Final Report on the clinical benefits delivered by the ACGT project

Project Number: FP6-2005-IST-026996

Deliverable id: D12.7

Deliverable name: Final Report on the clinical benefits delivered by the ACGT project

Submission Date: 07/09/2010

COVER AND CONTROL PAGE OF DOCUMENT	
Project Acronym:	ACGT
Project Full Name:	Advancing Clinico-Genomic Clinical Trials on Cancer: Open Grid Services for improving Medical Knowledge Discovery
Document id:	D12.7
Document name:	Final Report on the clinical benefits delivered by the ACGT project
Document type (PU, INT, RE)	PU
Version:	3.0
Submission date:	07/09/2010
Editor: Organisation: Email:	Christine Desmedt IJB christine.desmedt@bordet.be

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

ABSTRACT:

This document presents the ACGT deliverable **D12.7**: *Final Report on the clinical benefits delivered by the ACGT project.* To evaluate the clinical benefits of ACGT, we have used the TOP trial as an example, as it will also be done in the final demonstration. This deliverable will present and discuss several procedures and tools set up during the *ACGT* project in the context of the TOP trial. In practice, this deliverable unfolds into eleven- (11) main chapters. The eight first chapters will describe some tools, procedures and results obtained in the context of the TOP trial and the final two chapters will report on conclusions and perspectives of ACGT.

KEYWORD LIST: cancer, breast cancer, clinical trials, final demonstration

MODIFICATION CONTROL			
Version	Date	Status	Author
1.0	12/08/2010	Draft	C. Desmedt
2.0	23/08/2010	Pre-final	C. Desmedt
3.0	07/09/2010	Final	C. Desmedt

List of Contributors

- **Alberto Anguita,** UPM
- **Nikolaus Forgo,** University of Hannover
- **Brecht Claerhout,** Custodix
- **Stefan Rueping,** Frauenhoffer Institute
- **David Bernasconi,** Swiss Institute of Bioinformatics
- **Georgios Stamatakos**, National Technical University of Athens
- **Andreas Persidis,** Biovista
- **Regine Kollek,** University of Hamburg
- **Anca Bucur,** Philips Research
- **Manolis Tsiknakis,** FORTH
- **Norbert Graf,** Universitätsklinikum des Saarlandes
- **Christine Desmedt**, IJB, Brussels, Belgium

Contents

Figures

1. Executive Summary

This deliverable will evaluate the clinical benefits of ACGT. We have used the TOP trial as an example, as it will also be done in the final demonstration. This demonstration has been designed as a step by step process simulating how the TOP trial data is introduced and analyzed in the ACGT infrastructure to identify the targeted biomarkers.

In practice, we will first describe and illustrate the **legal framework** which is necessary to take into account the needs of modern scientific genetic research and the needs of the patients regarding data protection and privacy. This legal framework specifically involves the informed consent procedure which was carried out in the TOP trial, the collaborative contracts which had to be developed and agreed by the data providers and data users, and the anonymization of the patient data.

Second, we will illustrate the importance of the ***ACGT* Master Ontology** for the semantic integration of heterogeneous data (clinical, imaging, genomic, proteomic, …).

Third, we will demonstrate how tools developed within *ACGT* can facilitate the **identification of predictive markers of response/resistance** for anthracyclines chemotherapy using microarray-based gene expression profiling as well as genotyping technology.

Fourth, we will report the progresses and advances made by the **in silico oncology** working group. This group evaluates the reliability of in silico modelling as a tool for assessing alternative cancer treatment strategies; especially in the case of combining and utilizing mixed clinical, imaging and genomic/genetic information and data.

Since, until now, we were lacking **patients' views on and experience of involvement in clinico-genomic research**, Grid structures, and Europe-wide data flows, we will finish by reporting the results of the large empirical survey, carried out within *ACGT*, on perspectives and needs of persons who did consent to take part in tissue-based cancer research in several European settings.

Finally, we will report on perspectives of ACGT in the context of new clinico-genomic trails and conclude on the lessons learned from ACGT.

2. The TOP trial

The ultimate objective of the *ACGT* project is the provision of a unified technological infrastructure which will facilitate the seamless and secure access and analysis of multi-level clinical and genomic data enriched with high-performing knowledge discovery operations and services, in the concrete setting of **clinical trials** on **Cancer**. Pilot trials have been selected based on the presence of clear research objectives, raising the need to integrate data at all levels of the human being. This integrative view underlies the development of clinico-genomic models, showing that the combination of biomarkers and clinical factors are most relevant in terms of statistical fit and also, more practically, in terms of cross-validation predictive accuracy

The **TOP trial**, a trial which aims at identifying molecular markers that predict response/resistance to one of the most commonly administered chemotherapies in breast cancer, is one of those pilot trials for ACGT and has been chosen for the final demonstration of *ACGT*.

To date very little progress has been achieved in the field of biomarkers predictive of chemotherapy benefit in breast cancer. Consequently, the vast majority of patients considered to be at moderate or high risk of relapse are treated with the cytotoxic agents viewed as the most active “on average”, namely anthracyclines and taxanes chemotherapy agents. These drugs have significant side effects, the most worrisome of which are secondary leukemias and irreversible congestive heart failure for anthracyclines, and slowly reversible neurotoxicity for the taxanes.

In the pre-operative international TOP trial presented here, we focus on identifying molecular markers that predict response/resistance to anthracyclines, one of the most commonly administered chemotherapies in breast cancer (Figure 1).

