. The clinician has the ability:

- To download datasets of DICOM files from the server and load them.
- To select some of the images from the MRI datasets, or all of them.
- To load stored segmentation information that the DICOM files may have attached in them.
- To delete existing annotations of tumour contour or create new ones. Multiple annotations can also be performed.
- To use different algorithms for semi-automatic segmentation (Active Contours (Snake algorithm, Greedy algorithm)) and magic wand)
- To manually define the edges of the tumor.
- To use erasing/drawing tools for correcting an annotation
- To visualize the segmented tumour areas in different MRI slices in 3D.
- To store the processed files on the server.

7.2 DoctorEye for segmentation

DoctorEye can be downloaded from http://139.91.189.171:8891/, where DICOM images are also stored. DoctorEye is based on the .NET framework architecture and can be used in any Windows-based computer with the .NET framework ver. 2.0 or later. However, .NET framework permits, in concept, the porting of the application in many other Linux based environments, using the Mono project functionality and libraries, so the user can run an application using the No-Touch Deployment. Internet connection is necessary, since the program has to communicate with the DICOM server.

7.2.1 Loading the MRI images from DICOM

When loading DoctorEye, the program automatically connects to the DICOM server. A list of DICOM files is then presented. The user can choose some or all of them and then DoctorEye downloads it to the local machine.

The MRI images are then presented to the user. Next to each image, a tick box is presented, so that he/she can choose the images to process. After having chosen the images of interest, the respective thumbnails are presented.

If there are previously stored segmentations in the DICOM file, they also appear automatically on screen.

7.2.2 Segmentation of the MRI images

After selecting a specific MRI slice, the user can load the already existing annotations and manage them (add, rename, remove etc) or start to create new annotations of the tumor either manually or semi-automatically. DoctorEye gives the ability to create more than one annotation on the same MRI slice, remove unwanted ones, name them and draw them with different color.

When creating a new annotation using semi-automatic methods, the user could use three types of tools that can be used:

- The magic wand tool:

  This tool automatically selects the area around a chosen central pixel according to their gray level values. The pixels are selected if the difference in gray level with the

central pixel does not overcome a threshold. This threshold can be adjusted by the user.

- The active contours algorithm (in two different versions, the greedy algorithm and the active contours – snake one)

  This algorithm requires from the user to form an ellipse inside the tumour area, by using the mouse. The algorithm is creating a contour, which iteratively expands, in order to fit the tumour boundary. The evolution of the contour is accomplished by minimizing a properly defined snake energy function. The user can adjust four parameters, in order to change the snake's behaviour and control the result.

However, the user could manually annotate the tumour area or correct the segmentation results of the tools, by using the drawing tool and the erasing tool.

For follow-up cases, an histogram-based segmentation approach is to be investigated and incorporated into DoctorEye. The main concept is based on enhancing tumour detection and segmentation by analyzing changes of tissue intensity patterns over time, and relative to patient-specific healthy areas. The concept is illustrated in Fig. 7.2

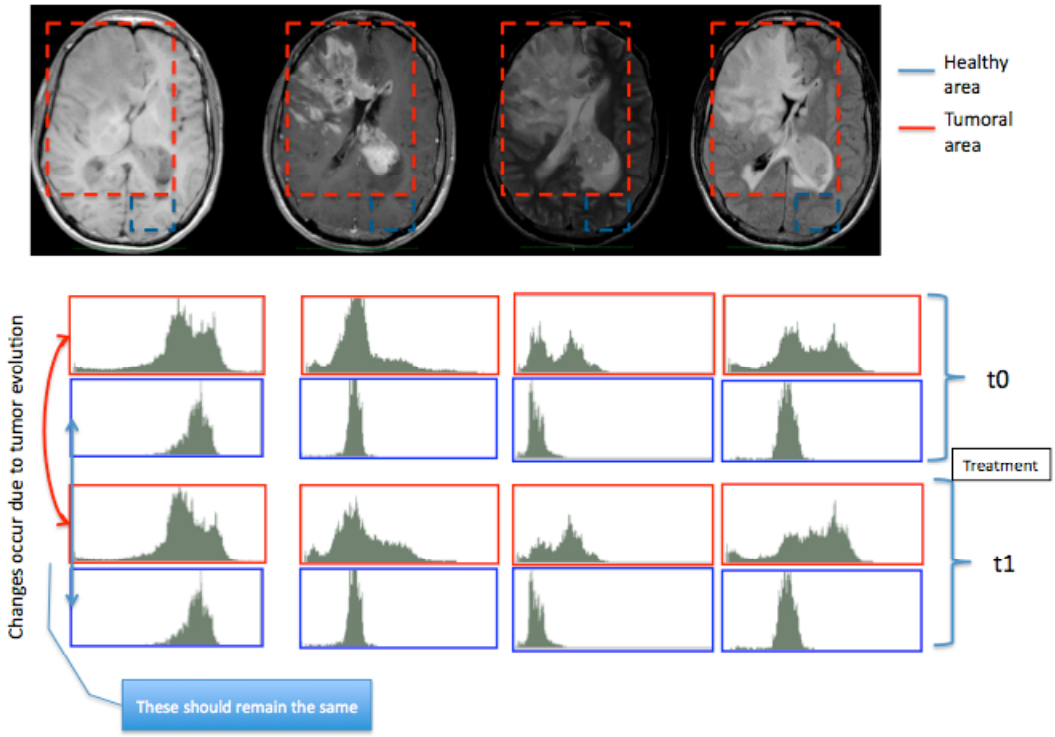

Fig 7.2 Tumour segmentation based on analysis of tumoral intensity pattern changes over time with respect to healthy tissue.

In addition, age-dependency is expected to add further information to the segmentation process since it known that myelinisation of the white matter in the brain is age dependent.

7.2.3 Moving from an MRI image to another

The user can move to any other MRI image, by selecting it from the thumbnail list of the selected images. This won't affect any annotation he has done to a specific image, since it is going to be presented when the user returns to this image.

7.2.4 Storing the annotations

The annotations can be stored by clicking the Store Work button, at any time the user wants. Then, the user will be asked to give the name of the new DICOM file that is going to be saved in the server. If the user selects the name that already exists on the server, a message for "overwriting file" confirmation will be presented. The new DICOM file stored in the server, can then be loaded by anyone that has DoctorEye installed locally.

