FONDAZIONE POLICLINICO UNIVERSITARIO AGOSTINO GEMELLI

UNIVERSITA' CATTOLICA DEL SACRO CUORE

RADIOTERAPIA ONCOLOGICA – GEMELLI ART

ULISSE: Umbrella protocoL ISSue for oncological patiEnts

Study co-coordinator:

Vincenzo Valentini MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Writing committee:

Vincenzo Valentini, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Andre Dekker, PhD. Department of Radiation Oncology (MAASTRO), Maastricht University Medical Center Maastricht (MUCM+), The Netherlands.

Elisa Meldolesi, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Andrea Damiani, PhD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Co-investigators:

Elisa Meldolesi, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Johan van Soest, PhD. Department of Radiation Oncology (MAASTRO), Maastricht University Medical Center Maastricht (MUCM+), The Netherlands.

Roberto Gatta, PhD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Nicola Dinapoli, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Maria Antonietta Gambacorta, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Maura Campitelli, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Anna Rita Alitto, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Giovanna Mantini, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Mario Balducci, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Daniela Smaniotto, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Stefano Luzi, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Giovanni Palazzoni, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Gian Carlo Mattiucci, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Francesco Miccichè, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Stefania Manfrida, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Francesco Cellini, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Vincenzo Frascino, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Barbara Corvari, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Adele Petrone, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Fabio Marazzi, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Mariangela Massaccesi, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Luca Tagliaferri, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Silvia Chiesa, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Rosa Autorino, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Valentina Chiloiro, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Luca Boldrini, MD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy.

Giuseppe Ferdinando Colloca, MD. Department of Radiation Gerontology, Neuroscience and Orthopedics, Catholic University of Sacred Heart, Rome, Italy.

Other expertise:

Vito Lanzotti – software programmer manager
Andrea Damiani – MatheMATICS
Pierre Marie Gori – Image mining (Astrophysicist)
Carlotta Masciocchi, PhD. Department of Radiation Oncology – Gemelli ART, Catholic University of Sacred Heart, Rome, Italy

Table of Contents

1. Summary

Aim of the study
The primary and general objective of the ULISSE Umbrella Protocol for oncological patients is to facilitate the development and validation of multi-factorial prediction models for different treatment outcomes. The long term aim is to build a Decision Support System (DSS) based on validated prediction models in order to be able to personalize treatments in terms of both treatment efficacy and toxicity control. The DSS has also the objective to identify patients to be included in future randomized clinical studies stratifying the different risk classes depending on the outcomes each times identified.

Hypothesis
Our general hypothesis is that we will improve the performance of the prediction models for survival and toxicity if we develop multifactorial models. The basic models will be based on patient related variables (e.g. age, sex), clinical presentations of the disease (e.g. staging, markers, imaging data), treatment data (e.g. chemotherapy, radiotherapy, surgery information, palliative care) and imaging data (diagnostic, treatment or follow-up images). The improved multifactorial models will include additional clinical and treatment imaging and/or genetic information even though no biological data will be actively collected in this project.

Study Design
This is a retrospective and prospective cohort study.

Inclusion criteria
All patients arriving at the participating Centers for oncological treatment, will be eligible for the inclusion in the ULISSE study. For the retrospective part of the study, patient data have already been stored in a local electronic database at each center. The data will be anonymized at the local treatment site and only be shared for research purposes. The patients enrolled into the prospective part of the study will be informed at the first visit about the standardized data collection by the treating physicians. Patient's written informed consent will be collected and archived.

Objectives
Development and validation of multi-factorial prediction models for different treatment outcomes. Based on the validated prediction models, the long term objective is to build a DSS that will be finally presented to the end-user in a variety of ways such as nomograms [1] or via interactive websites to easily calculate outcome predictions.

2. General introduction

2.1 ULISSE framework

- The development and validation of multi-factorial prediction models requires the availability of a large amount of data patient considered significant for present and futures studies.
- Each variable has to be included into a terminological system. Adding more variables in the future is possible, but starting early with the most important variables is fundamental.
- Collected data has to be reusable both in time (e.g. in the future) and in the space(across different institutions or research groups); this is possible only if everything about the data is correctly specified (e.g. denomination, measurement units, measurement modality)

- Reusability of legacy data is possible, on condition that suitable semantic remapping functions from old to new data are provided.
- Appropriate mathematical and statistical methods are needed in order to learn from a large collection of data (Large Database) and will help to suggest new modeling hypotheses to be tested.
- The Patients privacy protection has to be protected. This can be accomplished in two ways:
    - by anonymizing data before they leave the collecting institutions walls, making sure that no inverse remapping is available ("cloud" solution)

  or
    - by exploiting the so called "Distributed Learning" solution, in which no data ever leaves the collecting institution, but a regressive or classifying predictive model can be acquired exactly as if all data had been collected in the same place.

2.2 Individualized treatment and prediction of outcome

Over the past decade, remarkable advances in cancer care with the adoption of the newest diagnostic and treatment technologies has created new challenges [2]. Progress in computer technology with new diagnostic methods and treatment modality developments is responsible for advances in radiation oncology with radiotherapy planning and evolution of delivery facilities evolution. However, although the progress in computer technology has had an important influence in radiotherapy planning and delivery facility evolution allowing for remarkable precision in treatment delivery and better outcome, the dose escalation process can increase the severity and duration of side-effects [3]. While some patients may fail to complete their treatment, others will need medication or hospitalization and sometimes these side-effects will lead to late toxicity, which will negatively influence quality of life and well-being.

Long considered to be a physical intervention, radiation therapy is now more accurately conceptualized as a biological intervention with effects at the cellular and molecular level, modulated through cellular signaling pathways and the immunological axis [4,5]. Accordingly, combinations of radiation therapy with targeted biological agents have been proven to show increasing efficacy and hold promise for future advances [6,7]. Therefore, new, less toxic anti-cancer therapies are being developed. They include new approaches targeting cancer-specific pathways in the cell and intending to improve the treatment outcome in terms of survival as well as toxicity [8,9].

The use and role of medical imaging technologies in clinical oncology has also greatly expanded during the last decade from a primarily diagnostic, qualitative, tool to acquiring a central role in the context of individualized medicine with a quantitative value. Several studies have been developed to analyze and quantify different imaging features (e.g. descriptors of intensity distribution, spatial relationships between the various intensity levels, texture heterogeneity patterns, descriptors of shape etc.) and the relationships of the tumour with the surrounding tissues to identify a possible their relationship with treatment outcomes or gene expressions [10,11].