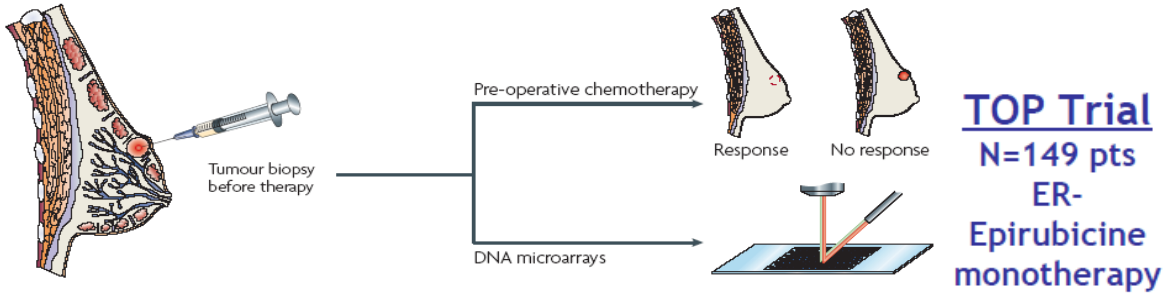

Figure 1: Schematic illustration of the TOP trial

Interestingly, the TOP trial avoided any possible confusion by including only women being treated with single-agent epirubicin. We also limited the patient population to women whose tumours did not over-express the estrogen receptor (ER). This is because, for these women, chemotherapy can stop the ovaries from working, offering an additional benefit that may distort the results.

The aim of the trial was to carry out the first prospective evaluation of the predictive value of topoisomerase II alpha (TOP2A) gene aberrations and expression. TOP2A is a key enzyme in DNA replication, one of the molecular targets of anthracyclines, and it is amplified in 24% to 54% of HER2-amplified tumors. Although TOP2A is considered by some investigators to be a promising marker for predicting the activity of anthracycline-based regimens, inconsistent results have been reported about TOP2A amplification/expression and response to anthracyclines.

The study protocol also included exploratory analysis to identify gene expression signatures that correlate with pCR. We therefore aimed to develop a gene expression signature (the Anthracycline-Score, Figure 1) to identify those patients who would *not* benefit from anthracyclines and could therefore be spared the non-negligible risks of this type of chemotherapy.

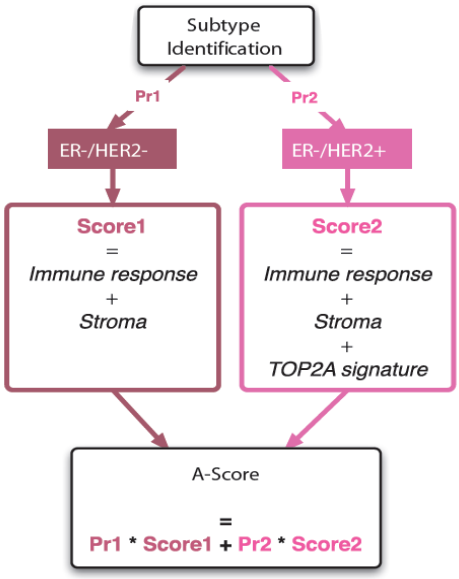

Figure 2: Schematic representation of the Anthracycline-Score. Pr1 and Pr2 is the posterior probability of a tumor to be ER-negative/HER2-negative and ER-negative/HER2-positive respectively.

The results of this trial are currently under second revision at the Journal of Clinical Oncology.

3. The re-consent procedure in the context of the TOP trial

Today, the doctrine of informed consent is widely acknowledged as one of the main principles in ethics and bioethics by protecting persons concerned and their fundamental rights to integrity and self-determination in medical interventions. In principle, the doctrine states that any medical treatment as well as scientific research involving human subjects is only to be carried with the prior, free, and informed consent of the person concerned. Therefore, it is from an ethical point of view indispensable that the patient concerned is informed adequately prior to his consent, that the consent is explicitly expressed and may be withdrawn at any time and for any reason without disadvantage or prejudice. Besides this, the informed consent is one way, among others, to legitimate the data processing needed in the ACGT project in the sense of the European Data Protection Directive 95/46/EC. From a data protection point of view, there are several legal preconditions that have to be fulfilled to make an informed consent valid. However, these preconditions are very similar to the ethical ones. The consent has to be explicit, freely given, e.g. not led by external influences, for a specific case and in awareness of the factual situation, which means, it has to be an "informed" consent.

The consent to the processing of the patients' data within ACGT therefore was necessary for two reasons. Firstly, consent is required from an ethical point of view, as the consent to the processing safeguards the patients' autonomy. Secondly, the consent serves as a fallback-scenario if ever the context of anonymity within the ACGT data protection framework is impaired and therefore the data cannot be considered being de facto anonymous.

In more specific terms, the data that will be used for the final demonstration is that of patients of the Institute Jules Bordet. These patients consented to participate in the TOP trial. However, the consent to participate in the trial at the hospital did not include the processing of the patients' data within ACGT. Therefore, the patients that had already consented to participate in the TOP trial were additionally asked to consent to the processing of their data within ACGT. Therefore, they were asked to re-consent.

At the time being, 67 patients gave their consent for having their data shared in the context of ACGT and only 2 patients refused.

4. Collaborative contracts

The essential idea of the ACGT Data Protection Framework is the implementation of a safety net that ensures compliance with data protection regulation on different levels. The first step of this safety net is the de facto anonymisation of all patient data processed within the ACGT network. To reach this goal it is essential to create an environment that ensures contextual anonymity. This has been achieved among others by the signing of contracts by all participants ensuring the compliance with the ACGT data protection and data security policies.

The contracts have to be seen in the context of “The ACGT ethical and legal requirements” (D 10.2). The Data Protection architecture for data flows within ACGT is set up with the prior aim to work with anonymous data wherever this is possible. Anonymisation is the best way to protect patients’ privacy. The architecture of data flows combined with data security measures as well as the contracts guarantee that data that is used, stored and exchanged within ACGT is de facto anonymous. To ensure compliance with this Data Protection Framework it is necessary to bind all participating partners by contracts to the ACGT policies and procedures. However, to put ACGT into the position to be able to conclude contracts with the participating partners, at first it was essential to establish a legal body that will be able to conclude such contracts. This was achieved by founding the “Center for Data Protection” (CDP), a non profit organisation under Belgian law.

Two contracts had to be signed. On the one hand the “Data Transfer Agreement” dealing with the transfer of patient data to the ACGT network was concluded between the CDP and the Institute Jules Bordet as the healthcare organisation delivering patient data. On the other hand the “Contract on data protection and data security within ACGT” concerning the data processing within the ACGT network was concluded between the CDP and all ACGT end-users doing research on this data.

