PRODIGE: PRediction models in prOstate cancer for personalized meDIcine challenGE

Authors: Alitto AR[1], Gatta R[1], Vanneste BGL[2], Vallati M[3], Meldolesi E[1], Damiani A[1], Lanzotti V[1], Mattiucci GC [1], Frascino V[1], Masciocchi C[1*], Catucci F[1], Dekker A[2], Lambin P[2], Valentini V[1], Mantini G[1]

Corr. Author*: Masciocchi Carlotta

annarita.alitto@policlinicogemelli.it

Tel N: +39 0635034736

Fax N.: +39 0630155908

Affiliations: [1]Radiation Oncology Area, Gemelli-ART, Catholic University of the Sacred Heart, Rome, Italy.

[2]Department of Radiation Oncology (MAASTRO), GROW – School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, The Netherlands.

[3]School of Computing and Engineering, University of Huddersfield, Huddersfield, United Kingdom.

ABSTRACT

Background
Identifying the best care for a patient can be extremely challenging. To support the creation of multifactorial Decision Support Systems (DSSs), we propose an *Umbrella Protocol*, focusing on prostate cancer.

Materials and Methods
The PRODIGE project consisted of a workflow for standardizing data, and procedures, to create a consistent dataset useful to elaborate DSSs. Techniques from classical statistics and machine learning are adopted.
The general protocol accepted by our Ethical Committee can be downloaded at HTTPS://doi.org/10.17195/candat.2016.09.1 (www.cancerdata.org).

Results
A standardized knowledge sharing process has been implemented by using a semi-formal ontology for the representation of relevant clinical variables.

Conclusions
The development of DSSs, based on standardized knowledge, could be a tool to achieve a personalised decision-making.

Key Words: Decision Support System, Individualized medicine, Large Database, Machine Learning, Ontology, Predictive Model

INTRODUCTION

Over the past decades, many advances have been made in cancer care, as in radiation oncology [1-2]. Thanks to these advances, "Personalized medicine" is gaining importance, and it is becoming one of the challenges faced by clinicians. In order to adequately support the resulting decision-making process, there is a need to develop new tools.

Traditionally, clinical practice has been based on evidence-based guidelines, crafted by considering meta-analyses and randomized trials. Generally, the population subgroup enrolled shows homogeneous features, often without considering costs, and is impossible to include all possible characteristics and values [3]. For this reason, resulting evidence are sometimes hard to adapt in daily clinical practice, where actual patients may significantly differ from those enrolled in the subgroups. Moreover, trials need a long follow up: resulting evidence could be outdated at the time of publication [3].

Even if large randomized trials and meta-analyses or systematic reviews play a key role, they need the integration of emerging new different approaches, also the findings of observational studies and the variability of patients' features [4]. Moreover, a large amount of different types of data, with their increased complexity, and also the technologies progress need to be considered in the decision-making process [4]. Owing to the heterogeneous features of tumours and patients, the decision-making process needs to consider a lot of different variables, without the possibility to trail every combination [1]. This increasing amount of covariates is hardly analysed by human cognitive capacity, which discriminates a limited number of factors per each decision process [5]. To reach a "personalized medicine" level, there is therefore a growing need of decision support systems.

A clear systematic data-collection, and the identification of variables of interest, are two essential steps to create large databases, that can be used for fostering personalized medicine. Collecting data for such purposes implies a standardized way to represent the meaning of any variable, and the adoption of a well-defined methodology, to share such meaning, addresses the point of frequent innovation in cancer care. This problem can be tackled using an explicit representation of the involved ontology. In this paper we will adopt the "ontology" definition proposed by Gruber [6]: "ontology is a (formal) specification of concepts, relations and functions in a domain and hence focus on concepts".

The ontology-based methodology supports the creation of large databases. Over the last twenty years, Computer Science research has been carried out in order to develop personalized medicine goals, providing tools for diagnosis, treatment, supporting decision-making process and knowledge representation [7]. Due to the limited number of variables analysed by human capacity [5], solid DSSs become relevant in clinical decisions and a significant amount of research is focusing on this aspect [8-11]. In a recent publication, the TRIPOD statement introduces recommendations on quality of prediction models reporting [12].

Several interactive DSSs (Partin Tables, Kattan nomograms, D'Amico tables, CAPRA score, CaPSURE/CPDR Recurrence Equation) and many different ontologies (as the Unified Medical Language System (UMLS), National Cancer Institute Thesaurus (NCI), etc) have been developed for prostate cancer issues in clinical practice, but none of them are: a) specific for radiation therapy issues and b) designed to deal with the frequent innovation in radiotherapy and in the broad area of oncology.

The strategy to collect data in a standard and consistent manner, and to analyse data in a way that suits decision support purposes, is called "umbrella protocol" [13].

The aim of PRODIGE project is to elaborate an Umbrella Protocol related to prostate cancer, able to: collect a standardized large amount of heterogeneous features in large databases; use both retrospective and prospective data; analyse variables by modern and advanced statistical techniques; be flexible, for being able to deal with different non pre-determined endpoints.