Therefore, as these new strategies and therapies are being tested, it becomes more and more apparent that certain subgroups of patients may benefit from a specific treatment, while others don’t or may even have a worse outcome [12]. The same is observed for the toxicity of the treatment. Some patients suffer from severe side-effects while others are relatively unaffected [13]. This means that there is a complex interplay of different factors

which has not yet been unraveled yet. These differences between individual patients are not only observed in case of treatment with medication or chemotherapy, but they also occur both during radiotherapy treatment - implying that the decision to escalate the radiation dose should be individualized - and surgery, modulating the extension of the surgical intervention.
Many publications have shown that the dose distribution can easily fitted to complex anatomical shapes enabling dose distribution optimization; however no one has actually shown consistent outcomes in terms of tumor control or organ at risk owing to both the small series of patients and the lack of homogeneity in the data collection used in such clinical trials. Hence, the necessity to create large databases, realized by crossing and combining multiple data already recorded in specific storage archives, to provide sufficient statistical power to act as acceptable decision supporting tools.
The amount of available information to explain these observations is expanding enormously owing to new diagnostic tools such as genomic and proteomic profiling (e.g. based on the patient's blood or saliva), and anatomical and functional imaging techniques (e.g. CT, MRI, PET).
This knowledge will enable the prediction of the outcome for a certain patient in combination with a specific treatment with more accuracy. It will lead to better identification of risk groups, which results in stage migration trying to find new treatment options or other combinations of treatment options for these subgroups. It can be expected that treatment will be more personalized, which will not only save patients from unnecessary toxicity and inconvenience, but will also facilitate the choice of the most appropriate treatment. Currently, this choice is based on general guidelines that only take into account a low number of variables. These guidelines are developed for groups of patients and therefore can lead to over-treatment in some patients and inadequate therapy in others, resulting in major expense for individuals and society.
However, prediction of outcome in order to choose the optimal treatment is complicated in view of the very complex, dynamic nature of cancer and organs at risk. In a systematic review it was concluded that physicians' predictions of survival of terminally ill cancer patients tended to be incorrect in the optimistic direction [14]. This is in agreement with a study, investigating the accuracy of radiation oncologists in predicting survival [15]. Studies, investigating the performance of physicians in predicting side-effects of radiotherapy treatment, are currently lacking. However, the ability of humans, and thus physicians, to assess the risks and benefits associated with a specific combination of patient, tumor and treatment characteristics, that will ultimately include many thousands of parameters, is limited. Therefore, treatment can only become more personalized if accurate, scientifically based decision aids are developed, that can offer assistance in clinical decision-making in daily practice.

2.3 Population-based research

To date, the standard efforts in the medical field and inherently also in oncology are to consider the outcomes of randomized clinical trials (RCTs) as having the key role in the definition of clinical guidelines, protocols, and research. However, patients participating in RCTs only represent a selective subgroup of the general population, resulting in an inherent limiting factor when interpreting results, as the characteristics of a population seen in routine clinical practice is very different compared to the population included in RTCs [16]. Furthermore, some patient groups are under-represented in RCTs, including the elderly, those with comorbidities [17,18], and patients from under-represented ethnic and socioeconomic backgrounds [19-21]. Thus,, small benefits observed in highly selected trial patients are likely to disappear when the same treatments are applied in routine practice.
Beside RCTs, population-based observational studies are progressively emerging as a complementary form of research, especially to ensure that the results of clinical trials

translate into tangible benefits in the general population [22]. Observational studies are essential to identify whether practice has changed appropriately, to document the harms of therapy in a wider population, in patients of different age and with different comorbidities, and to determine whether patients in routine clinical practice are reaching the expected outcomes [23-25] with the expected toxicity.
Models for any outcome could benefit from extra information. Therefore, using the data from many patients will facilitate model building also for toxicity [26]. As physicians try to avoid severe side-effects as much as possible the number of events is generally low, making it barely possible to develop accurate models for these side-effects.
At the moment, models are usually based on a restricted number of variables, often limited to one type of information. Some models use genetic information only, others are solely based solely on clinical factors. Different types of variables could offer complementary information and thus improve the performance of models [27,28]. Furthermore, the usage of a distributed learning approach allows a model to learn from all these data without the need for data to leave the individual hospital, achieving a high quality research level.

2.4 Rationale for implementation of Standardized Data Collection (SDC) in cancer

SDC will improve the quality of the data by defining which variables should be collected preferentially and how these variables should be measured. Variables will be collected and organized into an Ontology, according three different tiers: Registry level, Procedures level and Research level [29]. The prospective collection of patient, tumor and treatment characteristics will facilitate the development of prediction models for survival as well as toxicity outcomes among participating Centers. In addition, data on survival and toxicity can be used to compare the results of new and emerging radiation delivery techniques, targeted therapies or chemotherapy regimens after they have been clinically introduced to the results obtained with the standard treatment.

3. Objectives of ULISSE

3.1 General objective

The primary and general objective of ULISSE Umbrella Protocol for oncological patients is to facilitate the development and validation of multi-factorial prediction models for different treatment outcomes. The long term aim is to build a Decision Support System (DSS) based on validated prediction models in order to be able to personalize treatments in terms of both treatment efficacy and toxicity control. A DSS also has the objective of identifying patients to be included in future randomized clinical studies through the stratification of the risk's classes depending on the outcomes each times identified.

3.2 Specific objectives

- To develop, validate, and improve prediction models for overall survival, local control, disease-free survival, and metastasis-free survival;
- To develop, validate, and improve prediction models for acute and late radiation-induced side effects relevant for cancer patients;
- To use the prediction models to better inform patients about the risks (acute and late toxicity) and benefits (overall survival) of the treatment;
- To use the outcome of the prediction models to individualize the treatment;
- To use the outcome of the prediction models for the development and investigation of the potential benefit of new and emerging radiation delivery techniques or other treatment options;

- To compare the outcome of new treatment options that are clinically introduced with the current standard in terms of radiation-induced toxicity, patient-rated symptoms and quality of life and overall survival.
- To develop, validate, and improve prediction models about QoL in the involved population

3.3 Inclusion criteria

All patients arriving at the participating Centers for cancer treatment, will be eligible for the inclusion in the SDC.

4. SDC data

4.1 SDC features

Minimal requirements of each Center to participate in the SDC exercise are:

- To provide an Electronic Medical System (EMS) for cancer to record patient's information.
- To be able to 'translate' local data into the ontology based archives
- To be able to anonymize local data
- To use technology able to developed advanced multicenter researches
- To provide patient written informed consent according with local National legislation.