5. Anonymization of the data

Data Export and Anonymisation

Before data can be accepted into the ACGT environment it needs to be properly de-identified according to the rules set out in the ACGT contracts (cf. WP10 deliverables). One of the tools that ACGT provides for data protection compliance is CAT (Custodix Anonymisation Tool). This tool was used for de-identifying the TOP trial data.

It was decided that Custodix would perform the actual de-identification with the CAT tool, i.e. the configuration of the "privacy protection profile" (see a.o. the autumn 2008 newsletter of ACGT for the working principles behind CAT[1]). The TOP-trial data consisting of "CEL" and "CSV"-formatted files, was thus sent to Custodix. The data was to be hosted on the ACGT environment by UPM.

The privacy protection profiles of the files are simple as the personal content is rather limited.

The CEL files contain two types of identifying data, the patient's identifier and the sample date.

Figure 3 shows the CAT data model that represents this. Subsequent to the definition of the generic data model in CAT, the privacy profile (processing of data) is defined (shown on Figure 4). During de-identification of the CEL files, CAT replaces the patient identifiers by a random pseudonym (a same patient is always mapped to the same pseudonym). The dates are made relative to a fixed reference date[2].

[1] http://eu-acgt.org/fileadmin/newsletter/autumn2008/Acgt_Newsletter.html, article "Grid news: Custodix Anonymisation Tool (CAT)"

[2] A better approach is to make dates relative versus the date of birth (DOB) of a patient. However, DOB was not present in the previous set of TOP-trial data and use of the same data protection approach was desired for this iteration.

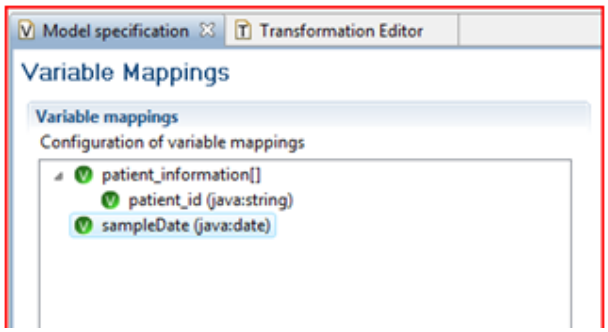

Figure 3: Model for CEL Files

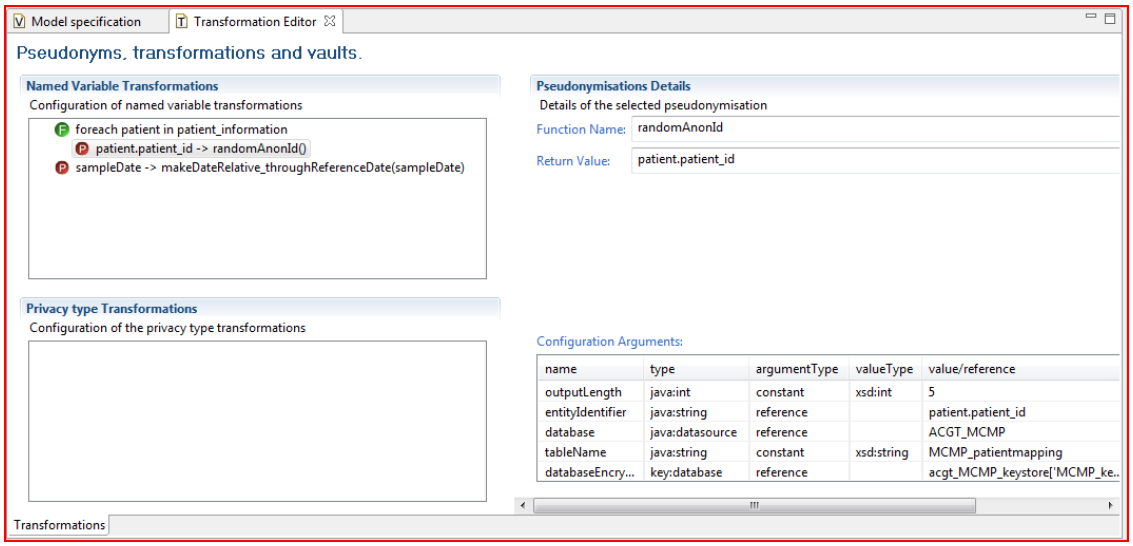

Figure 4: Privacy transformations for the CEL files

Two CSV formatted datasets were delivered, one containing only the patient identifiers as identifying information, and one ” containing a column with the patient identifier, one with the inclusion date and one with the patient’s date of birth.

The former was processed similarly to the CEL files. CAT was configured to process the latter by replacing the identifier with the same random pseudonym as used in the CEL files and the above CSV. The inclusion date is processed like a sample date. This is by making it relative to the same reference date. The patient’s date of birth is removed.

After the de-identification process, the data was transferred to UPM where it is hosted (in the ACGT infrastructure).

Access Management

The TOP trial data classifies as sensitive data that should only be made available to a restricted set of people (cf. section on legal aspects). Access management in the ACGT environment is achieved through the use of Virtual Organizations (cf. D3.3 “The ACGT technical architecture: Final Specifications”.). A Virtual Organization (VO) refers to a dynamic set of people who have decided to share resources across organizational boundaries around an agreed set of sharing rules and conditions. Within ACGT, the different dynamic groups of collaborating partners that originate from the different trials and research initiatives are reflected in the formation of such “virtual” organizations.

The TOP trial demonstration VO has a simple setup with all VO participants having the same access rights. The “TOP-Trial_VO” was setup by a VO manager according to the procedures explained in D11.6 “ACGT guide with administrative documentation of ACGT security and VO management”. The TOP trial dataset has been made available on the VO by the resource manager of the hosting organization (in this case UPM) through the ACGT OGSA DAI data wrappers (the resource manager granted “the VO”, i.e. everyone belonging to it, access to the OGSADAI Resource defined as “Local-TOP-Database”).