This methodology would cover all aspects of prostate cancer care through the mentioned collection of heterogeneous data from patients in large database, using the "semi-formal ontology" developed.

Furthermore, after testing and validation, DSSs will be delivered according to specific needs and used in clinical workflow to choose the better way to treat patient.
Application of a model into daily clinical practice requires also the comparing, in a "controlled way", with the results of pre-existing trials: future perspectives will also include this comparison between DSSs results and "regular" human decisions.

MATERIALS AND METHODS

The Umbrella Protocol workflow developed (Fig. 1) is characterized by the following phases:
-Standardized Knowledge Sharing (SKS), that is the definition of a system to collect heterogeneous data in a standardized way, to create large databases;
-Standardized DSS Development (SDSSD), concerning the definition of a specific study, analysis method identification, model validation, model delivering, and respecting ethical issues at every step.

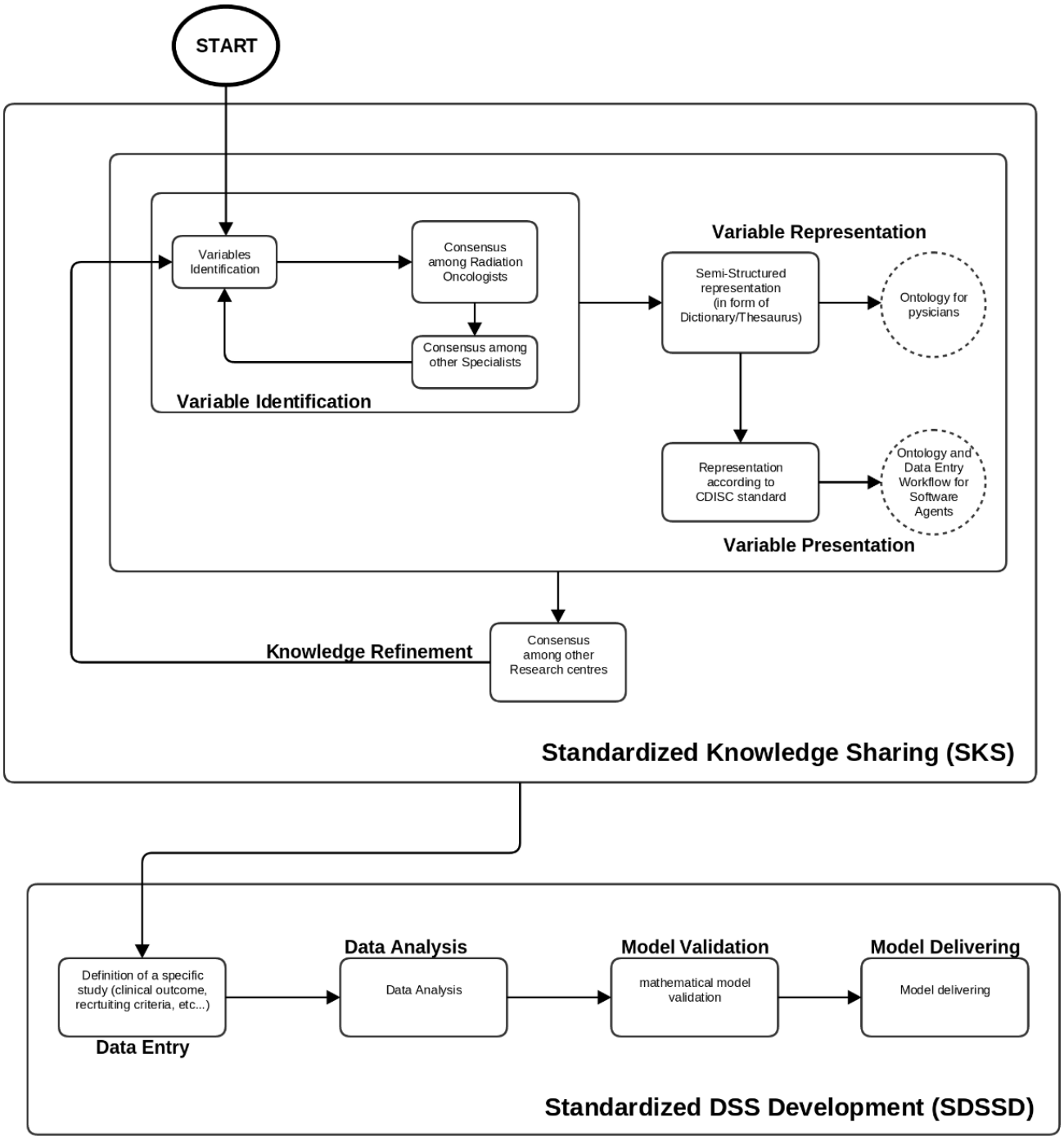

Fig. 1: The general workflow of PRODIGE Project

Standardized Knowledge Sharing
The steps of this phase are:

- *Variable identification*: clinical variables, anthropometric measures, clinical outcome, etc. are listed and shared among Radiation Oncologists and, subsequently, with other Specialists (Radiologists, Urologists, etc.). In this task, the focus is on collecting the largest number of widely applicable and significant variables.