4.2 SDC general

The SDC includes two different steps:

A. Retrospective analysis of baseline characteristics, treatment-related factors (including dose distribution parameters, acute and late radiation-induced toxicity, local control, disease-free survival, overall survival).
B. Prospective assessment of baseline characteristics, treatment-related factors, (including dose distribution parameters, acute and late radiation-induced toxicity, local control, disease-free survival, overall survival and health-related quality of life).

In the following paragraphs, the assessments will be described in more detail.
It has to be highlighted that the investigator will be responsible for inclusion of patients and day-to-day management of the patient treatment according to local policies and the patient's need and will monitor the progress of the SDC in an ethical and scientific manner. A web based Electronic CRF will be used. In each participating centre a data manager will be responsible for the data collection. Patients will be included in the SDC by the treating physician.

4.2.1 Baseline characteristics (Registry Tier)

The baseline patient and tumor characteristics that are considered relevant are outlined and organized into the Registry level, the first and most general level that includes the minimal information (age, gender, ethnicity etc), used for epidemiological analysis only.

4.2.2 Treatment-related characteristics (Procedure Tier)

The baseline treatment and radiotherapy characteristics that are considered relevant are also defined. These variables are organized into the Procedures level that includes treatment information with related toxicities and the evaluation of outcome in terms of achievement of patient goal' as well as DFS and acute and late toxicities. Additional information on radiotherapy will be extracted in an automated way from the record and verified system. More detailed information regarding dosimetric parameters can be

calculated using the 3D dose matrix and the imaging information. This information will be retrieved from the PACS (Picture archiving and communication system) system, also in an automated way. This will not be any burden to data managers, treating physicians or patients.

4.2.2.1 Acute and late toxicity characteristics

Acute and late toxicity will be scored according to the RTOG scale (for the retrospective analysis) and to the CTCAE v3.0 or CTCAE v4.0 (for the prospective analysis).

4.2.2.2 Patient-rated quality of life

Quality of life will be measured using the EuroQol-5D-5L, EORTC QLQ-C30 and EORTC QLQ specific for each cancer type. EuroQol-5D-5L is a small, standardized generic quality-of-life questionnaire consisting of two parts. The first part is a 5-dimensional questionnaire (5 questions), the EQ-5D. The five dimensions are mobility, self-care, usual activities, pain/discomfort, and anxiety/depression [30]. With regard to each of those dimensions, the patient is asked to indicate if he or she experiences no problems, some problems, or major problems. The resulting profile of answers (one of 243 possibilities) can be transformed to a value given by the general population: the EQ-5D index [31]. The second part of the EuroQoL questionnaire is a visual analogue scale, the EQVAS, which represents the patient's judgment of his own health state. The advantage of the EuroQoL-questionnaire is its ability to provide utility scores expressing the health state of patients, which can be used to calculate Quality Adjusted Life Years (QALYs). QALYs combine the number of life years gained and the quality of life during these years in one single measure.
All scales and single-items measures in both questionnaires are linearly transformed to give a score from 0 to 100 according to the algorithm recommended by the developers. A high score for a functional scale represents a high level of functioning, a high score for the global health status represents a high QoL, and a high score for a symptom scale represents a high level of symptomatology or problems. These evaluations could be effected by paper questionnaires or mobile applications technology.

4.2.3. Imaging (Research Tier)

Diagnostic, treatment and follow-up imaging information will be retrieved from the PACS system in an automated way and organized in the third and most detailed level, the research level, to be used for advanced research projects. The use and role of medical imaging technologies in clinical oncology has moved from a primarily diagnostic, qualitative, tool to occupying a central role in the context of individualized medicine with a quantitative value. Several studies, such as radiomics [10,11], have been developed to analyze and quantify different imaging features (e.g. descriptors of intensity distribution, spatial relationships between the various intensity levels, texture heterogeneity patterns, descriptors of shape etc.) and the relationships of the tumour with the surrounding tissues to identify a possible their relationship with treatment outcomes or gene expressions.

4.2.3.1 Biological data characteristics

No biological data will be collected in this project. Analysis of biological data will be realized only using information properly collected in previous Ethical Committee approved clinical trials where a valid informed consent has been signed by the patient.

5. ULISSE strategies to implement prediction models for cancer

The availability of multiple clinical data, together with improved imaging modalities, leads to unprecedented amounts of medical and biological data, which can only be dealt with using

computational methods, not only for storing data, but also for integrating, analyzing, displaying and eventually understanding it. Beside traditional statistical tools (e.g. linear models, generalized linear models, survival models), machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from the data, machine learning allows computers to find hidden insights without being explicitly programmed where to look. This approach is offered by a number of different techniques for these purposes, mainly Bayesian networks [32,33], Support Vector Machines [34] or Cox regression [1]. These techniques can overcome problems encountered with conventional statistical methods especially if data is highly correlated, many variables are available but a limited number of patients (high-dimensional data), or many different models have to be tested for their predictive value. In the field of radiotherapy and especially for the prediction of treatment responses, machine learning is an upcoming modality. Successes over traditional statistics have already been published [35] and the first promising results for building predictive models concerning survival of cancer can already be found in the literature [34]

5.1 Main ULISSE strategies

To accomplish the challenge of collecting a large amount of data, two different strategies will be used dependent on the research's purpose and Centers' agreement. A centralized data record consolidation approach requires a conversion of the data archives according to a global data dictionary and then, the anonymous reproduction of the clinical data into a cloud-based large database. Distributed learning is a very flexible approach that allows the system to learn from the data without the need for data to leave the individual hospital. In the following paragraphs, these two approaches will be described in more detail.