6. Semantic integration of the TOP trial in the ACGT platform

The demonstration involves the testing of the ACGT Platform with part of the database of the TOP clinical trial. The goal was to use the tools provided by the platform to analyze the data from the Case Report Forms (CRFs)—actually, a subset of it—to obtain new knowledge. In order to set up the clinical data from the TOP trial in the ACGT platform, a series of steps were performed. These steps are listed below:

1. Selection of key fields for the demo.
2. Setting up of a database containing the selected fields.
3. Mapping of the database with the ACGT Master Ontology.
4. Query definition.

The next subsections describe in detail the work performed to achieve these steps.

Selection of key fields for the demo

The CRFs of the TOP trial contained over two hundred fields. In an ideal case, we would have built a database resembling the CRFs in every field. However, for practical reasons we decided to select just a few fields that would allow us to demo the ACGT infrastructure with the TOP trial. These fields were picked up in the preparation stage of the demo, as a fruit of discussions between the trial chairman and biostatisticians working in ACGT. The goal was to pick the minimum set of fields that would allow extracting some valuable knowledge when analyzed together with genetic data, while at the same time fully exploiting the potential of our technological platform. Finally, eight fields were selected for the demo, namely: i) the patient's identifier in the trial, ii) the patient's birth data, iii) the patient's diagnosis date, iv) the tumor's histopathologic grade, v) the tumor's T classification, vi) the tumor's N classification, vii) the end of treatment reason and viii) the pathological complete response status. The first field was needed to establish a link of the clinical data with the genomic data. The second and third fields allowed computing the patient's age at the diagnosis time. The rest of the fields gave details of how the treatment had resulted in each patient.

Setting up of a database containing the selected fields

The second step consisted on setting up a database that would store the data of the selected key fields. As this database would be access by the ACGT Semantic Mediator, it had to be configured as a database wrapper. A simple RDF schema was designed for the database, and a new OGSADAI resource was configured for storing it. This data resource was physically deployed at UPM machines.

Given that the data to store in the database included sensitive fields, the wrapper was configured to work in secured mode, collaborating with the GAS Security framework to ensure credential-based data accesses. A new VO was set up for the data resource, including only the credentials of the people that were allowed to access the data—in this case, the partners that signed the confidentiality contracts.

Mapping of the database with the ACGT Master Ontology

After setting up the new database, we configured the semantic mediation layer to be able to access it. This meant creating a new mapping of the database with the Master Ontology. For this task, MO experts collaborated with the trial clinicians to identify the semantic equivalences in both schemas. During this process no missing concepts were identified in the MO, so there was no need for updating it. As a result, a new mapping document was submitted to the Semantic Mediator, enabling access to the TOP data from data analysis tools from the ACGT platform.

Query definition

Having set up the access to the TOP data through the Semantic Mediator, the final step was to define the necessary queries to retrieve the data needed by the analysis tools. One single query was designed to retrieve all fields from the database. The ACGT Query Tool was used for this task, just like an end user would have done. The final query was uploaded to the query repository, so workflows designed in the ACGT Workflow Editor could use it. This query was successfully tested by generating an initial workflow that contained it. The data was correctly retrieved from through the Semantic Mediator, while it was confirmed that only users included in the VO for this data source were able to run the workflow correctly.

7. Identification of molecular markers associated with the efficacy of treatment

Overview

One of the scientific objectives of the TOP scenario is to find a list of genes whose RNA expression or DNA aberration is significantly associated with response/resistance to the epirubicin pre-operative treatment and verify the predictive capacity of TOP2A as a marker for response/resistance.

In order to achieve this objective, data from the 65 patients from the TOP trial who re-consented (cfr Chapter 3) are analyzed. Genes identified as significantly associated with response are then compared with relevant publications through the literature mining tool.

In an ACGT's point of view, the objective for this scenario is to demonstrate the ability of the tool to work with various types of data. Indeed, this is the first scenario using SNP microarray data in the analysis workflow. It also illustrates the possibility of integration of external services such as command line tools or literature mining tools within the ACGT analysis environment.

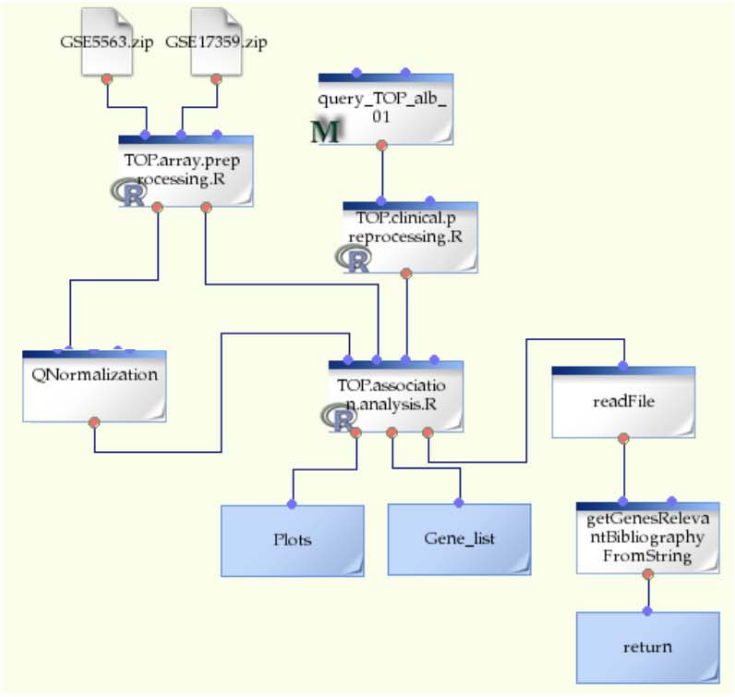

Figure 5: The TOP scenario's workflow

Data pools used in the scenario

The scenario takes advantages of three types of data, namely clinical data, gene-expression microarray data, and SNP (single nucleotide polymorphism) microarray data.

- Clinical trial data accessed via mediator queries: For this scenario, 7 queries are needed for constructing the clinical information matrix. The fields required for the analysis are the following:

1. Patient's identifier in the trial
2. Patient's birth data
3. Patient's diagnosis date
4. Tumor's histopathologic grade
5. Tumor's T classification,
6. Tumor's N classification
7. End of treatment reason
8. Pathological complete response status

- Gene-expression microarray data retrieved from DMS: Gene-expression information come from an Affymetrix GeneChip Human Genome U133 Plus 2.0 array and are stored in the DMS in compressed zip directories

- SNP microarray data retrieved from the DMS: SNP information come from an Affymetrix Genome-Wide Human SNP Array 6.0 and are stored in the DMS in compressed zip directories.