- *Variable representation*: this step reduces ambiguity in sharing the knowledge and allows a software system elaboration, representing the domain of interest, by writing a dictionary for the previously identified variables. The main output is the production of an "ontology". A Medical Ontology is a linguistic/logical model used to represent the concepts which compose the knowledge of a clinical domain; it contains all relevant concepts, related to a clinical field, organized in a formal way (or informal, in any explicit case), that allows to perform "reasoning" by automatic inference. The developed prostate ontology was written in a semi-formal language for Specialists. Attention should be paid to re-using existing and validated ontologies, giving a clear reference to collected concepts.

- *Variable presentation*: to present the chosen variables in a formal way, compatible with the state of the art in Medical Informatics, for developing a software architecture supporting data entry by specific interfaces and/or interacting directly with existent EHRs, avoiding human involvement in data entry.

- *Knowledge refinement*: to share the output of these steps with other medical centres, collect the suggestions, and repeat the entire loop. The success of this step is not the blind "agreement" of the proposed representation but the suggestions, feedback, and improvement requests obtained from it.

The main goal of SKS is to provide a more formal representation of clinical knowledge, overcoming the limitations of natural language ambiguity, which is to attribute the same meanings to the variables. The aims are to: a) build a shared knowledge, b) build a semi-formal representation of the knowledge, c) enlighten current standards and the feasibility of an IT infrastructure supporting the study, d) share the knowledge with other centres, e) support a constant update, to be aligned with the advances of the state of the art.

Standardized DSS development
The main steps of this phase are:

- *Data Entry*: it can be performed manually into specific electronic forms, assisted by software agents (SA) checking that data is correct, or can be totally performed by SA, i.e. by an integration with an existing Electronic Health Record (EHR). In this scenario, the documents produced in the *variable representation* and *variable presentation* steps are fundamental.

- *Data Analysis*: this step has to face several important tasks:

    - *Pre-processing*: the actual features to be used for the generation of the models are chosen, according to both their mutual correlation or by exploiting some selection strategy techniques. Data quality is improved by identification and correction of missing data, outliers and bias [14-15]. In order to validate the model, training and validation sets can be considered: the training set is used to train the mathematical models, and the

testing set to measure performance and confirm the usefulness of models. If an external validation set is not available, available data could be split into two groups. Models trained can also be used to improve the quality of the chosen features, i.e. by the adoption of greedy forward or backward elimination approaches.

- *Data Quality Assurance*: a formal ontology allows the implementation of appropriate data quality assurance policies to: a) detect lexical (i.e.: error in data format), or semantic errors (i.e.: the end of a therapy registered before the beginning) b) identify and reduce the missing data effect, c) find hidden bias in the enrolled population.

- *Computation* of predictive models is based on two families of data analysis tools: techniques from Classical Statistics and from Machine Learning (ML). Classical Statistics include inferential regression analysis tools (linear and non-linear), survival models, etc.; ML methods include, for instance, Bayesian Network, Support Vector Machines, Random Forests, Artificial Neural Networks, etc.
  ML is a branch of artificial intelligence frequently used in cancer diagnosis and detection, and more recently also in prognosis and prediction [16-17], modelling the main outcomes of cancerous conditions.

- *Patients' Privacy* Protection: the local Ethics Committees (EC) shall approve the protocol before patient's accrual, according to the legislation of each country. Written informed consent for anonymized treatment data collection and approval of related research will be collected from each patient, according to local practice.

  Two methods were adopted to preserve patient's privacy:

  (a) Centralized consolidation of data records,
  (b) Distributed Learning approach.

  In the former (a), patient's privacy is protected by the architectural design: data from a local repository is anonymously transferred, through an encrypted pipeline, to a main repository, either by internet or other channels, without the possibility of associating clinical data to the patient. In this scenario, the mapping between data record and patients is protected by software procedures and such association never leaves the original centre. In this way, the centre's endpoint, queried by other research group member's out the institute, does not expose with any method to associate clinical data to the patient [13].

  When encryption is considered insufficient, a Distributed Learning approach (b) allows no patient data transfer out of the centre, but only the transfer of results, which are obtained via the computation on a big set of clinical data (for example the regression curve coefficients). This approach, for some algorithms, has been proven to have the same performance reached by joining all the datasets [18]. Moreover it can guarantee the highest level of patient's privacy, because no clinical data leaves the centre.

- *Model Validation:* Every model should be built using a training set and validated using an independent internal or external testing set, in accordance to TRIPOD statements [12]. The choice between internal or external depends on the data availability and the predictor aims. In any case, an analytic form of mathematical models is provided in order to rearrange the predictor using a different representation (for instance, for internal purposes, an external independent testing set is not the best choice). The coefficients of the mathematical models

are provided, in order to allow a user to rearrange the predictor using a different configuration. Similarly, residual analysis, performance indexes like c-statistics, Area Under the Curve (AUC), Receiver Operating Characteristics (ROC), calibration plot, F-score, etc., are provided.
Furthermore, the application of a model into daily clinical practice requires the matching between the results of pre-existing trials and meta-analyses, to compare the standard care with the personalized care [11]. All the process needs to be clear and published.