5.2 Centralized consolidation of data records approach (BOA CLOUD)

Università Cattolica del Sacro Cuore (UCSC) in Rome developed a software called "BOA" (Beyond Ontology Awareness), which is part of an EMR system the company itself has developed for Radiation Oncology wards (Annex 1). The IT architecture of BOA (System for Patient Individual Data Entry and Recording Beyond Ontology Awareness) converts legacy pathology archives of a Center according to a global data dictionary and replicates anonymously just the clinical data into a cloud-based large database (fig.1). The Global Data Dictionary is designed to be compatible with the standard CDISC Operational Data Model to exchange data in a common format.
The cloud-based large database is the only asset that is shared among the participating centers; this sharing is only temporary, research-bound and lasting through the life span of a particular study. The system guarantees that nothing, except anonymous and non-referable clinical data – with no link to the original local archives – will become part the large database.
Furthermore, to investigate a predictive model by using information from one or more institutes, the data will be run through statistical algorithms, in a process which exchanges only aggregated data but no individual records between the participating institutes nor gives external access to individual records or to multiple records regarding the same physical individual of a participating centre. The Supervisor Center can directly query the shared large database only, complying with the requests of research investigators and giving back results accordingly to the policies of the participating centers. At the end of the study life span, the cloud database is deleted.
Each center can make local queries on its own pathology database, much like the way the Research Supervisor can run queries on the cloud large database and compute outcomes for each participating center to use. The Research Proxy is designed to give back only anonymous data when a minimum threshold of cases is reached. Furthermore, it never provides the patient's anonymized ID. Data subsets produced by the Research Proxy will be

used to build and validate investigation models, incorporating in them the open source R statistical software.
In a second stage, the system will evolve towards a Distributed Learning approach: the Supervisor's Learning Analyzer Proxy will send algorithms directly to Local Research Proxies, taking back from them only the results of each iteration step, with no need to work with shared data in the Cloud anymore.
Presently, due to reasons strictly related to the algorithms, some of the most common predictive models used in statistics and machine learning cannot be run under a Distributed learning framework, thus making the cloud-based shared large database the only feasible solution for research.

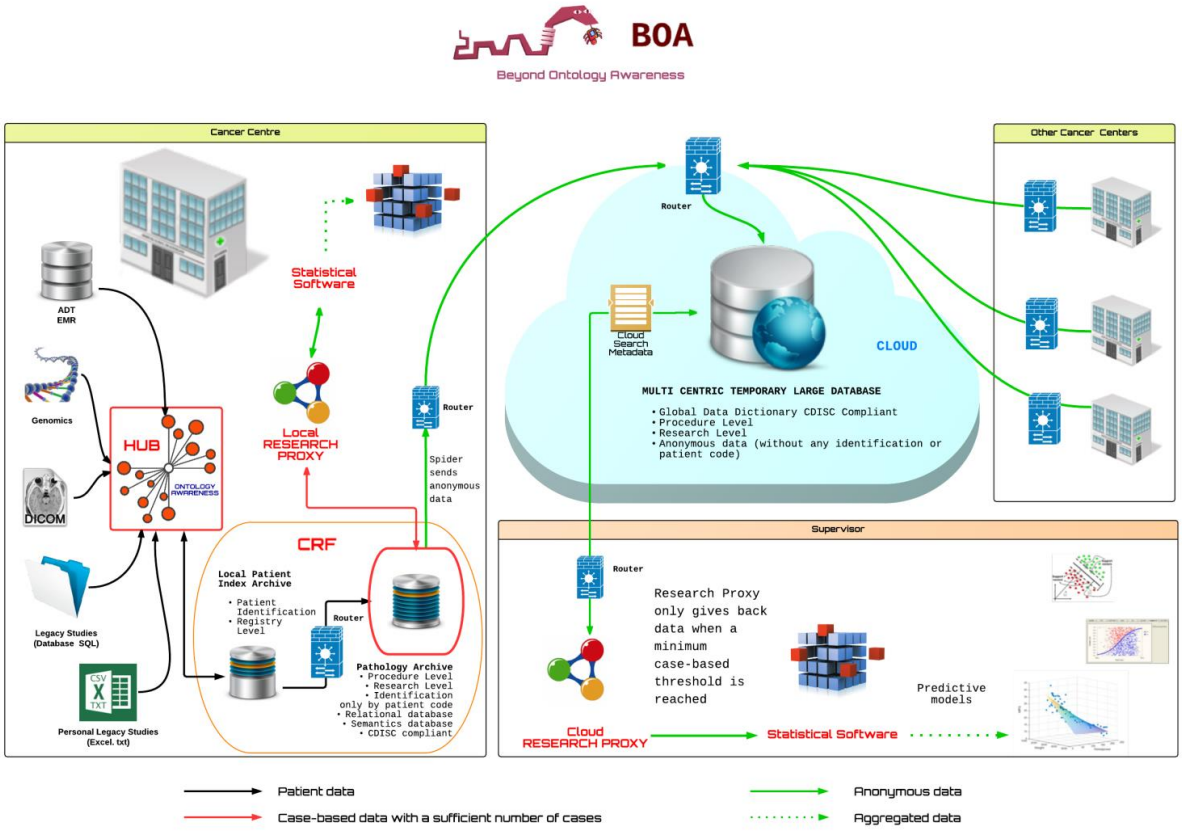

Fig. 1 BOA CLOUD

5.3 Distributed learning approach

The aim of distributed learning is to learn a model from the data without the need for data to leave the individual hospital. A distributed machine learning algorithm is split up into two parts (fig.2):

1. One master application which is installed on a central server (called proxy) and coordinates the learning between the hospitals.

2. The second part is a local learning application which is installed at each hospital. It has access to the local data and performs learning tasks but does not share patient data with the outside world.

The local application learns a model from local data. This local model is sent to the proxy where it is compared with the models from the other hospitals. A consensus model is generated and sent back to the hospitals for refinement. After preset convergence criteria are met, a final consensus model is created. This method works for a variety of models as described in literature [36].

The information exchanged between proxy and local nodes is limited to aggregated values (e.g. parameter weights, general statistics, coefficients) and contains no patient data. All traffic between proxy and local nodes is managed, monitored and audited by the infrastructure. An entire learning run is an iterative process that usually requires many cycles (~500) until the master application determines that the learning process has been completed.

In Distributed Learning mode, Local Research Proxies do not move data around: they only apply iterative algorithms that the Supervisor will use to build a consensus and estimate the model's parameters.