Analysis pipeline

In the scientific analysis pipeline, we attempt to find a list of genes whose expression or DNA aberration is associated with response/resistance to the preoperative epirubicin treatment using clinical data and gene-expression data from 65 patients of the TOP clinical trial. The specificity of this scenario is the inclusion of SNP (single nucleotide data) in the analysis process. These data are used to assess the association between gene-expression and copy number variants (CNV) in the genome. The second specificity of the scenario is the integration of external services in the workflow.

In order to fulfill the analysis, the workflow is composed by three R scripts:

1. **TOP.array.preprocessing.R**: In this R script microarray data are prepared for the further analysis steps. Microarray .CEL files are stored in a compressed directory. The R script download, uncompress and store the microarray data in the current working directory. Gene-expression intensities then are computed but not normalized as this step is performed by the command line tool (QNormalization box in Figure 1). SNP intensities data are computed, normalized and summarized according to the response to the pre-operative treatment. Two intensities matrices constitute the output of the script.

2. **TOP.clinical.preprocessing.R**: This R script prepares the clinical information gathered by mediator queries under the form of a matrix in order to be used in the following steps of the analysis.

3. **TOP.association.analysis.R**: This R script investigates the association between the genomic data and the clinical data. Inputs of this R script are microarrays normalized intensities as well as clinical data.

The workflow also contains two external services: a command line tool and a literature mining tool. These tools not only illustrate the capacity of integration of third party tools within the ACGT environment but also give insight of the relevance of our results by comparing our list of genes with the relevant publications.

1. **The command line tool:** A web service executing a QNormalization of the gene-expression microarrays using several processors (MPI library).

2. **The literature mining tool:** A literature service from Biovista returning the publications referencing one or more of the genes given as input. The first 100 results are returned.

Analysis output:

Outputs of this analysis workflow are the following:

1. List of genes whose expression is significantly associated with response/resistance to the treatment. The genes are displayed in a html file and ordered according to the significance of their association with the response. For each gene reported, a log2 ratio of the CNV variant is presented in order to assess the relation between gene-

expression and gene copy number.

2. Heatmap of the identified genes.

3. Gene copy number comparison between good and bad responders. For each chromosome, log2 ratios of intensities for each SNP are reported. The plots allow the identification of chromosomal region where copy number variation may be related with response to the treatment.

4. Literature associated with identified genes. The final workflow's output consists in a list of publication related with the list of genes identified as associated with the response to the treatment. The list of publication contains hyperlink to the Pubmed database.

8. The Oncosimulator

The In Silico Oncology Group from National Technical University of Athens recently published some interesting work and results based on the TOP data in the context of ACGT[3].

In this paper, an advanced, clinically oriented multiscale cancer model of breast tumour response to chemotherapy is presented. The paradigm of early breast cancer treated by epirubicin in the context of the TOP trial has been addressed. The model, stemming from previous work of the In Silico Oncology Group is characterized by several crucial new features, such as the explicit distinction of proliferating cells into stem cells of infinite mitotic potential and cells of limited proliferative capacity, an advanced generic cytokinetic model and an improved tumour constitution initialization technique. A sensitivity analysis regarding critical parameters of the model has revealed their effect on the behaviour of the biological system. The favourable outcome of an initial step towards the clinical adaptation and validation of the simulation model, based on the use of anonymized data from the TOP clinical trial, is presented and discussed. Two real clinical cases from the TOP trial with variable molecular profile have been simulated. A realistic time course of the tumour diameter and a reduction in tumour size in agreement with the clinical data has been achieved for both cases by selection of reasonable model parameter values, thus demonstrating a possible adaptation process of the model to real clinical trial data. Available imaging, histological, molecular and treatment data are exploited by the model in order to strengthen patient individualization modelling. The expected use of the model following thorough clinical adaptation, optimization and validation is to simulate either several candidate treatment schemes for a particular patient and support the selection of the optimal one or to simulate the expected extent of tumour shrinkage for a given time instant and decide on the adequacy or not of the simulated scheme. The main parts of this work will be presented in the context of the final demonstration.

[3] Stamatakos GS, Kolokotroni EA, Dionysiou DD, Georgiadi ECh, Desmedt C. An advanced discrete state-discrete event multiscale simulation model of the response of a solid tumor to chemotherapy: Mimicking a clinical study. J Theor Biol. 2010

9. Patient's attitudes towards clinico-genomic research

A growing number of studies in tissue-based research aim to explore the roles of genes and gene activities in order to improve treatment and prognosis of cancer. Such studies conducted with patients currently affected by cancer raise a number of questions concerning informed consent and the attitude of research subjects towards the handling and processing of data and of data protection. These questions have been discussed intensively in the theoretical discourse on ethical and legal aspects of modern biomedical research. We are, however, unaware of empirical reports on the participants' views on and experience of involvement in clinico-genomic research, Grid structures, and Europe-wide data flows.

We have therefore designed an empirical survey on perspectives and needs of persons who did consent to take part in tissue-based cancer research in several European settings (Great Britain, Belgium, Germany, and Greece). In particular, the survey aims at elucidating patients' understanding of and motivation for taking part in tissue-based research, their attitudes towards future research, their expectations concerning confidentiality of medical information and the feedback of study findings.

The study population comprises breast cancer patients who gave a tumour sample for tissue analysis (e.g. profiling of gene expression and proteomics) and are usually treated in clinical trials (especially patients involved in the MINDACT study organized by EORCT[4]). To get access to cohorts of different European countries we have co-operated with the clinical partners within ACGT and with the "West-German Study Group" which coordinates the MINDACT study at several breast cancer centres in Germany. In Belgium, data were collected in co-operation with the Jules Bordet Institute. The questionnaires were distributed from February 19, 2009 until November 2, 2009. Data of 159 completed questionnaires are processed and analysed. The respondents of were breast cancer patients; over half of them were initially diagnosed between 2003 and 2005. 67% were aged 50 years and older. Most of them were married and had one or two children. One-fourth had no school-leaving certificate, whereas one-fifth had a school-leaving certificate. One-fourth finished an apprenticeship and one-fourth had a university degree.