- *Model Delivery*: final model, optimised and validated, can be delivered through many channels, such as a nomogram, interactive website, scientific paper, app for smartphone, etc., according to specific needs. The product is delivered with clear instructions about the population it refers to, methods and results of the experimental phase of its development.

RESULTS

The proposed workflow has been developed to provide results for the Variable Identification and Representation steps, as showed before (Figure 1). The SKS considers two kinds of partitioning: horizontal, where patients' data come from different centres, and vertical, where data, regarding each patient, refers to different features (clinical, pathological, imaging, etc).

Variable Identification and Representation

All prostate cancer patients can be enrolled, including both retrospective and prospective information related to diagnosis and treatment.

Each feature was included in a terminological system with measurement units' specification, acquisition modality, range, etc., based on pre-existing ontologies, when available.

Despite these preliminary results, this methodology makes features meaning clearer and shareable, allowing data's re-usability both in space (among different research groups), in time and also in different research aims.

A team of prostate cancer specialists selected more than 200 features, organized into a *dictionary* divided in three tiers, according to the level of granularity. Each concept has been described with a unique reference, preferably correlated to a published coding system (e.g. NCI Thesaurus, CTCAE, SNOMED-CT etc.) and a trade-off has been adopted between the formal explication of the ontology and its effective usability, to increase simplicity. Therefore, this ontology is explicit, furthermore not formal, and designed to be "easily" formalized by one of the available languages (i.e. RDFS, OWL, etc.).

In detail, the first and more general tier, Registry level, includes all demographical-epidemiological information (Tab. 1).

The second Procedure level, records all clinical, pathological and treatment information and also several outcome features.

General patient characteristics are more detailed (Tab. 2a, i.e. height, weight, BMI, Prostate Antigen Specific-PSA, Testosterone). Tumour features include (Tab. 2b): imaging information; clinical and pathological classification according to TNM (American Joint Commission on cancer) [19]; histological (ICD-O) [20] and Gleason score classification [21-22], etc. Moreover, all prostate cancer treatment characteristics (i.e., hormonal, systemic, radiotherapy or surgical treatments), toxicities, according to CTCAE v3.0 or 4.03 [23-24], and RTOG scale [25] are recorded. Finally, several outcome features are reported (i.e., Biochemical Disease Free Survival, Overall Survival, etc).

The third Research level (Tab. 3), includes clinical and imaging data, for advanced research projects, such as Radiomics [26]. Pre-existing general and prostate quality of life questionnaires (EuroQol-5D-5L, EORTC QLQ-C30, EROTC QLQ-PR25, IPSS or EPIC score) and other tumour features are collected. Diagnostic imaging and radiotherapy planning information are uploaded for future re-elaboration, feature extraction and dose distribution analysis.

In parallel, we began to explore DSSs validation, to clearly identify and describe performance and limitations [28].

In particular, our methodology and tools have been verified on a small sample of 123 prostate cancer patients, to provide a validation of our software, that we will use for our next analysis on a big sample [29] to elaborate DSSs.

We focused our tests on developing techniques and methodologies to train DSS in multi-centric environment ensuring patient's privacy, without exchanging patient's data [18].

The text of the whole umbrella protocol developed for all cancer sites and approved by our Ethical Committee is at https://doi.org/10.17195/candat.2016.09.1 (www.cancerdata.org).

DISCUSSION

The PRODIGE project created a prostate cancer Umbrella Protocol supporting DSS' development, with the proposal of a procedure and an ontology, in a multi centric/specialistic environment. Umbrella Protocol requires a flexible strategy to: collect a large amount of heterogeneous data, and also in a flexible manner; data mining; develop DSSs and report outcome [3].
Building DSSs is a complex task, due to multidisciplinary professionalism interaction, heterogeneous data (clinical data, images, molecular/genomic data, etc.) and geographically distributed data sources (Clinical Databases, Image Repositories, Excel data sheets, ECG/EEG/ABPM, etc.).

After identification of the main features related to prostate cancer in the Variables Identification step, patients' history variables were assorted in categories (e.g., diagnostic variables, staging, or treatment) into a dictionary in the subsequent Variable Representation step. Because of the possible ambiguity of some terms, we tried to base our dictionary on pre-existing ontologies, looking for a reference for each variable.
Selected variables were encoded as a tabular representation, and defined using a table in natural language composed by: <class name, variable name, definition, measurement>. No lexical/syntactical rules were defined.