7.2.5 DICOM tag protocol used for the annotations

The annotations are saved in the DICOM file, and more specifically in the "Text Value" tag. The following XML structure is used for annotating:

<annotation>

<id> </ id>
<points> </ points>
</ annotation>

The "id" tag (value as text) contains the name of the object to be annotated (eg. glioma). If the MRI contains 2 or more areas of interest (multiple annotations) then for each area a new annotation object is defined with the same "id". Therefore multiple annotations with the same "id" make reference to the same object. The points" tag (value as text) contains the sequence of points which define the contour of the annotation in serial pairs of coordinates (eg. x1 y1 x2 y2 x3 y3 ...). There is no need to implement a tag for the colour. The colour of each annotated area can be defined within the programs. Each annotation with different "id" tag will have a different colour. The ID numbering is unnecessary, as in every DICOM file is an MRI and we make reference to it. The structure of XML to ensures the separation of annotations. The XML information is finally stored in the tag TextValue of the DICOM format.

7.3 Workflow of DICOM data in the lung cancer scenario

Figure 7.3 gives an overview of the workflow of DICOM data in the lung cancer scenario. Data will be send from the Radiology PACS at UdS to the ContraCancrum server at UdS and anonymized. After anonymization the data will be uploaded to the ContraCancrum DICOM server managed at UCL. WP7 partners can retrieve the DICOM data for processing them. Most important is the segmentation of the tumour done by using DoctorEye. After processing of the data they will be sent back to the radiologists/clinicians for checking the correctness of the segmentation. After confirming correct segmentations these Data will be uploaded on the ContraCancrum DICOM Server as validated data and can be used in the Oncosimulator.

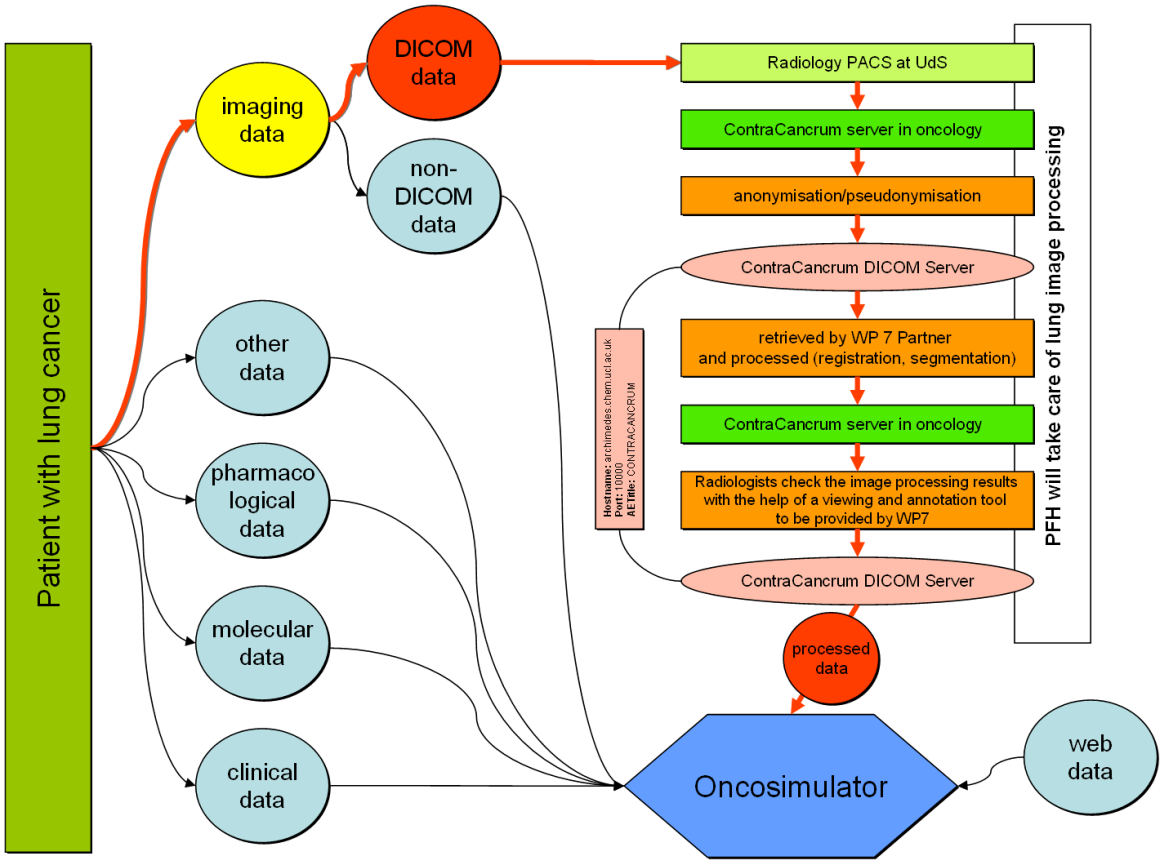

Fig 7.3 Workflow of DICOM data in the lung cancer scenario

7.4 Workflow of DICOM data in the glioma scenario

Figure 7.3 gives an overview of the workflow of DICOM data in the glioma scenario. Data will be send from the Neuroradiology PACS at UdS to the ContraCancrum server at UdS and anonymized. After anonymization the data will be uploaded to the ContraCancrum DICOM server managed at UCL. WP7 partners can retrieve the DICOM data for processing purposes. Most important is the segmentation of the tumour done by using DoctorEye. After processing of the data they will be sent back to the neuroradiologists/clinicians for checking the correctness of the segmentation. After confirming correct segmentations these Data will be uploaded on the ContraCancrum DICOM Server as validated data and can be used in the Oncosimulator.