[4] For further information please visit: http://www.eortc.be/protoc/details.asp?protocol=10041

Patients' memory of consenting to participate in the clinical study and motivation for consenting

About 92% of the patients remembered that they were inscribed in a clinical trial on breast cancer, whereas 8% denied that they were participants or were uncertain. According to the information of Jules Bordet Institute, all respondents were in fact participants (most of them participated in the TOP trial). From this follows that only 8% of the respondents have not appropriately remembered that they took part in a clinical study. 15% of the participants did not remember that the consent or non-consent respectively is documented in writing; 90% felt adequately informed in the consent process. However, we wanted to verify how well informed the respondents were about the tissue and data use in breast cancer research and asked for the different types of data that are generally used in tissue-based research. 52% had the opinion that socio-demographic data are not used in research, 56% thought that biological data and 57% thought that genomic data are not used in research. That is, of course, a misconception of the data processing in tissue-base research, in particular in clinico-genomic trials. In addition, only 60% of the respondents confirmed that tumour or blood analyses may provide information about a patient's hereditary condition.

The respondents had different motives for participating in research. The percentages are more or less evenly distributed across the given reasons to take part in the clinical trial. Outstanding were only the altruistic statements *"make a contribution to medical progress"* and *"benefit other patients"* with nearly 90% acceptance.

Nearly half of the respondents took the view that researchers should ask the patient again if researchers want to use stored medical data in future research projects besides the one for which they have given consent. About 30% of the respondents only wanted to be asked again if the data are not anonymously stored. However, 20% did not want to be asked for re-consenting if the data is used in future research. This result is in line with the subsequent question concerning the surrogate consent by an ethics committee. 20% of the respondents did not accept the vote of an ethics committee instead of personal consent; 27% would agree with the surrogate consent, and 55% were uncertain regarding this issue.

Attitudes of patients towards data protection

In the Belgian survey, almost half of the patients were uncertain whether or not the existing laws adequately protect medical records. However, 40% of them thought that medical data are protected by law and 12% thought they are not. Asked for groups and institutions separately, the respondents expected that hospital workers (74%), relatives (48%), and public health authorities (44%) are able to get access to patient records without his or her knowledge. However, they did not agree that medical information is passed to someone without the patient's permission. In particular, the transfer of medical data to the

public and professional sphere without consent was not accepted by nearly 90% of the respondents.

Attitudes of parents towards the communication of research results

Only 12% of the respondents did not want to be informed about aggregate research results; the overwhelming majority (88%) would like to receive this kind of information. 60% of the respondents preferred to receive aggregate results via information letters by mail. One-third of them wanted to be informed via a website on the internet and almost one-third wanted the information via a flyer that is available in the public (e.g. at the clinic). Only few respondents would like to receive aggregate research results at a meeting (13%) or by reading scientific journals (18%). Fewer patients – only 60% – would definitely like to be informed about individual research results, whereas only a handful of respondents (0,7%) would not like to be informed at all. Others wanted to be informed only if the information is validated (20%) or if treatment or preventive intervention is available (13%). 65% of the respondents preferred to be informed about individual results by the attending physician, whereas about one-fourth of them wanted to receive this information by mail.

In D10.6.2 ("First results of the international and national empirical survey on patients' and parents' perspectives and needs"), we have given a first insight into the results of the survey on patients' and parents' perspectives, preferences and needs regarding informed consent and data protection. As the presented data of the Belgian sample show, it provides some unexpected results even at this early stage of data analysis. Hence, these first results illustrate that the survey on patients' attitudes and expectations will deliver important information towards the implementation of informed consent, data protection, and disclosure in ACGT and comparable projects.

10. Reality check with the neoBIG requirements

Besides the demonstration of tools, procedures and results in the context of the TOP trial, we also have carried out a reality check of the different ACGT tools and procedures in the context of a new family of clinical trials.

We therefore have investigated how the expertise, and potentially also the tools, developed in ACGT could be used to support a large real-life multi-centric clinical trials program, such as NeoBIG, the new research program of the Breast International Group (BIG). To suit the ACGT scope, our focus was on the IT needs of the neoBIG research program, specifically with respect to secure privacy-preserving data management and sharing as these are issues at the core of ACGT. We have evaluated the suitability of our solutions by first collecting and analyzing the requirements of BIG concerning the data sharing platform needed to support their future clinical trials, and based on that briefly evaluating potential alternatives in which ACGT could support this program by making use of ACGT tools and infrastructure, but also of relevant expertise.

We have collected requirements by carrying out interviews and discussions with the main stakeholders of the data sharing platform, i.e. with representatives of BIG for the clinical aspects and with representatives of the Breast European Adjuvant Studies Team (BrEAST), the data centre of the BIG, for an IT perspective. Central to the discussions were requirements and scenarios concerning the building of a data sharing platform to support the NeoBIG program of the Breast International Group, that will be sustainable far beyond the NeoBIG program and provide the clinical research community with common methodologies and standards, data models and consolidated datasets that could be used for further research.

To provide a platform that enables data sharing and collaboration between cancer research centres, NeoBIG requires a robust, secure IT solution that is compliant with a wide set of regulations and laws in the context of security, safety and privacy protection. The platform needs to be able to store, manage, and share the various types of data that will be generated by NeoBIG trials.

Security is an important aspect of the NeoBIG data sharing infrastructure. NeoBIG deals with personal data obtained from patients, whose privacy needs to be protected (both from an ethical and a legal perspective). Secondly, future prospective clinical trials with targeted therapies will require a system capable of dynamically setting up collaborations of organizations around specific data sets. Data shared within such a group needs to be well protected. Therefore, the NeoBIG data sharing platform needs to assure secure data sharing, such as authentication of users (secure logon), authorization (access control),

encryption (to guarantee confidentiality), trust establishment, and Virtual Organization Management. The interactions with the NeoBIG data sharing platform need to be fully audited to enable traceability.