Considering Variable Presentation, we are still developing a model, an ontology and data entry workflow, for SA. The dictionary is designed to be compatible to exchange data in a common format, for possible future certification.
To end SKS and complete and integrate the semi-structured knowledge's representation, a step of knowledge optimization will take place towards a consensus achievement of the dictionary among other centres.
The choice of the language to represent ontology was the first critical problem. While many formal languages (such RDF, OWL) and software tools are available for this purpose [27], a lot of them are not designed to be used by physicians, and will increase the complexity of writing, checking, and upgrading [6]. After an in-depth analysis of existing approaches, performed by a multidisciplinary team (clinicians, engineers, mathematicians), we identified the best trade-off in terms of simplicity and a structured representation of the interested concepts, even if it does not use formal representation. For this reason, we used pre-existing formal ontologies (like NCI, etc) to build this "semi-formal" tool to collect data in a standardized way. "Semi-formal" is a technical term, concerning with the level of "ambiguity" allowed by the language. The so built ontology is explicit, even not formal, and can be "easily" formalized by one of the available languages for this purpose (i.e. RDFS, OWL, etc.).

In parallel, exploiting SDSSD, another key point was related to the distributed learning architecture: due to the high heterogeneity of hospital technologies and policies, in terms of patient's privacy and technicalities (firewall rules and IT offices), a team of engineers and mathematicians proposed a flexible solution adaptable to local needs and able to work in the general multi-centric framework.

A small sample of 123 prostate cancer patients was used only to validate our methodology and our developed software [29]: DSSs elaboration will need a bigger sample, even from multiple centres.

After the previous experience in colo-rectal cancer [3, 7, 13], we are adopting umbrella protocol framework for prostate and further for all cancer sites, even if few centres are investing in these new tools and methodology; to overcome this possible limitation, a larger network is going to be created, to share and consolidate this methodology and elaborate and validate predictive models.

The future of cancer research is based on a deeper multidisciplinary collaboration, for a hybrid discipline encompassing oncology. The common challenge is to effectively exploit the massive amount of data generated by researchers and clinicians, in order to develop accurate and scientifically-based decision tools for a shared decision-making process [30]. These decision tools will allow moving towards participative medicine [15] and, in the case of expensive treatments, involve a-priori individualized cost effectiveness analysis [31].

CONCLUSIONS

Nowadays, emerging observational studies, the so-called “Rapid Learning Approaches”, are crucial to confirm trials and meta-analyses results, identifying new population risks groups and check whether practice has appropriately changed [32].
By these research pathways interactions, predictive models could integrate existing guidelines and consensus, overcoming risks of patient over/under-treatment [3] hence having an impact also on the cost [4]. The analysis of cost-effectiveness will be an important endpoint for further investigations and it is a challenge to better address resources. Through designing, developing and testing a framework to represent data in a re-usable way, DSSs’ development will be possible, based on automatic extraction of the appropriate features for considered outcome. Obtained DSSs will provide a practical support to clinical choices for a specifically tailored medicine, by combining routinely collected clinical treatment data and innovative features (i.e. outcome information, diagnostic and treatment images). It will be an opportunity to move towards participative medicine, with evidence level 1 [11, 33].
According to the emerging approaches in this field, the DSSs will be able to overcome the limitation of classical clinical data and analyse innovative features (i.e. features extracted from images, etc.). An emerging necessity of multidisciplinary integration with different figures beside the clinicians is a crucial step to answer the need of care of often complex and puzzling diseases as cancer [34]. Even a multicentric collaboration is needed to realize this methodology and obtain robust DSSs.
It is pivotal to bear in mind that a predictor can be useful and can show great performance, but it remains only a tool; it is not the decision maker that will be the multidisciplinary equip together with the patient.

Variables	Definition	Measurement
Eligibility criteria		
Prostate Cancer Classification	According to the ICD-9 classification	http://www.icd9data.com/2014/Volume1/140-239/default.htm
	According to the ICD-10classification	http://www.icd10data.com/ICD10CM/Codes/C00-D49/C51-C58
General characteristics		
Institute	Hospital/Institute where patient was treated	Europe: EU-Country code (CC)-Institute number (IN) North America: AN-CC- IN South America: AS-CC-IN Asia: AA-CC-IN Australia: AU-CC-IN
Age@at primary Diagnosis	At diagnosis	Years
Date@Diagnosis	At diagnosis	Day/Month/Years
Age@RT	At start of any types of radiotherapy treatment (first fraction)	Years
Ethnicity		***Table 1***
Age at first recurrence diagnosis		Day/Month/Years
Age at first metastasis diagnosis		Day/Month/Years
Outcome		
Death		0: No – last FUP data (Day/Month/Year) 1: Yes – data of death (Day/Month/Year)
Cause of death		***Cause of death Table***
Date of death		Day/Month/Years