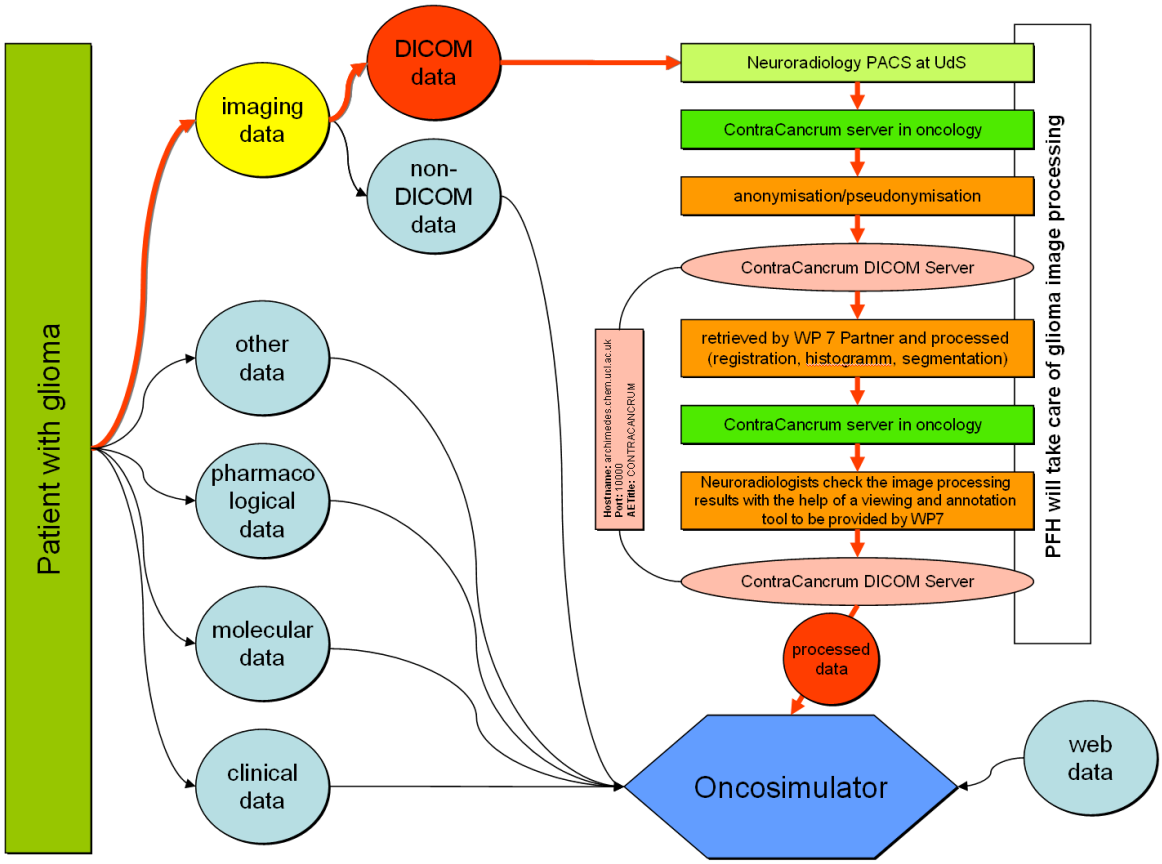

Fig 7.3 Workflow of DICOM data in the glioma scenario

8 Workflow of non-DICOM imaging data

8.1 Introduction

Part of this workflow is described in deliverable 4.1. in section 6 and 7 explaining the modelling of imageable tumour growth and response to treatment and the modelling of normal tissue response to treatment.

8.2 Workflow of non-DICOM imaging data in the lung cancer scenario

8.2.1 Collective Sampling Data

All biopsy samples from back to 2007 until now and also prospectively that were/are suspected as primary lung tumour by the clinicians have been preliminary retrieved. We have collected 277 cases, which then are evaluated and finally included in the definite biopsy series. Among lung tumour resection samples only those that are sent freshly and unfixed during surgery are prospectively collected. From each resection case fresh tumour tissue from the centre and the periphery as well as non-neoplastic lung tissue are snap-frozen with liquid nitrogen and then stored in a freezer at -80°C. We have collected 81 resection cases that are also evaluated as described beneath (*Histopathologic Data*).

8.2.2 Histopathological data

All available slides of the retrieved cases are consecutively reviewed microscopically. Until now, the diagnosis of NSCLC has been verified in 60 biopsy cases and 78 resection cases. TNM stage and tumour grade has also been re-evaluated and revised in these cases, if necessary.

8.2.3 Macroscopic and microscopic image data

Macrophotographs: Macroscopic photographs are taken from fresh, unfixed resection samples. We have macrophotographs from 22 informative NSCLC resection cases.

Microphotographs: In both biopsy and resection cases two digital micrographs were taken at low power (1:100) and three digital micrographs at medium power (1:200). Complete collections of microphotographs have been accomplished in 27 biopsy cases and in 46 resection cases.

Ki-67 labelling index: Tumour cell proliferation is estimated by the Ki-67 labelling index (LI). Immuno-labelling of tumour cells with an antibody directed against Ki-67 (clone MIB1) has yet been done for 14 resection cases. Determination of the Ki-67 LI is described beneath (*Image analysis*).

Angiogenesis: The endothelium of blood vessels are immunohistochemically labelled with an antibody directed against CD-31. Angiogenesis is quantified by assessing the mean microvessel counts (microvessel density *MVD*) by two independent investigators and by determination of CD-31 positive microvessel surface area (*MSA*) in hot spot areas. CD-31 immunohistochemistry has been done in 14 resection cases. Determination of the MSA is described beneath (*Image analysis*).

Image analysis: Segmentation and analysis of immuno-labelled structures in digital images is done by image analysis devices based on ImageJ. Basically, colour-deconvolution, iterative threshold-technique, and watershed-segmentation are applied for segmentation. Percentages of Ki-67 immuno-positive tumour cells are determined in five high-power-fields in the area of highest labelling and separately in the tumour centre and periphery, respectively. MSA is determined in five microphotographs of CD-31 immuno-labelled slides at medium power (1:200) in the areas of highest labelling. Image analysis for Ki-67 LI and MSA has yet been performed in 20 images and yielded high accuracy and reproducibility.

8.2.4 Imaging data

Biopsy cases that are included in the definite biopsy series are consecutively reviewed in the patient information system for available CT scans before and after treatment. Until now, there are 49 cases with only one CT scan available, 36 cases with two CT scans, and 23 cases with at least three CT scans at different times.