Strong requirements on the data sharing platform are production-level reliability and availability and full maintenance. The data sharing platform will be productively used by the BIG community and needs to be available long beyond the end of the clinical trials, as the data is highly valuable for further research. Additionally, data interoperability and adherence to widely accepted international standards are important requirements which will enable the collaboration between BIG and other cancer organizations world-wide. In that context, well-known standards (HL7, DICOM, MIAME, MAGE, etc.) and terminologies (SNOMED, LOINC, etc.) are relevant, but also new standards emerging with the development and adoption by the US research community of relevant NeoBIG tools.

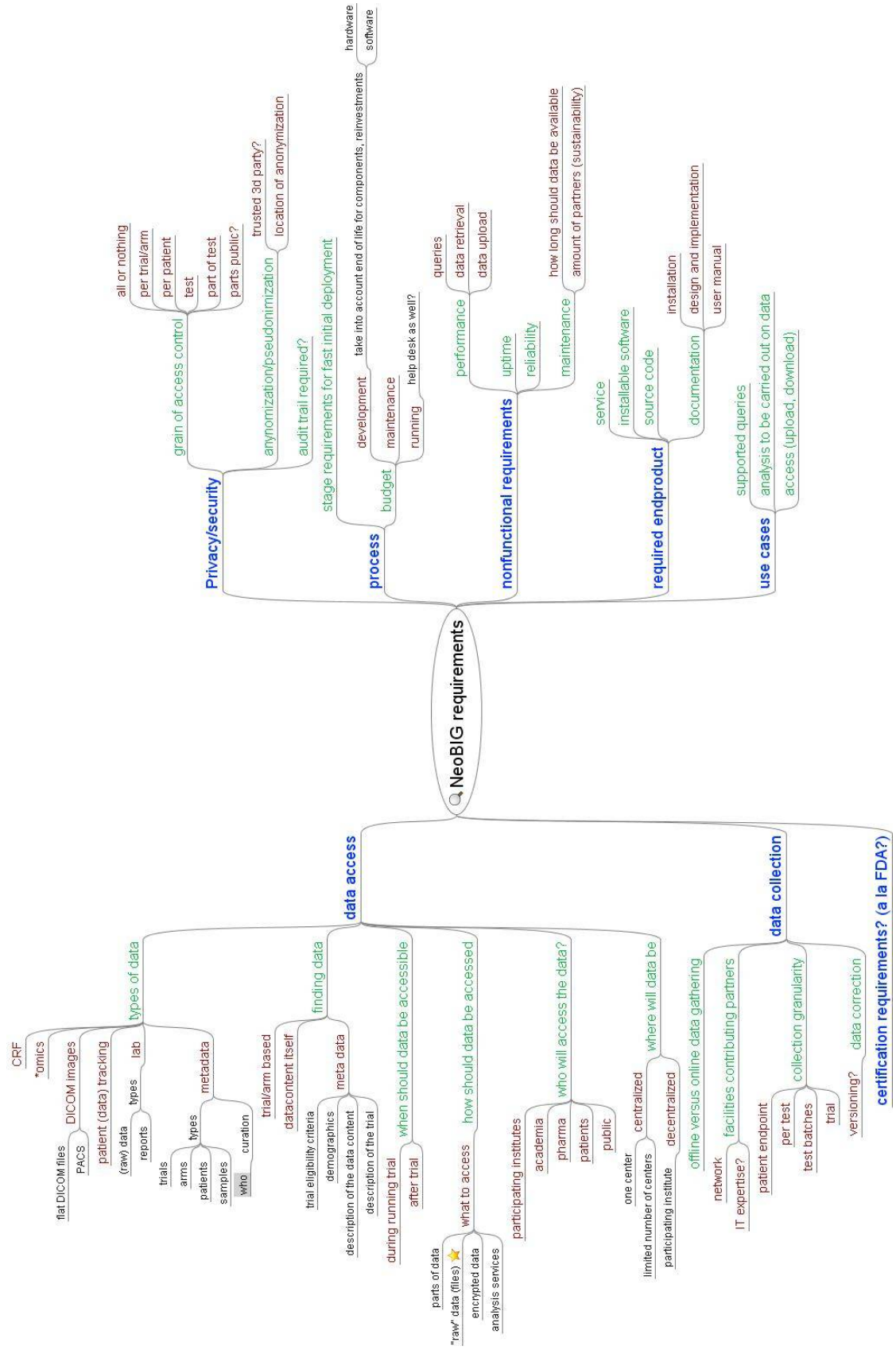

Figure 6: Requirements of the NeoBIG program

As collaboration with the US cancer research community is desired and the US market is important for the pharma organizations participating in the NeoBIG trials, additional requirements need to be extracted from regulatory frameworks (such as FDA 21 CFR part 11) to which compliance needs to be assured. The figure above identifies all the aspects that we have considered relevant in the requirements collection.

During the study, we have concluded that there is a lot of ACGT expertise that could be used for the NeoBIG data sharing platform, especially with respect to data storage, management and sharing, VO management, and with respect to privacy and security. At the same time, we have understood that while accessing external data out of heterogeneous repositories is highly relevant, there is also very high value in supporting the NeoBIG community to build comprehensive datasets including all the wealth of data collected in the NeoBIG trials, and to provide infrastructure enabling large scale collaboration and sharing. The advantage of building such consolidated data sets under a single authority in charge of their maintenance is that coherence, adherence to common methodologies, standards and ontologies, and availability can be ensured. While a solution maintaining all the data at the institutions generating that data is feasible and provides flexibility and scalability, it does not guarantee adherence to the same methodologies, common data models and standards, or long term maintenance, which makes the use of that data by large communities of users more difficult.

Due to the very strict requirements for a production-level system, with available documentation and user support, commercial deployment and long term maintenance, we have also concluded together with BIG that current ACGT prototype tools and services cannot be directly used for the NeoBIG project, however many of them should be part of a further targeted solution.