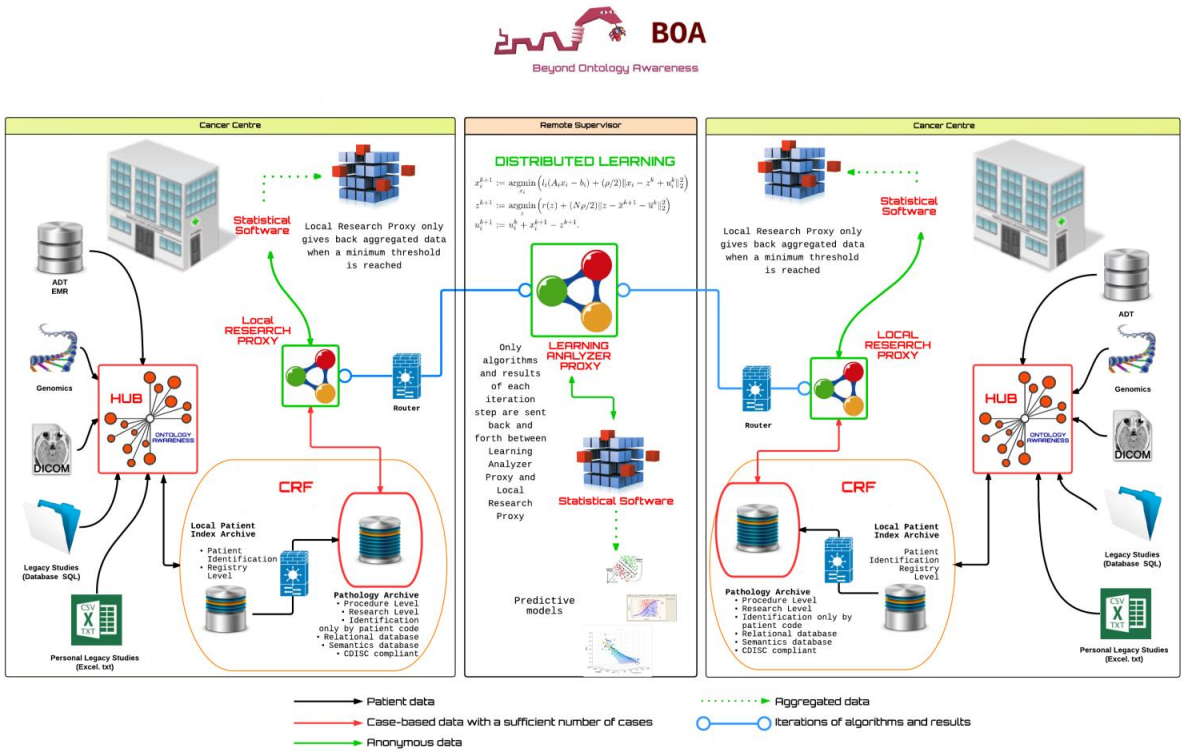

Fig.2 Distributed learning process.

5.4 Semantic Web technology

The transition from a list of variables identified as relevant for cancer to an ontology, requires an increase of both the complexity and the formalism of the language with the usage of more complex constructs to represent relationships between variables thus enriching their knowledge contents. Semantic Web technology is the model used to represent data distribution. For the Semantic Web technology, data is represented by triplets (subject, predicate, object) using the Resource Description Framework (RDF)

language [37]. The interaction between elements of multiple triplets is defined inside an ontology through a different language (RDFS or OWL) allowing informatics system to automatically generate inferences from any exploitable data source. Software agents can easily parse and make inferences on large data repositories applying formal-ontologies on explicitly declared facts to infer the entire set of facts logically inferable.
Semantic Web technology is integrated in our Institution with all the storage data in order to be able to exploit the data source and automatically generate inferences from them.
The power of the semantic web is the extremely simple, however flexible RDF representation (one table with three columns), as well as the federated nature of the web where both data and knowledge can reside at multiple locations on the internet and can be queried using SPARQL, the query language of the Semantic Web [38].

6. ULISSE Statistical analysis

6.1 Data analysis features

Prediction models will be built using two large families of data analysis tool:

1. Inferential regression analysis tools, mainly based on the relationship between outcomes (binary, continuous or multinomial) and covariates, or elements in the dataset, that establish a data-to-outcome one-way link, investigated using traditional statistical tools such as linear models, generalized linear models, survival models etc;
2. Machine learning analysis tools, used to create a recursive relationship between outcomes and generating data, with a complex automation background, that can resolve complex relationships between elements in the dataset and final results, too complex, in some situations, to be investigated by using the tools of the family 1.

Each model, however defined, must undergo to a strict evaluation process mainly based on internal and external validation [39] in order to become a reliable tool to be used in clinical contexts.
The methodological process to learn, i.e. to go from data to useful decision support as follows; experts determine which features should be included in the learning process; in the pre-processing step data quality is improved by imputation for missing data and outliers and bias detection and correction. Then the data is split into a training and a validation cohort. The training cohort is used in a feature selection and classification algorithm to train a model.
The machine learning approaches can vary but are typically Bayesian networks [32,33], Support Vector Machines [34] or Cox regression [1]. The final model can be presented to the end-user in a variety of ways such as nomograms [1] or via interactive websites.
The performance of the models will be assessed in terms of discrimination as well as calibration. External validation cohorts will be used for this purpose. Discrimination will be assessed using the c-statistic or Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC). The c-statistic is comparable to the AUC for dichotomous outcomes, but can also be used for Cox regression analysis. A graphical assessment of calibration will be performed by plotting the expected versus the observed outcome. In addition, the Hosmer-Lemeshow test will be used. The clinical value of the models will be assessed using decision curve analysis [40,41]. This will make it possible to compare the clinical value of different models over a number of decision thresholds (or cut-off points for probability of outcome). Using this method, there is no necessity to choose an a priori cut-off point (for a clinical decision).

6.2 Missing data

The system will take advantage of a missing data tool[1] with the basic requirement to detect, report and impute missing data. For the detection a solution based on shape expressions [42] will be further developed. In this approach a data shape can be intuitively defined including taking full advantage of ontologies (e.g. if a patient shall has a T-stage specified, all T stages (1, 1a, 1b, 2 etc.) will be accepted automatically). This data shape can be converted into a SPARQL query that can query a local RDF store for all patients that fit the data shape. It can also query for the inverse (i.e. all patients that do not fit a data shape) and which data elements (e.g. T-stage is missing) are the problem. The results of these data queries are the basis of the detection and reporting of missing data. The imputation of missing data is a research field in itself. A literature review will be performed to implement all common imputation methods (e.g. the investigators have used median, mean and Bayesian imputation in prior work) available and configurable in the context of a specific research question. This will allow a learning algorithm to call the tool with a data shape and imputation configuration (method and parameters). The imputed data elements will be stored in a separate graph in the local RDF store including their provenance so that the imputed elements can be separated from the asserted data elements. Interfacing between the tool and the calling learning application will be defined during the project based on user requirements.