8.2.5 Workflow of non-DICOM data in the lung cancer scenario

Figure 8.1 shows the workflow of non-DICOM data in the lung cancer scenario. Non-DICOM imaging data will be sent from the ContraCancrum Server in Pathology after anonymization to the ContraCancrum DICOM Server at UCL for use in the Oncosimulator.

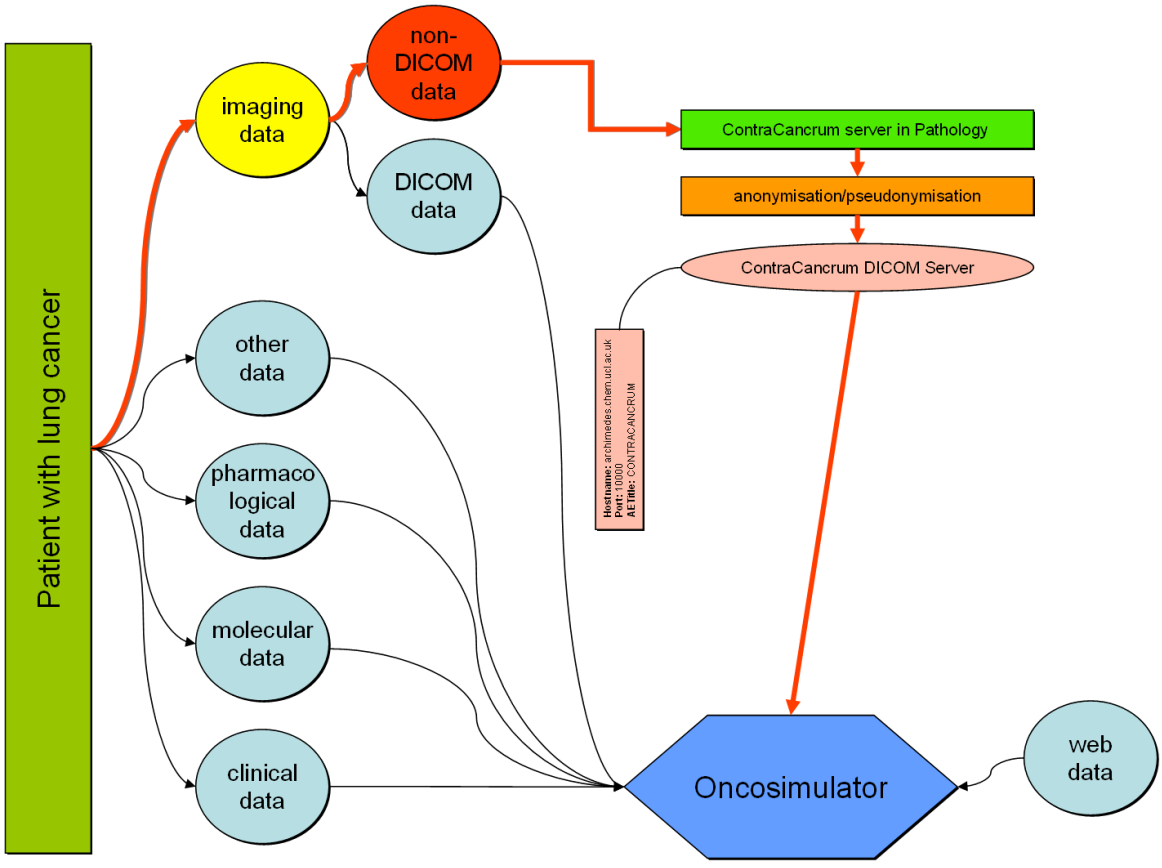

Fig 8.1 Workflow of non-DICOM data in the lung cancer scenario

8.3 Workflow of non-DICOM imaging data in the glioma scenario

Figure 8.2 shows the workflow of non-DICOM imaging data in the glioma scenario. Non-DICOM imaging data will be sent from the ContraCancrum Server in Pathology after anonymization to the ContraCancrum DICOM Server at UCL for use in the Oncosimulator.