Tab. 1: Extract from Prostate Registry Level

Variables	Definition	Measurement
Body height	before start of treatment	cm
Body weight	before start of treatment	kg
BMI	Body Mass Index	BMI: mass (Kg) / (height (m))^2
ACE-27: COMORBIDITIES SCORING	ACE-27: COMORBIDITIES SCORING http://www.rtog.org/LinkClick.aspx?fileticket=oClaTCMufRA%3D&tabid=290	ACE-27: COMORBIDITIES SCORING
Previous Oncological History	Site	Specify
	Treatment	0: no 1: Yes (if yes, specify and complete relative fields) 999: missing data
	State of previous disease(according to RECIST criteria; if not applicable, refer to specific disease' ontology) RECIST: http://www.recist.com/recist-comparative/01.html	0: NED 1: Stable complete response 2: Stable partial response 3: progression disease 999: missing data
Multidisciplinary (MDT) management	yes/no	0: no 1: MDT discussion only without patient 2: MDT discussion with patient 999: missing data

Tab. 2a: Extract from Prostate Procedure Level – General characteristics

Variables	Definition	Measurement
Staging System		0: AJCC – TNM v5.0 1: AJCC – TNM v6.0 2: AJCC – TNM v7.0 999: Missing data
Tumour Location	Different site of the tumour in the prostate gland and combinations	1: Left Lobe 2: Right Lobe 3: Apex 4: Seminal vesicles 5: Basis 7: Central part 8: Peripheral part 9: Other specify 999: missing data Combination
Histology modality	Method used to obtain histology	0: TURP 1: Adenomiectomy 2: Needle biopsy 3: Citology 4: Surgery procedure 999: missing data
Date of Histology		Date: dd/mm/yyyy
Histology	Specification of histology (also subtypes if specified) http://whqlibdoc.who.int http://bioportal.bioontology.org/ontologies/NCIT/?p=classes&conceptid=http%3A%2F%2Fncicb.nci.nih.gov%2Fxml%2Fowl%2FEVS%2FThesaurus.owl%23C7378 http://bioportal.bioontology.org/ontologies/NCIT/?p=classes&conceptid=http%3A%2F%2Fncicb.nci.nih.gov%2Fxml%2Fowl%2FEVS%2FThesaurus.owl%23C2919	0: Adenocarcinoma, NOS 1: Adenocarcinoma 2: Neuroendocrine tumors 3: Other (specify) 999: Missing data others subtypes -Adenocarcinoma, NOS - Adenocarcinoma tipo acinare - Ductal Carcinoma ì - Mucinous Carcinoma - Signet RING cells Carcinoma -Neuroendocrine carcinoma - Oat-cell carcinoma - Carcinoma Undifferentiated bot oat cells - Squamous and Adenosquamous Carcinoma -Sarcomatoid Carcinoma (carcinosarcome)
Gleason Score 1	The first types of cancer cell present in the samples, numbering each type from 1 for the least affected up to 5 for the most affected. *Gleason DF. Classification of prostatic carcinomas. Cancer Chemother Rep. 1966; 50: 125-128.* *Gleason DF, Mellinger GT. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urol. 1974 Jan;111(1):58-64.* *Epstein JI, Allsbrook WC Jr, Amin MB et al. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. Am J Surg Pathol 2005 Sep; 29(9): 1228-42* *Montironi R, Cheng L, Lopez-Beltran A, et al. Original Gleason system versus 2005 ISUP modified Gleason system: the importance of indicating which system is used in the patient's pathology and clinical reports. Eur Urol 2010 Sep; 58(3): 369-73*	0:1 1:2 2:3 3:4 4:5 999: missing data
Gleason Score 2	The second types of cancer cell present in the samples, numbering each type from 1 for the least affected up to 5 for the most affected. *Gleason DF. Classification of prostatic carcinomas. Cancer Chemother Rep. 1966; 50: 125-128.* *Gleason DF, Mellinger GT. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. J Urol. 1974 Jan;111(1):58-64.* *Epstein JI, Allsbrook WC Jr, Amin MB et al. The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. Am J Surg Pathol 2005 Sep; 29(9): 1228-42* *Montironi R, Cheng L, Lopez-Beltran A, et al. Original Gleason system versus 2005 ISUP modified Gleason system: the importance of indicating which system is used in the patient's pathology and clinical reports. Eur Urol 2010 Sep; 58(3): 369-73*	0:1 1:2 2:3 3:4 4:5 999: missing data

Tab. 2b: Extract from Prostate Procedure Level – Tumour characteristics

Variables	Definition	Measurement
Study/Trial number	Protocol number	Number
Medication	Concomitant medication (not therapeutic)	According to *the Anatomical Therapeutic Chemical (ATC) Classification System* ***Table 4*** http://www.whocc.no/atc_ddd_index/
Pre-existing QoL general challenges	Record the worst grade of general complaints according the EORTC QLQ-C30 and EQ-DL5, FACIT_D (Version 4), which occurred within 4 weeks before the date of histology	
Pre-existing QoL prostate challenges	Record the worst grade of rectal complaints according the EORTC QLQ – PR 25**,** PROMs, EPIC scoring 2 EPICE 26, IPSS which occurred within 4 weeks before the date of histology	
Tumour		
Tumour Markers		0: none 1: K-ras positive 2: EGFR positive 3: HER-Neu 4: p53 5: CEA 6: Cromogranin A 7: CDX2 8: CK20 9:MUC2 999: missing data
Tumour Markers - specimen		0:Biopsy 1:Surgical specimen
Imaging	Types	1: Trans Rectal Ultrasonography (TR EUS) 2: MRI (pelvis) 3: PET 4: CT 5: Bone Whole Body Scan (BWBS) combinations 999: Missing data
	Date	Dd/mm/yy
	DICOM Files	

Tab. 3: Extract from Prostate Research Level

Acknowledgments, Financial support and Conflict of interest

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in, or financial conflict, with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending or royalties.
No writing assistance was utilized in the production of this manuscript.