6.3 Control of data consistency

The Data Manager will perform computerized and manual consistency checks by ad hoc retrieval services. A 'Continuous Data Quality Assurance' process will be able to identify inconsistencies inside the data collection in three different ways:

- Identifying impossible data (e.g. patient's weight=500Kg)
- Identifying conflicting combinations (e.g. stage I and presence of metastasis)
- Through the usage of Bayesian network analysis with data shaping

Queries will be issued in case of inconsistencies. Consistent forms will be validated by the Data Manager. Inconsistent data will be kept "on hold" until resolution of the inconsistencies.

7. ULISSE objectives' representation

DSS will be presented to the end-user in a variety of ways. Graphical calculating devices such as nomograms [1,43] are one of the most common forms of predictive device, besides the even more appealing interactive website. Furthermore, in this era of technological progress, the possibility to create specific applications for devices of new generation is also very interesting (e.g. cell-phones, tablet etc).

8. Ethical considerations

Accruement will be conducted according to the principles of the Declaration of Helsinki (version of 2004) and in accordance with the Medical Research Involving Human Subjects Act (WMO) and consequent guidelines, regulations and Acts.
The local Ethics Committees (EC) of the participating centres shall approve the protocol before patient accrual phase starts, according to legislation of each country.

[1] Note that when a 'tool' is mentioned, no assumption is made on how that tool will be deployed. E.g. as a service, or stand-alone application or embedded inside a learning algorithm, etc…

Patient accrual will be conducted by each Center, selecting those who meet the inclusion criteria. The physician will explain the aim of the study to the patients and questionnaires they have to complete. Written informed consent for anonymized treatment data collection and approval of related research will be collected from each patient according to local practice and following the rules listed in the present document (See annex 2). Patients will be given at least three days before actual data collection begins, to decide whether to agree or not.
Each Center involved in the research is the only party responsible for data collection and it guarantees that all related procedures are carried out according to the present document and that each patient has signed a suitable informed consent. Participating Centers can share the data within the large-database only after approval of this protocol by the respective EC. Each Center indemnifies, the Catholic University of Rome, the data manager and the individuals who will analyze the data and holds them harmless from any and all claims and exempts them from all responsibilities regarding the collection of informed consent.
Obtaining an informed consent is mandatory for accrual of prospective patient. For the analysis of patients accrued retrospectively before prospective patient accrual starts, such informed consent is not required (See appendix 1).

9. ULISSE Management

9.1 Privacy protection of patients

In case of a distributed learning approach data does not need to leave the institute in the process of distributed machine learning. This is possible as the central ”master” sends possible prediction models rather than fetching the data from remote nodes. Only statistical indexes totally unrelated to specific patients are exchanged between nodes and their master [36,44,45].
In a centralized data record consolidation approach, the patient’s privacy will be protected at the architectural level because all data transfer will happen through a fully encrypted pipeline, and data records will be anonymized before leaving the local center’s walls. The mapping between data records and individuals will also be protected via software procedures and will never leave the originating center, thus rendering virtually useless any attempt at tampering with data transmission and even accessing the actual data records. This already high degree of protection will be raised even further, where appropriate, by the adoption of secure communication channels (e.g.: virtual private networks over secured connections) and, should the necessity arise in order to comply with local regulations or specific policies at the centers’ level, decentralized data processing and/or data obfuscation will be added as a further layer of security.

9.2 Data Privacy Strategy

The risk of privacy infringement for participating institutes will not be increased by the project. All data, originating from local repositories at the institute’s site, will be routed to a local repository in a totally anonymous form, because all names and links to traceable information will be removed before entering the repository, while unique identifiers that could point directly or indirectly to individual patients will be remapped to a different code.
As a consequence, the local endpoint, which is queried by other research group member’s from outside the institute during normal activity, will not expose any method to reconcile clinical information to the relevant patient. From this point of view, our data will have a higher degree of protection than the same data stored in the institute’s databases.
It must also be noted that in investigating a predictive model by using information from one or more institutes, the data will be run through statistical algorithms, in a process which

exchanges only aggregated data but no individual records between the participating institutes nor gives external access to individual records or to multiple records regarding the same physical individual of a participating centre. The inherent statistical nature of this activity adds yet another level of protection to the system.
Model validation activities will happen locally, and will result in statistical indices totally unrelated to specific patients.
When endpoints are added locally, in order to enable data mining activities in a Semantic Web approach, all data will be anonymized and, when information is collected horizontally across different institutes, further levels of isolation will be achieved through data sectioning (e.g.: stripping away the geographical localization of the patient and contributing institute) as considered necessary.
Ethical committees ask for patients' privacy to be secured not only from an architectural point of view, i.e. through cyphered data exchange, but also as the information content itself: data have to be completely anonymized in order to allow them to safely leave the Hospital and feed multicentre research archives. It is not enough, though, to exclude registry-level information such as first and last name, addresses, telephone numbers and the like; it is an agreed upon requirement that no backwards path can be followed to reconstruct the clinical records from which anonymized data came from. Data sharing among different Centres is thus fostered by a system that extracts and harmonizes legacy data while making them available under these strict anonymity constraints.

9.3 Patient Privacy Data Mining (PPDM)

In both "Cloud" and "Distributed" pathways, patient Privacy is enforced at the architectural level:

- in a Distributed Learning approach, data never leave the hospital walls
- in a centralized consolidation of data records, any reference to personal information is physically detached from each individual data record before it leaves the hospital walls. Each record is assigned a new, randomly generated ID whose only purpose is to join records belonging to the same patient once the data records are collected at the master application level. The remapping function is deleted before the dataset leaves the originating hospital, hence the relation is broken and no backward mapping (i.e.: from data to physical patient) is possible.

It should also be noted that the system, already at the local node level, does not allow the execution of a selection query which only recovers a small number of records (under a predefined, hard-coded security threshold), because a malicious combined query approach could be exploited in order to pinpoint a specific patient through the partial knowledge of a subset of features that uniquely identify him/her.

10. Publication policy

All information resulting from this study is considered to be confidential. Study coordinators and a statistician will complete a data check before data can be analysed.
Any publication, abstract or presentation comprising results from the study must be submitted for examination and approval to the study coordinators. Publication policy is in accordance with CCMO regulations.
The first, the second and the last author of the publication will be chosen in accordance with the principal investigators. Other authors will be the investigators of the main recruiting centres listed in order of decreasing number of included patients. The responsible statistician of the trial will be always included in the author list.