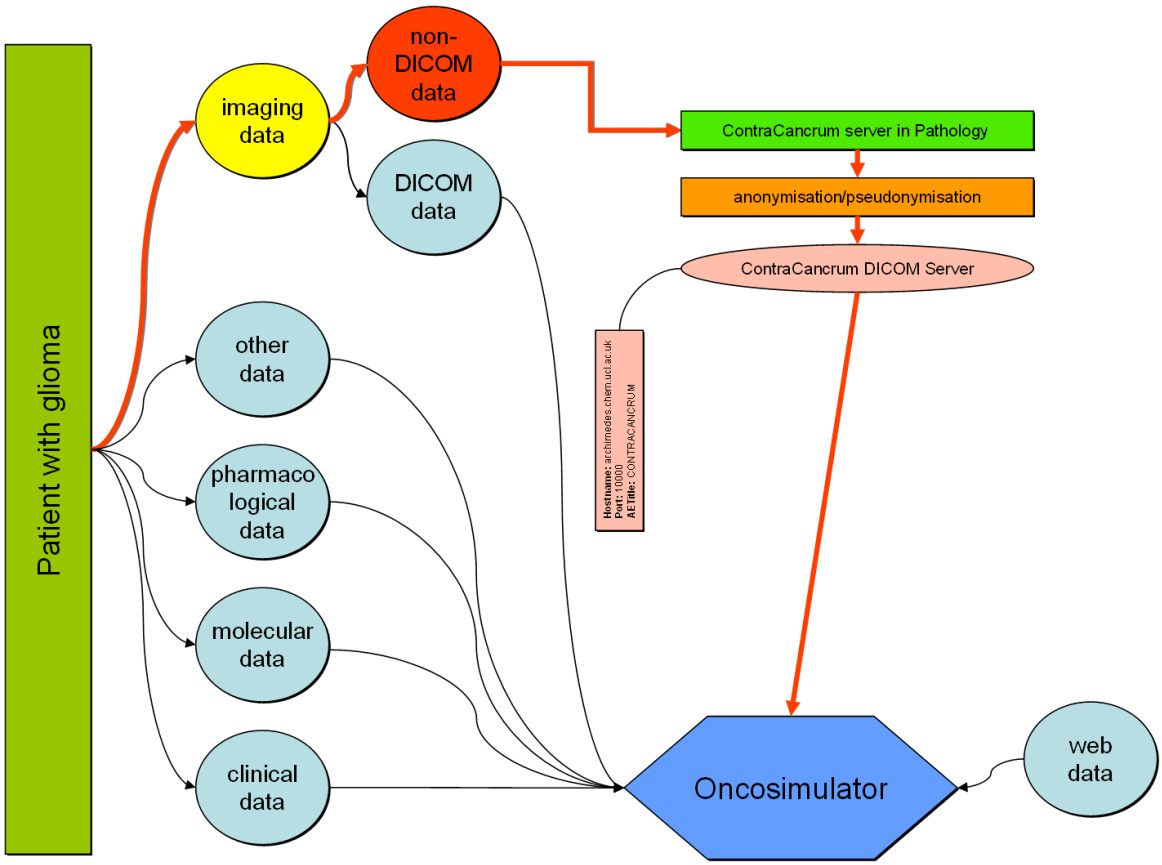

Fig 8.2 Workflow of non-DICOM data in the glioma scenario

9 Workflow of molecular biological data

9.1 Introduction

Part of this workflow is described in deliverable 4.1 in section 6 and 7 explaining the modelling of imageable tumour growth and response to treatment and the modelling of normal tissue response to treatment. Molecular biological data are of utmost importance in the models developed in ContraCancrum (see chapter 12 and fig. 12.1 of this deliverable).

9.2 Workflow of molecular biological data in the lung cancer scenario

The workflow for the lung cancer scenario is described in D2.2 & D9.1. Molecular biological data that are used in the lung cancer scenario are given in 3.1.2:

- EGFR Gene sequencing data and
- Tumour specific antigen data

9.2.1 EGFR Gene Sequencing data

Up to date EGFR gene sequencing for four exons (18, 19, 20, and 21) was performed in 26 biopsies and in 16 resection specimens, respectively. Beforehand, tumour cells were dissected by laser microdissection in 9 biopsy specimens, in the other cases tumour tissue was dissected macroscopically. Until now, we have identified EGFR gene mutations in two specimens (mutation L858R on Exon 21). These mutations were detected in two biopsy specimens (adenocarcinomas, respectively).

EGFR Gene Sequencing data will be sent to UCL for further use in the simulation model.

9.2.2 Tumour specific antigen data

Seroreactive tumour specific antigens will be detected according the description provided for gliomas in D2.2 & D9.1 (page 44 - 47). Data of tumour specific antigens will be ranked according to the information content by correlation to clinical parameters. These data will be provided for use in the simulation model.

9.3 Workflow of molecular biological data in the glioma scenario

The workflow for the glioma scenario is described in detail in D2.2 & D9.1. Molecular data that are used in the glioma scenario are given in 3.1.1 in this deliverable.

- Glioma specific antigen data
- Cytogenetic data will be provided for those gliomas that can be used for short term tissue culture.
- cDNA expression data will be provided for those gliomas that are available as frozen tissues.

All these data will be provided for further use in the simulation model for glioma.

10 Data Mining of public databases

10.1 Public databases

Most of the needed information regarding this chapter is provided in D5.1. The method of mining public databases for lung cancer and gliomas is identical. The following data will be searched for:

- Gene expression data
- Drug sensitivity data
- Radiobiological data
- Clinical data

Data sources that are used for gene expression, drug sensitivity and radiobiology include as described in D5.1 the following:

- NCI-60 Panel
  - Microarray data
  - Drug sensitivity data
  - Radiation sensitivity data
- Prospective Glioma Cell Line Data (University of Cambridge)
- Gene Expression Omnibus (GEO) database
- CMAP Cell Line
- REpository for Molecular BRAin Neoplasia DaTa (REMBRANDT)
- The Cancer Genome Atlas (TCGA)

For clinical data and literature search the following resources will be used:

- MEDLINE/PubMed
  - http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed
- EMBASE/Ovid
  - http://www.ovid.com/site/catalog/DataBase/903.jsp
- The Cochrane library
  - http://www3.interscience.wiley.com/cgi-bin/mrwhome/106568753/HOME?CRETRY=1&SRETRY=0

Data mining will be done by different groups of the consortium and do depend on the necessity of the simulation models for lung cancer and gliomas.

11 ContraCancrum Integrated Data Environment server

11.1 Introduction

As described in chapter 7.1 of this deliverable, the ContraCancrum Integrated Data Environment is the platform used for storing and retrieving data of any kind. The test deployment of the platform is available under at http://test.contracancrum.eu/. A screen shot of the interface of the development version of this tool is shown in figure 7.1. Personal data are only allowed to be uploaded after pseudo-anonymization. In most cases further processing of data is necessary before they can be used in the modelling scenarios.

This integrated data environment aims to serve four primary purposes:

1. To serve as a secure central repository for all pseudoanonymised data (clinical and computational) related to ContraCancrum,
2. To provide clinicians and researchers with an easily accessible open source internet-based platform from which they can upload, view, edit, retrieve, visualise and report clinical data and derived computational data,
3. Provide access to data and resources using standardised protocols and middleware, and wherever possible, lightweight protocols,
4. Provide access to computational resources via web-services and other standardised middleware, in particular to generate patient-specific results within clinically relevant time frames (Section 4).

This results in an integrated data environment that provides a rich and flexible data interface allowing clinicians to explore different patient-specific treatment scenarios and the validation through comparison to patient outcomes. By using multiple standard protocols to access data, researchers and clinicians can access the same data pool using their own methods. For example, the clinicians can upload imaging data to the central repository from the USAAR PACS workstations using the DICOM protocol, at which point researchers can immediately search this data and download DICOM images using ubiquitous HTTP via the central website.

This environment will ultimately be released to and validated for use by clinicians, therefore as new ideas in data representation and user interfaces are developed it is done in consultation with the project clinicians. Furthermore, the emphasis is on integrating data of different types and different scales. In the remainder of this section, various examples of data flows and interacting with data are discussed from the viewpoint of the environment implementation and usage.

11.2 Interacting with clinical data

Fundamental to the integrated environment is the ability for clinicians and researchers to interact, upload, edit and download data. But also the ability to generate reports and run simulations using clinical data within clinically relevant time frames (Section 4).