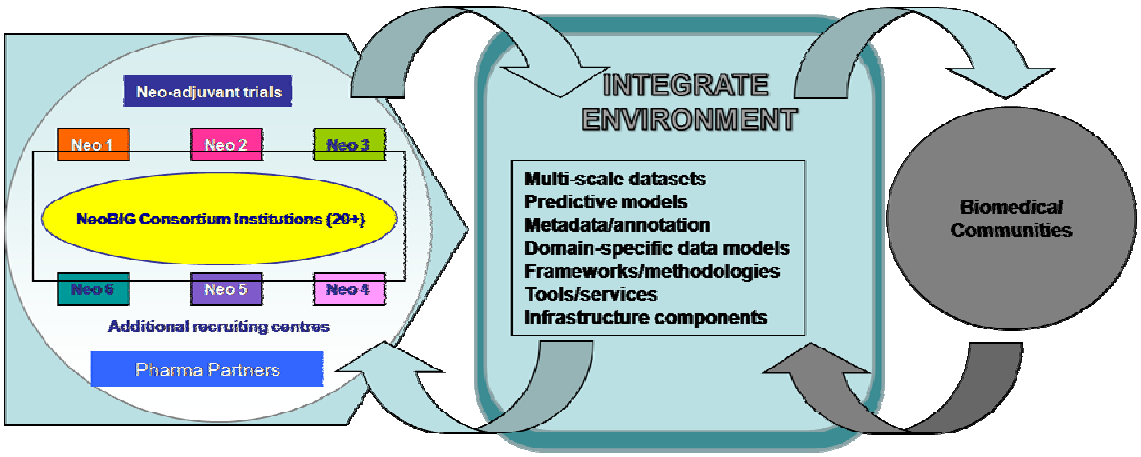

Figure 7: The INTEGRATE concept: Sharing and collaboration among clinical research and biomedical communities

In this context, we have defined INTEGRATE, a new collaborative project that aims to build solutions that support a large and multidisciplinary biomedical community ranging from basic, translational and clinical researchers to the pharmaceutical industry to collaborate, share data and knowledge, and build and share predictive models for response to therapies, with the end goal of improving patient outcome.

To address the needs identified during the NeoBIG requirements analysis, the INTEGRATE project will develop flexible infrastructure components and tools for data and knowledge sharing and wide scale collaboration in biomedical research. Our infrastructure will bring together heterogeneous multi-scale biomedical data generated through standard and novel technologies within post-genomic clinical trials and seamlessly link to existing research and clinical infrastructures, such as clinical trials management systems, eCRFs, and hospital EHRs, and to relevant external biomedical infrastructures. The fundamental conceptual difference to the ACGT approach is the need to build repositories of data, annotated models, and metadata and provide tools to extract and manage content, add and update data and models, and link to external sources for complex analyses. On top of this flexible infrastructure and using the available multi-level data, the project will focus on developing and validation of models and simulators predicting therapy sensitivity for individual patients.

Next to bringing together data and knowledge, our solutions will join a wide multi-disciplinary community of biomedical and clinical researchers committed to work together, to establish common methodologies and clinical protocols, to collaboratively build predictive models, carry out research and select the most suitable integrative workflows. The infrastructure and tools developed by the INTEGRATE project will support BIG to promote in the clinical community new methodologies and define standards concerning the collection, processing, annotation and sharing of data in clinical research and improve the reproducibility of results of clinical trials.

INTEGRATE aims to build an environment providing to its users full support for collaboration and sharing of complex multi-level datasets and models, but also access to relevant external data, knowledge and services. At the same time, we aim to enable the biomedical research community to benefit of the comprehensive datasets preserved by the INTEGRATE environment, and of our predictive models and tools. We are aware that the value of our tools and infrastructure would be only limited without sufficient data. To this end, an important goal of the project will be to enable long term sustainability of the project solutions, with specific focus on long term maintenance of large datasets built with common methodologies and using standardized models and terminologies.

According to the plan of work of the INTEGRATE project, ACGT tools, services but also the legal and ethical framework needs to be further developed in a way that the results of ACGT but also of a related VPH project called ContraCancrum can be exploited and used as productive tools and services in concrete clinical scenarios and trials.

11. Conclusions and Perspectives

A number of important lessons have been collected in the context of the ACGT TOP trial. These relate to both scientific issues, procedural issues as well as usability of the ACGT tools and platform related issues.

Specifically,

a) Although processing individual data is complex (for example, uncovering functional DNA variation in multiple cancer samples using whole-genome sequencing), the true challenge is in integrating the multiple sources of data. Mining such large high-dimensional data sets poses several hurdles for storage and analysis. Among the most pressing challenges are: data transfer, access control and management; standardization of data formats; and accurate modelling of biological systems by integrating data from multiple dimensions.

b) It is important to efficiently move big data sets around the internet. In ACGT we approached the problem based on the notion “leave data where it is produced” and enable access to it on a “need to be basis”. This approach is technically achievable, but creates a number of “administrative tasks” for each and every research center. This is inefficient and presents a barrier for data exchange between groups.

c) Standardizing data formats. Different centres generate data in different formats, and some analysis tools require data to be in particular formats or require different types of data to be linked together. Thus, time is wasted reformatting and re-integrating data multiple times during a single analysis. There seem to exist significant advantages in building consolidated data sets under a single authority in charge of their maintenance, so that coherence, adherence to common methodologies, standards and ontologies, and availability can be ensured. While a solution maintaining all the data at the institutions generating that data is feasible and provides flexibility and scalability, it does not guarantee adherence to the same methodologies, common data models and standards, or long term maintenance, which makes the use of that data by large communities of users more difficult.

d) One solution to these problems is to house the data sets centrally (which would require data format standardisation too) and bring the high-performance computing (HPC) to the data. Although this is an attractive solution, it also presents access control challenges, as groups generating the data may want to retain control over who can access the data before they are published. This is the conceptual, scientific and technical approach adopted in the context of the NeoBIG studies to be pursued by the INTEGRATE project.

e) Finally, due to the very strict requirements for a production-level system, with available documentation and user support, commercial deployment and long term maintenance, we have also concluded that current ACGT prototype tools and services cannot be directly used for production related clinical research. However many of them should be part of a further targeted solution, under the assumption that a working model for their maintenance and long term support is defined.