References

1. Valentini V, van Stiphout RGPM, Lammering G, Gambacorta MA, Barba MC, Bebenek M, et al. Nomograms for predicting local recurrence, distant metastases, and overall survival for patients with locally advanced rectal cancer on the basis of European randomized clinical trials. J Clin Oncol [Internet]. 2011 Aug 10 [cited 2014 Mar 28];29(23):3163–72. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21747092

2. Lambin P, van Stiphout RGPM, Starmans MHW, Rios-Velazquez E, Nalbantov G, Aerts HJWL, et al. Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol [Internet]. Nature Publishing Group; 2013 Jan [cited 2014 Mar 26];10(1):27–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23165123

3. Bentzen SM, Trotti A. Evaluation of early and late toxicities in chemoradiation trials. J Clin Oncol [Internet]. 2007 Sep 10 [cited 2015 Jan 29];25(26):4096–103. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17827459

4. Bentzen SM. Preventing or reducing late side effects of radiation therapy: radiobiology meets molecular pathology. Nat Rev Cancer [Internet]. 2006 Sep [cited 2014 Mar 22];6(9):702–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16929324

5. Fowler JF. 21 Years of Biologically Effective Dose. Br J Radiol [Internet]. 2010 Jul [cited 2014 Mar 28];83(991):554–68. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3473681&tool=pmcentrez&rendertype=abstract

6. Glynne-Jones R, Hadaki M, Harrison M. The status of targeted agents in the setting of neoadjuvant radiation therapy in locally advanced rectal cancers. J Gastrointest Oncol [Internet]. 2013 Sep [cited 2014 Mar 28];4(3):264–84. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3712302&tool=pmcentrez&rendertype=abstract

7. Bonner J a, Harari PM, Giralt J, Cohen RB, Jones CU, Sur RK, et al. Radiotherapy plus cetuximab for locoregionally advanced head and neck cancer: 5-year survival data from a phase 3 randomised trial, and relation between cetuximab-induced rash and survival. Lancet Oncol [Internet]. 2010 Jan [cited 2014 Mar 21];11(1):21–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19897418

8. Bentzen SM, Harari PM, Bernier J. Exploitable mechanisms for combining drugs with radiation: concepts, achievements and future directions. Nat Clin Pract Oncol [Internet]. 2007 Mar [cited 2015 Jan 16];4(3):172–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17327857

9. Provencio M, Sánchez A, Garrido P, Valcárcel F. New molecular targeted therapies integrated with radiation therapy in lung cancer. Clin Lung Cancer [Internet]. 2010

Mar 1 [cited 2015 Jan 16];11(2):91–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20199974

10. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer [Internet]. 2012 Mar [cited 2014 Mar 26];48(4):441–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22257792

11. Kumar V, Gu Y, Basu S, Berglund A, Eschrich S a, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging [Internet]. Elsevier Inc.; 2012 Nov [cited 2014 Mar 28];30(9):1234–48. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3563280&tool=pmcentrez&rendertype=abstract

12. Mok TS, Wu Y, Thongprasert S, Yang C-H, Chu D-T, Saijo N, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med [Internet]. 2009 Sep 3 [cited 2015 Jan 18];361(10):947–57. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19692680

13. Bentzen SM1 HJ. Variability in the radiosensitivity of normal cells and tissues. Report from a workshop organised by the European Society for Therapeutic Radiology and Oncology in Edinburgh, UK, 19 September 1998. Int J Radiat Biol. 1999;75(4):513–7.

14. Chow E, Harth T, Hruby G, Finkelstein J, Wu J, Danjoux C. How accurate are physicians' clinical predictions of survival and the available prognostic tools in estimating survival times in terminally ill cancer patients? A systematic review. Clin Oncol (R Coll Radiol) [Internet]. 2001 Jan [cited 2015 Jan 29];13(3):209–18. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11527298

15. Chow E, Davis L, Panzarella T, Hayter C, Szumacher E, Loblaw A, et al. Accuracy of survival prediction by palliative radiation oncologists. Int J Radiat Oncol Biol Phys [Internet]. 2005 Mar 1 [cited 2015 Jan 29];61(3):870–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15708268

16. Zietman AL. Falsification, fabrication, and plagiarism: the unholy trinity of scientific writing. Int J Radiat Oncol Biol Phys [Internet]. Elsevier Inc.; 2013 Oct 1 [cited 2014 Mar 28];87(2):225–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23958142

17. Tyldesley S, Zhang-Salomons J, Groome P a, Zhou S, Schulze K, Paszat LF, et al. Association between age and the utilization of radiotherapy in Ontario. Int J Radiat Oncol Biol Phys [Internet]. 2000 May 1;47(2):469–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10802375

18. Faivre J, Lemmens VEPP, Quipourt V, Bouvier a M. Management and survival of colorectal cancer in the elderly in population-based studies. Eur J Cancer [Internet]. 2007 Oct [cited 2014 Jul 16];43(15):2279–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17904353

19. Bach PB1, Cramer LD, Warren JL BC. Racial differences in the treatment of early-stage lung cancer. N Engl J Med [Internet]. 1999;341(16):1198. Available from: http://www.nejm.org/doi/full/10.1056/NEJM199910143411606

20. Boyd C, Zhang-Salomons JY, Groome P a, Mackillop WJ. Associations between community income and cancer survival in Ontario, Canada, and the United States. J Clin Oncol [Internet]. 1999 Jul;17(7):2244–55. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10561282

21. Hershman D, McBride R, Jacobson JS, Lamerato L, Roberts K, Grann VR, et al. Racial disparities in treatment and survival among women with early-stage breast cancer. J Clin Oncol [Internet]. 2005 Sep 20 [cited 2014 Jul 16];23(27):6639–46. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16170171

22. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. Br J Cancer [Internet]. Nature Publishing Group; 2014 Feb 4 [cited 2014 May 28];110(3):551–5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3915111&tool=pmcentrez&rendertype=abstract

23. Pearcey R, Miao Q, Kong W, Zhang-Salomons J, Mackillop WJ. Impact of adoption of chemoradiotherapy on the outcome of cervical cancer in Ontario: results of a population-based cohort study. J Clin Oncol [Internet]. 2007 Jun 10 [cited 2014 Jul 16];25(17):2383–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17557951

24. Booth CM. Evaluating patient-centered outcomes in the randomized controlled trial and beyond: informing the future with lessons from the past. Clin Cancer Res [Internet]. 2010 Dec 15 [cited 2014 Jul 16];16(24):5963–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21169249

25. Sanoff HK, Carpenter WR, Stürmer T, Goldberg RM, Martin CF, Fine JP, et al. Effect of adjuvant chemotherapy on survival of patients with stage III colon cancer diagnosed after age 75 years. J Clin Oncol [Internet]. 2012 Jul 20 [cited 2014 Jul 16];30(21):2624–34. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3412313&tool=pmcentrez&rendertype=abstract