11.2.1 Accessing data

Users within the project are authenticated using a centralised identity server. Logins are provided to researchers and clinicians within the project by contacted the WP3 leader.

11.2.2 Displaying, adding and editing data

When a user first logs into the online integrated environment, they will be presented with a summary of the patients (identified by pseudonyms) that are stored the database. At a glance the summary will contain information about the type of data available (imaging data, histopathological data, molecular data, etc. as discussed in Section 3) for each patient. This also provides the clinicians who are supplying data with an overview of missing data.

A user can click through to a specific patient and see a more detailed overview of their clinical data, treatment plans, etc. At this point data can be added. For example, if a clinician wishes to add a new histopathological image to that patient's online records (microscope image of a tumour biopsy, for example), they can click on 'add file'. They are presented with an interface where they select the local file to upload, provide the relevant metadata such as date of acquisition and a description, and upload the file. Data is automatically added to the repository and it is then immediately available to other researchers for further study (cell counts, etc.).

11.2.3 Notification services

When using the environment over a period of time, clinicians and researchers will often want to be notified of events related to the data environment. For example, a researcher who is working on a group of patients may wish to be notified (by email, for example) of new data that is added to those patients, or a clinician may wish to be notified of new patient-specific molecular binding affinity reports which are ready. The technical platform provides a tailored subscription service such that users can receive notification of data being editing or added (related to particular patients and/or data types).

11.2.4 Visualising data

One fundamental aspect of integrating the different types of data is how the data can be displayed. Of value to clinicians is the ability to visually display a summary of data available, simulations related to that patient, as well as treatment schedules as well as patient-specific schedules designed as a result of simulations on the clinical data. The interface is designed to display traditional clinical data alongside computationally-derived patient-specific medical simulation data.

An example of an early implementation of a *timeline visualisation* is shown in Figure 11.1. This web interface queries the central project PACS repository and displays DICOM studies (MR, CT, CR and other imaging modalities) acquired over time for various glioma patients. The user can obtain further information on each study by clicking on its point on the timeline. This visualisation modality provides researchers and clinicians within the project with an immediate snapshot of the type of data available and frequency of acquisition. As soon as data is uploaded to the PACS repository it is available for viewing on this web interface.

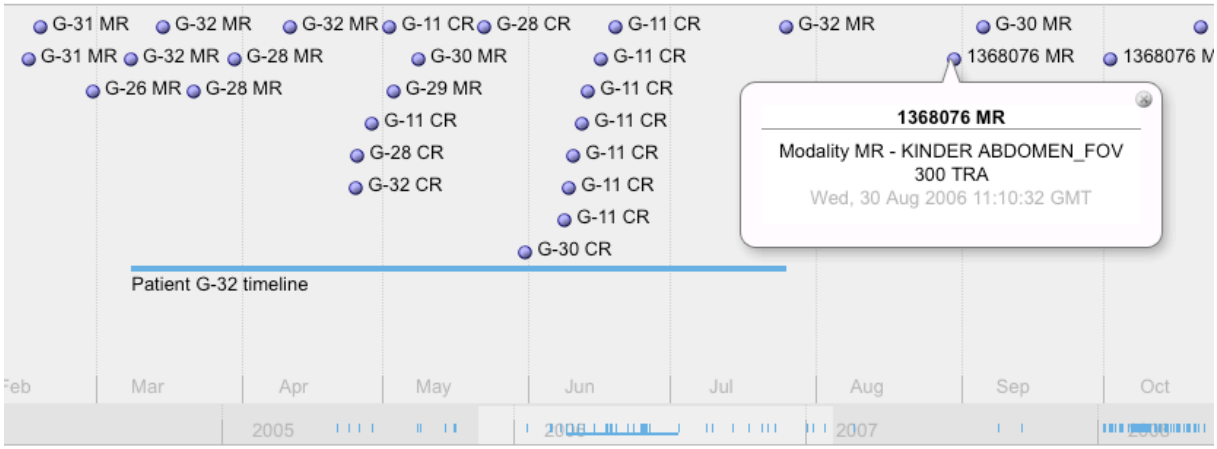

Fig 11.1 Timeline visualisation of imaging data acquisition. Each dot represents acqusition of a DICOM study from specific patients labelled 'G-11' etc., over time. Extra data, such as the study modality (CT - Computer Tomography, MR - Magnetic Resonance) and date are also shown.

11.3 Computable medical data and interacting with workflows

At the core of the integrated data environment lies the ability of clinicians and researchers to perform computations on clinical data. This is achieved by integrating the GSEngine workflow environment into the project technical environment (GSEngine has been discussed in detail in published deliverables D3.1 and D5.2). This workflow environment provides interfaces for researchers to develop what are termed virtual experiments, and publish them on the technical environment for other ContraCancrum members to refine and use.

The result of this is that the end-users (clinicians and researchers) can run computations on the data present in the technical environment. For example, when a user is viewing the DICOM data for a particular patient within the web interface, they can select a 3D DICOM series and perform a 3D segmentation. Once the segmentation is complete the user is notified by email that a VTK mesh file is ready for collection from within the environment (in the background this is performed using web-services or grid resources where necessary). A more complicated example is simulating patient-specific tumour growth in response to chemotherapy or radiotherapy (see published Deliverable 4.1 for more details). This requires macroscopic imaging data, microscopic imaging, molecular quantities and derived data, which are all stored within the central repository. A clinician can perform this virtual experiment by selecting the relevant data related to a specific patient within the web interface. The necessary computations are then performed in the background and the clinician is notified when they are complete. The outcome of the simulation shows the tumour growth in both non-treated and chemotherapy-treated cases as 3D and 4D graphs. Since this derived data is temporal and graphical in nature, it then interoperates with the existing medical data, and can be superimposed on graphs such as Figure 11.1, facilitating the comparison of actual predicted patient outcome.

11.4 Integration with ObTiMA and the ACGT-MO

The current back-end implementation of the database treats various patient-related events such as radio-diagnostic procedures, surgical procedures, biopsies etc., as unrelated events (although they all relate to the same patient). The implementation of the ACGT Master Ontology (Section 5) and relevant parts of the ObTiMA tool (Section 6) will ensure that *relationships* between events are concretely defined, for example - a microscope image obtained from a biopsy of a lung tumour. This will not significantly alter the technical environment interface, and will ensure more precise relationships between events, better data management and better interoperability.