SUMMARY POINTS

Personalized Medicine

- "Personalized medicine" is defined by the National Cancer Institute (NCI) as a "form of medicine that uses information about a person's genes, proteins, and environment to prevent, diagnose, and treat disease. In cancer, personalized medicine uses specific information about a person's tumour to help diagnose, plan treatment, find out how well treatment is working, or make a prognosis".
- The tendency towards individualised medicine and the increasing amount and complexity of data, makes extremely difficult to identify which clinical decisions are better for a specific patients.
- In daily clinical practice, Decision Support Systems (DSSs) could help to personalize clinical choice.
- We propose a general conceptual/procedural framework (an Umbrella Protocol) which can help to represent and share the knowledge in clinical domain and reduce misunderstanding and improve efficacy in predictors development, in particular in studies among different institutes using large databases. We focus our attention on a specific implementation of such framework for prostate cancer.

Umbrella Protocol

- "The strategy to collect data in a standard and consistent manner and to analyze them properly for decision support is called <umbrella protocol>." [13]

PRODIGE

- The main features of an Umbrella Protocol, created in Radiotherapy Division of the Fondazione Policlinico Universitario A. Gemelli in Rome, for standardizing data and procedures to create a consistent dataset useful to obtain a trustful analysis for a Decision Support System (DSS) for prostate cancer are reported.
- It is a part, specific for prostate cancer, of a whole protocol for all cancer sites, named ULISSE, approved by the Ethical Committee of Fondazione Policlinico Universitario A. Gemelli, in Rome

Standardized Knowledge Sharing process

- A phase to realize a formal or semi-formal representation of knowledge, in order to overcame the limitations of the ambiguity of the natural language.
- This phase will benefit of an "ontology": a linguistic/logical model used to represent the concepts which composes the knowledge of a clinical domain. An ontology contains all the

relevant concepts, related to a clinical field, organized in a formal (or informal, in any case explicit) way that allows to perform reasoning by automatic inference.
It includes:

- Variable identification: list of variables of major interest.
- Variable representation: describing in a non-ambiguous way the identified variables.
- Variable presentation: to present the identified variables in a formal and structured way, compatible with the state of the art of Medical Informatics.
- Knowledge tuning: for supporting a continuous verification, upgrade and correction of the previous steps.

Standardized DSS development (SDSSD)

It concerns with the development of Decision Support Systems

- Data Entry: manually or assisted, it is crucial and it requires the variable representation and presentation steps, for a correct input of data in a non-ambiguous way.
- Data Analysis: ensuring patients privacy protection, it includes: a pre-processing step, to correct bias and missing data, and to identify a training and a validation sets to model development and test; data quality assurance, and; a computation step, with technique form classical statistics and machine learning.
- Model Validation: every model has to be built using a training set and evaluated by an independent internal or external testing set, for validation, with various performance measures.
- Model Delivery: provides a mean for delivering the generated models, under the form of nomogram, interactive website, scientific paper, app for smartphone, etc.

Future Perspectives

- Personalized cancer treatment is a challenge for the modern radiotherapy and for cancer disciplines in general. The development of Decision Support System, based on a Standardized Knowledge, represent the corner stone of a highly individualized, shared and participative decision making process.

References

1-Lambin P, van Stiphout RG, Starmans MH, et al. Predicting outcomes in radiation oncology – multifactorial decision support systems. Nat. Rev. Clin. Oncol. 10, 27–40 (2013).

2-Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer 48, 441-446 (2012).

3-Meldolesi E, van Soest J, Dinapoli N, 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. 112(1), 59–62 (2014).

4-Valentini V, Dinapoli N, Damiani A. The future of predictive models in radiation oncology: from extensive data mining to reliable modeling of the results. Future Oncol. 9 (3), 311-313 (2013).

5- Abernethy AP, Etheredge LM, Ganz PA et al. Rapid-learning system for cancer care. J. Clin. Oncol. 28(27), 4268–4274 (2010).

6- Gruber TR. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Int. J. Hum. Comput. Stud. 43 (5-6), 907–928 (1995).

7-Meldolesi E, van Soest J, Dinapoli N, et al. Medicine is a science of uncertainty and an art of probability (Sir W. Osler). Radiother. Oncol. 114(1), 132–4 (2015).

8-Lambin P, Roelofs E, Reymen B, et al. A. Rapid Learning health care in oncology' - An approach towards decision support systems enabling customised radiotherapy'. Radiother. Oncol. 109(1), 159-64 (2013).