26. Deasy JO, Bentzen SM, Jackson A, Ten Haken RK, Yorke ED, Constine LS, et al. Improving normal tissue complication probability models: the need to adopt a "data-pooling" culture. Int J Radiat Oncol Biol Phys [Internet]. 2010 Mar 1 [cited 2015 Jan 29];76(3 Suppl):S151–4. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2854162&tool=pmcentrez&rendertype=abstract

27. Acharya CR, Hsu DS, Anders CK, Anguiano A, Salter KH, Walters KS, et al. Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer. JAMA [Internet]. 2008 Apr 2 [cited 2015 Jan 29];299(13):1574–87. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18387932

28. Nevins JR, Huang ES, Dressman H, Pittman J, Huang AT, West M. Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. Hum Mol Genet [Internet]. 2003 Oct 15 [cited 2015 Jan 29];12 Spec No(2):R153–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12928487

29. Meldolesi E, van Soest J, Dinapoli N, Dekker A, Damiani A, Gambacorta MA, et al. An umbrella protocol for standardized data collection (SDC) in rectal cancer: a prospective uniform naming and procedure convention to support personalized medicine. Radiother Oncol [Internet]. 2014 Jul [cited 2014 Nov 6];112(1):59–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24853366

30. Brooks RG, Rabin R DCF. The measurement and valuation of health status using EQ-5D: a European perspective: evidence from the EuroQol BIOMED Research Programme. Boston: Kluwer Academic; 2003.

31. Dolan P. Modeling valuations for EuroQol health states. Med Care [Internet]. 1997 Nov [cited 2015 Jan 30];35(11):1095–108. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9366889

32. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst [Internet]. 1993 Mar 3 [cited 2015 Jan 22];85(5):365–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8433390

33. Whistance RN, Conroy T, Chie W, Costantini A, Sezer O, Koller M, et al. Clinical and psychometric validation of the EORTC QLQ-CR29 questionnaire module to assess health-related quality of life in patients with colorectal cancer. Eur J Cancer [Internet]. 2009 Nov [cited 2015 Jan 30];45(17):3017–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19765978

34. Gujral S, Conroy T, Fleissner C, Sezer O, King PM, Avery KNL, et al. Assessing quality of life in patients with colorectal cancer: an update of the EORTC quality of life questionnaire. Eur J Cancer [Internet]. 2007 Jul [cited 2015 Feb 3];43(10):1564–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17521904

35. Jayasurya K, Fung G, Yu S, Dehing-Oberije C, De Ruysscher D, Hope A, et al. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. Med Phys [Internet]. 2010 Apr [cited 2014 Mar 28];37(4):1401–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20443461

36. Oh JH, Craft J, Al Lozi R, Vaidya M, Meng Y, Deasy JO, et al. A Bayesian network approach for modeling local failure in lung cancer. Phys Med Biol [Internet]. 2011 Mar 21 [cited 2014 Mar 28];56(6):1635–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21335651

37. Van Stiphout RGPM, Lammering G, Buijsen J, Janssen MHM, Gambacorta MA, Slagmolen P, et al. Development and external validation of a predictive model for pathological complete response of rectal cancer patients including sequential PET-CT

imaging. Radiother Oncol [Internet]. 2011 Jan [cited 2014 Mar 28];98(1):126–33. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21176986

38. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform [Internet]. 2006 Jan [cited 2015 Jan 30];2:59–77. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2675494&tool=pmcentrez&rendertype=abstract

39. Boyd S. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Found Trends® Mach Learn [Internet]. 2010 [cited 2014 Mar 19];3(1):1–122. Available from: http://www.nowpublishers.com/product.aspx?product=MAL&doi=2200000016

40. Graham Klyne JJC. Resource Description Framework (RDF): Concepts and Abstract Syntax [Internet]. 2004. Available from: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

41. Assélé Kama A, Choquet R, Mels G, Daniel C, Charlet J, Jaulent M-C. An ontological approach for the exploitation of clinical data. Stud Health Technol Inform [Internet]. 2013 Jan [cited 2014 Mar 28];192:142–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23920532

42. Harrel F. Resampling, Validating, Describing, and Simplifying the Model. Springer-Verlag; 2001. p. 87–103.

43. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak [Internet]. 2008 Jan [cited 2014 Mar 28];8:53. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2611975&tool=pmcentrez&rendertype=abstract

44. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making [Internet]. 2006 [cited 2014 Mar 27];26(6):565–74. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2577036&tool=pmcentrez&rendertype=abstract

45. Boneva I, Gayo JEL, Hym S, Prud'hommeau EG, Solbrig H, Staworko S. Validating RDF with Shape Expressions. 2014 Apr 4 [cited 2015 Mar 9];1–35. Available from: http://arxiv.org/abs/1404.1270

46. Gorlia T, van den Bent MJ, Hegi ME, Mirimanoff RO, Weller M, Cairncross JG, et al. Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE.3. Lancet Oncol [Internet]. 2008 Jan [cited 2014 Nov 17];9(1):29–38. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18082451

47. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc [Internet]. 2012

[cited 2014 Mar 28];19(5):758–64. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3422844&tool=pmcentrez&rendertype=abstract

48. Liu K, Kargupta H, Member S, Ryan J. Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining. IEEE Trans Knowl Data Eng. 2006;18(1):92–106.

APPENDIX 1

According to the legislation in place, the informed consent can be omitted for:
- Ethical reasons attributable to the fact that the person ignores his condition.
- Organizational reasons: the number of patients who can not be contacted to inform them, compared to the total number of persons who will be involved in research, produce significant consequences for the study in terms of alteration of their results

Given the large number of patients (large database) the calling of patients for a informed consent would require an organizational effort exaggerated

Many of the patients could be deceased and a call to verify the possibility to sign an informed consent could determine, in the case of deceased patient, a psychological results in family (ethical reasons).

The signing of a generic informed consent is not accordance with laws but, in a large database, the type of analysis can be identified and changed over time. For this reason every time that a research proposal will be done, the patients should sign a new informed consent (organizational reasons).

It is not possible to sign a specific informed consent to the patient because the data are taken from a anonymized database, therefore also the system is not even aware of which patients will be part of the analysis (organizational reasons)

The informing of the patient about the need to use its data to know if his treatment has results inferior or superior to another treatment may create psychological issue in the patient since the treatment was already done (ethical reasons).