12 Integrated scenarios

12.1 Introduction

Since a central feature of the ContraCancrum project is the development of an integrated multilevel tumour dynamics simulator a brief delineation of the latter is provided below. This may help in acquiring a broad picture of the entire ContraCancrum project and will also serve as the basis for the integration of several modeling modules into the ContraCancrum Integrated Simulator. The latter will be used for the execution of the simulation tasks as submitted by the end users. Logical and technical validation of the composite simulator (integrator) will be performed before the clinical testing, optimization and validation and will take place in tight interaction with workpackage WP9. It is noted that there will be a strong interaction of WP8 with all the other simulation workpackages throughout most of the project's lifetime in order to optimally orchestrate and harmonize the development of all modules to be finally combined or fused.

12.2 Biocomplexity level jumping

In order to go beyond the state-of–the-art it is important to take under consideration as much information as possible stemming from different levels of biocomplexity for each individual patient. For this reason the project proposes a workplan that integrates molecular models of cancer with models of the cellular and higher biocomplexity levels thus integrating different scales. Additionally, the models will be individualised by modelling tissue biomechanics and extracting anatomical and functional information using sophisticated medical image analysis techniques. Figure 12.1 gives an example of the "summarize and jump" strategy aiming at a pragmatic biocomplexity "level jumping" and integration[9].This approach will serve as the core philosophy for the multilevel integration of biological data and mechanisms involved in ContraCancrum cancer modelling:

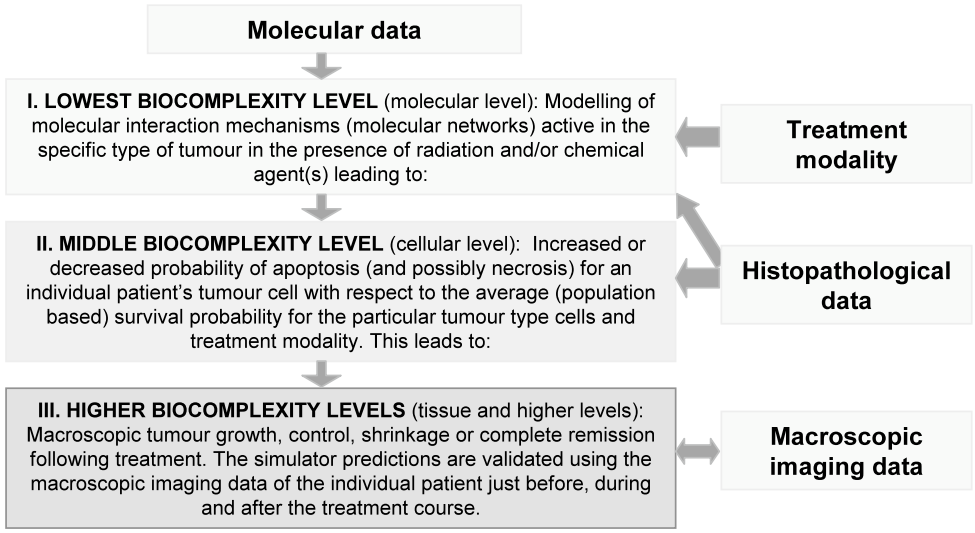

Figure 12.1. An example of the "summarize and jump" strategy as applied to the case of tumour response to teatment modlling in the clinical setting. Only three (clusters of) levels are depicted for simplification purposes.

[9] G. Stamatakos 2006, NCI CViT Ask the Expert Forum, password-controlled website https://www.cvit.org/node/128

Figure 12.1 outlines an example of the “summarize and jump” strategy aiming at a pragmatic biocomplexity “level jumping” and integration. The latter plays a fundamental role in the development of a multiscale simulator. This approach serves as the core philosophy for the multilevel integration of biological data and mechanisms involved in cancer modelling within the framework of ContraCancrum.

The ContraCancrum integrated simulation system will function briefly as follows. Provided that the system has been validated (retrospectively and prospectively) for a specific tumour type, the imaging, histopathological, molecular and clinical data of any given patient following pertinent preprocessing are introduced into the “MULTI-LEVEL CANCER SIMULATOR FOR TUMOUR AND NORMAL TISSUE RESPONSE SIMULATION” module. This module executes the simulation code for a defined candidate treatment scheme (Figure 12.2). The prediction is judged by the clinician and if a decision is made to test a further scheme *in silico* this is done in an analogous way. Alternatively a large number of candidate schemes can be executed concurrently on a cluster or grid platform. Finally the clinician decides on the optimal treatment scheme to be administered to the patient based on his or her formal medical education and knowledge and the predictions of the ContraCancrum integrated simulator. Subsequently comparison of the predictions with the real outcome provides a feedback signal to be exploited for the optimization of the ContraCancrum integrated simulator. The most fundamental processes to be implemented by ContraCancrum are the following: Processed molecular data is used in order to perturb the radiobiological or pharmacodynamic cell-kill parameters about their population-based mean values.