9-Lambin P, Petit SF, Aerts HJ, et al. The ESTRO Breur Lecture 2009. From population to voxel-based radiotherapy: exploiting intra-tumour and intra-organ heterogeneity for advanced treatment of non-small cell lung cancer. Radiother. Oncol. 96(2), 145-52 (2010).

10-Lambin P, Zindler J, Vanneste B, et al. Modern clinical research: How rapid learning health care and cohort multiple randomised clinical trials complement traditional evidence based medicine. Acta Oncol. 54(9), 1289-300 (2015).

11-Lambin P, Zindler J, Vanneste BG, et al. Decision support systems for personalized and participative radiation oncology. Adv. Drug. Deliv. Rev. 109,131-15314 (2017).

12-Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann. Intern. Med. 162(10), 735-6 (2015).

13-Meldolesi E, van Soest J, Damiani A, et al. Standardized data collection to build prediction models in oncology: a prototype for rectal cancer. Future Oncol. 12(1), 119–136 (2016).

14-Cismondi F, Fialho AS, Vieira SM, et al. Missing data in medical databases: impute, delete or classify? Artif. Intell. Med. 58 (1), 63-72 (2013).

15- Boneva I, Gayo JEL, Hym S, Prud'hommeau EG, Solbrig H, Staworko S. Validating RDF with shape expressions. arXiv 1404, 1270 (2014).

16-Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2, 59–77 (2007) .

17-Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13, 8-17 (2014).

18-Damiani A, Vallati M, Gatta R, et al. Distributed Learning to Protect Privacy in Multi-centric Clinical Studies. Artificial Intelligence in Medicine Lecture Notes in Computer Science Volume 9105, 65-75 (2015).

19-(AJCC) AJC on C. AJCC 7th Edition - Prostate Cancer Staging (2007). https://cancerstaging.org

20-Montironi R, Cheng L, Lopez-Beltran A, et al. Original Gleason system versus 2005 ISUP modified Gleason system: the importance of indicating which system is used in the patient's pathology and clinical reports. Eur. Urol. 58(3), 369-73 (2010).

21-Sobin L, Parkin DM. International Classification of Diseases for Oncology. Third Edition. World Health Organization, Geneva, Switzerland. http://whqlibdoc.who.int
http://bioportal.bioontology.org/ontologies/NCIT/?p=classes&conceptid=http%3A%2F%2Fncicb.nci.nih.gov%2Fxml%2Fowl%2FEVS%2FThesaurus.owl%23C2919

22-Epstein JI, Zelefsky MJ, Sjoberg DD, et al. A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score. Eur. Urol. 69(3), 428-35 (2016).

23-18-CTEP. Common Terminology Criteria for Adverse Events v3.0 (CTCAE) (2006).
http://ctep.cancer.gov

24-Common Terminology Criteria for Adverse Events (CTCAE). U.S.Department of Health and Human Services, National Institutes of Health, National Cancer Institute. v4.03: June 14, (2010)
Services H. Common Terminology Criteria for Adverse Events (CTCAE) (2010). http://evs.nci.nih.gov

25-Cox JD, Stetz J, Pajak TF. Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). Int J Radiat Oncol Biol. Phys. 31,1341-6 (1995).

26-Kumar V, Gu Y, Basu S et al. Radiomics: the process and the challenges. Magn. Reson. Imaging 30 (9), 1234-1248 (2012).

27-Min H, Manion FJ, Goralczyk E, et al. Integration of prostate cancer clinical data using an ontology. J. Biomed, Inform. 42, 1035-1045 (2009).

28-Vallati M, De Bari B, Gatta R, et al. Exploiting Machine Learning for Predicting Nodal Status in Prostate Cancer Patients. Artificial Intelligence Applications and Innovations IFIP Advances in Information and Communication Technology 412, 61-70 (2013).

29- Dinapoli N, Alitto AR, Vallati M, et al. RadioBio data: A Moddicom Module to Predict Tumor Control Probability and Normal Tissue Complication Probability in Radiotherapy. In: Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies. SCITEpress (In Press)

30-Jiang H, An L, Baladandayuthapani V, et al. Classification, predictive modelling, and statistical analysis of cancer data (a). Cancer Inform. 13 (2), 1–3 (2014).

31-Cheng Q, Roelofs E, Ramaekers BL, et al. Development and evaluation of an online three-level proton vs photon decision support prototype for head and neck cancer - Comparison of dose, toxicity and cost-effectiveness. Radiother. Oncol. 118(2), 281-5 (2016).

32-Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. Br. J. Cancer 110(3), 551–555 (2014).

33-Stacey D, Légaré F, Col NF, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst. Rev. 28 (1), CD001431 (2014).

34-Alitto AR, Gatta R, Meldolesi E, et al. Personalized Medicine in prostate cancer: future perspectives for tailored treatments. J. Cancer Prev. Curr. Res. 3(5): 00092 (2015).