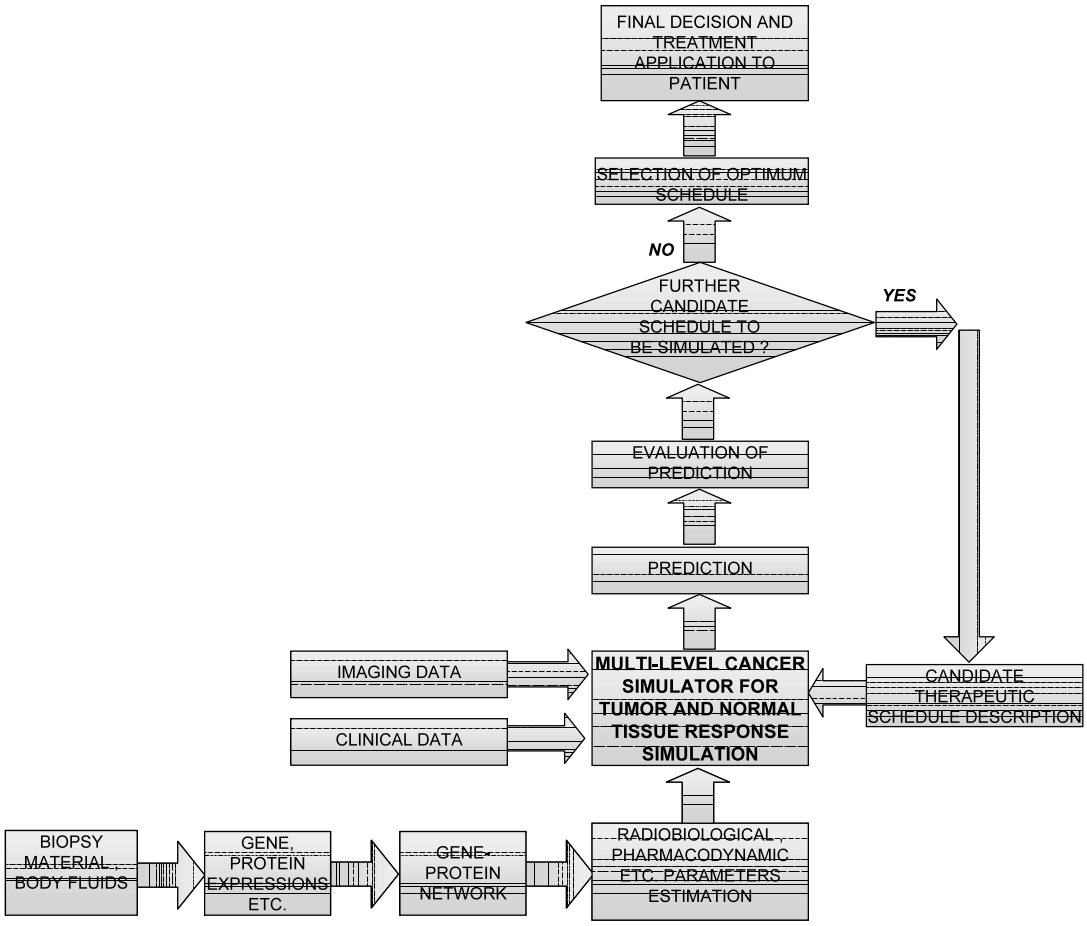

Figure 12.2. A generic functional outline of the ContraCancrum integrated simulator

At the heart of the simulation approach lies a prototype system of quantizing cell clusters included within each geometrical cell of a discretizing mesh, covering the anatomic area of interest. Cell-cycle phase durations and imaging-based metabolism distribution define the quantization equivalence classes considered. Several algorithms will be developed so as to simulate various macroscopic mechanisms such as tumour expansion or shrinkage and mechanics, as well as the effects of particular drugs and radiation on the tumorous and normal tissue under consideration.

13 Conclusion

This deliverable serves as a guideline for workflows of data needed in the project and is of great importance for all participants. It is described where and how to store data, as well as where to find data for the use in the simulation models for lung and gliomas. The complete data flow in the project will always be done according to European legal and ethical regulations. The integrated scenarios of ContraCancrum are described in this deliverable as well as in D2.2 & D9.1.

14 Appendices

Appendix 1 - Abbreviations and acronyms

CRF	Case Report Form
CTMS	Clinical Trial Management System
PDMS	Patient Data Monitoring System
RDE	Remote Data Entry
SOA	Service Oriented Architecture
UCL	University College London
UdS	University of the Saarland
DICOM	Digital Imaging and Communications in Medicine
PACS	Picture Archiving and Communication System
HTTP	Hypertext Transfer Protocol
SUSAR	Suspected Unexpected Serious Adverse Reaction
SAE	Serious Adverse Event
GCP	Good Clinical Practice

Appendix 2 – Master Protocol (contents)

This is the content of the Master Protocol that will be used for an ACGT clinico-genomic Trial

5.3 Study Drug
- 5.3.1 Description
- 5.3.2 Treatment Regimen
- 5.3.3 Method for Assigning Subjects to Treatment Groups
- 5.3.4 Preparation and Administration of Study Drug
- 5.3.5 Subject Compliance Monitoring
- 5.3.6 Prior and Concomitant Therapy
- 5.3.7 Packaging
- 5.3.8 Blinding of Study Drug
- 5.3.9 Receiving, Storage, Dispensing and Return

6. Study Procedures

6.1 Screening for Eligibility
6.2 Schedule of Measurements
6.3 Visit 1
6.4 Visit 2 etc.
6.5 Safety and Adverse Events
- 6.5.1 Safety and Compliance Monitoring
- 6.5.2 Medical Monitoring
  - 6.5.2.1 Investigator only
  - 6.5.2.2 expert to monitor
  - 6.5.2.3 Institutional Data and Safety Monitoring Board
  - 6.5.2.4 Data and Safety Monitoring Board
- 6.5.3 Definitions of Adverse Events
- 6.5.4 Classification of Events
  - 6.5.4.1 Relationship
  - 6.5.4.2 Severity
  - 6.5.4.3 Expectedness
- 6.5.5 Data Collection Procedures for Adverse Events
- 6.5.6 Reporting Procedures
- 6.5.7 Adverse Event Reporting Period
- 6.5.8 Post-study Adverse Event

6.6 Study Outcome Measurements and Ascertainment

7. Statistical Plan

7.1. Sample Size Determination and Power
7.2. Interim Monitoring and Early Stopping
7.3. Analysis Plan
7.4. Statistical Methods
7.5. Missing Outcome Data
7.6. Unblinding Procedures

8. Data Handling and Record Keeping

8.1. Confidentiality and Security
8.2. Training
8.3. Case Report Forms and Source Documents
8.4. Records Retention
8.5. Performance Monitoring

9. Study Monitoring, Auditing, and Inspecting

9.1. Study Monitoring Plan
9.2. Auditing and Inspecting

10. Study Administration

10.1. Organization and Participating Centres
10.2. Funding Source and Conflicts of Interest
10.3. Committees
10.4. Subject Stipends or Payments
10.5. Study Timetable

11. Publication Plan

12. Attachments

12.1. Tables
12.2. Informed consent documents
12.3. Patient education brochures
12.4. Special procedures protocols
12.5. Questionnaires or surveys

13. References