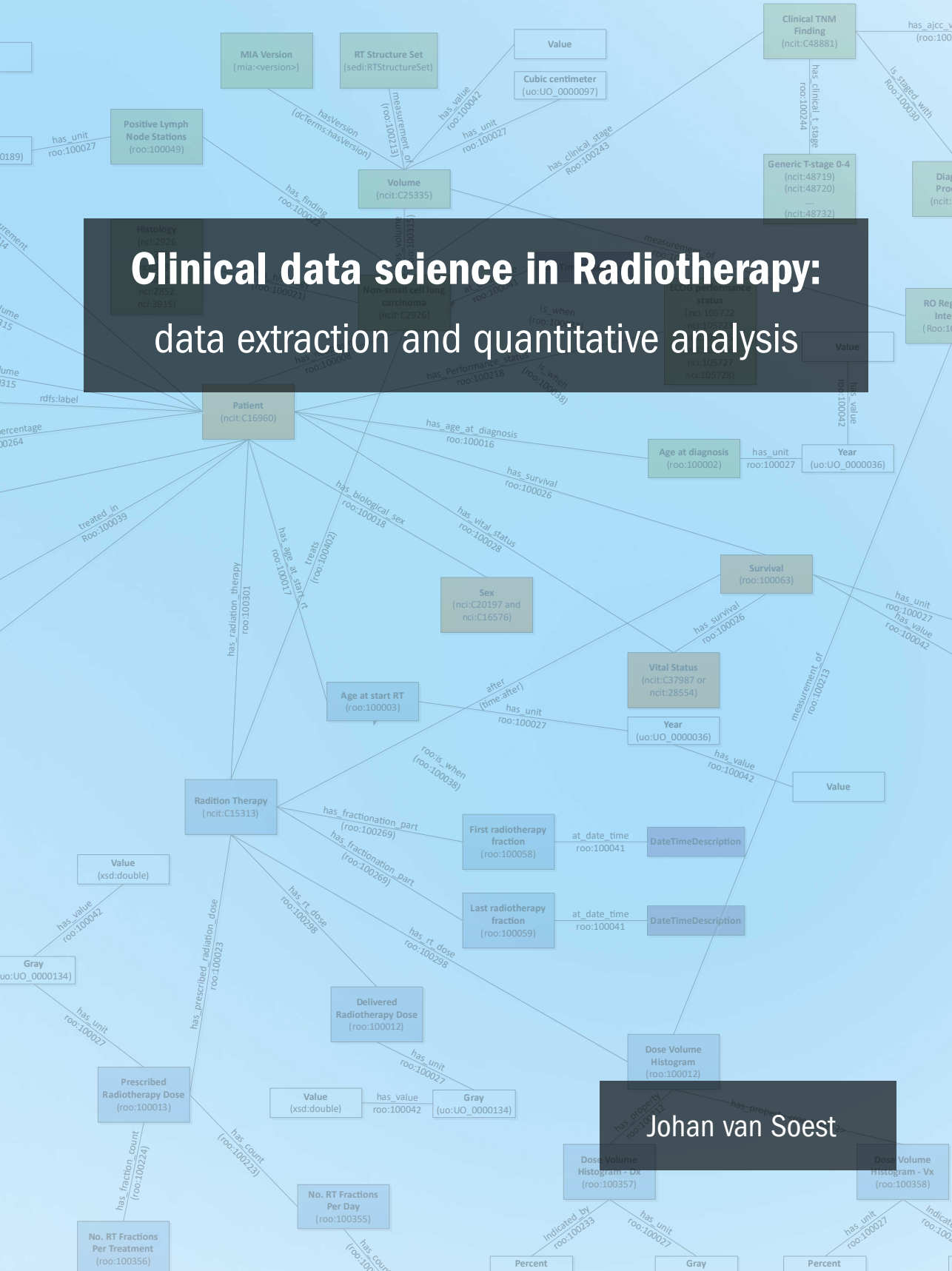
Clinical data science in Radiotherapy


Clinical data science in Radiotherapy:
data extraction and quantitative analysis
Johan van Soest
Patient
(ncit:C16960)
Radition Therapy
(ncit:C15313)
Prescribed Radiotherapy Dose
(roo:100013)
Delivered Radiotherapy Dose
(roo:100012)
Dose Volume Histogram
(roo:100012)
Volume
(ncit:C25335)
Survival
(roo:100063)
Vital Status
(ncit:C37987 or ncit:28554)
Sex
(nci:C20197 and nci:C16576)
Age at start RT
(roo:100003)
Age at diagnosis
(roo:100002)
Clinical TNM Finding
(ncit:C48881)
Generic T-stage 0-4
(ncit:48719)
(ncit:48720)
(ncit:48732)
Positive Lymph Node Stations
(roo:100049)
Non-small cell lung carcinoma
(ncit:C2926)
MIA Version
(mia:<version>)
RT Structure Set
(sedi:RTStructureSet)
First radiotherapy fraction
(roo:100058)
Last radiotherapy fraction
(roo:100059)
No. RT Fractions Per Day
(roo:100355)
No. RT Fractions Per Treatment
(roo:100356)
Dose Volume Histogram - Dx
(roo:100357)
Dose Volume Histogram - Vx
(roo:100358)
Value
Cubic centimeter
(uo:UO_0000097)
Year
(uo:UO_0000036)
Gray
(uo:UO_0000134)
DateTimeDescription
has_value
has_unit
has_finding
has_clinical_stage
has_age_at_diagnosis
has_survival
has_biological_sex
has_vital_status
has_radiation_therapy
has_fractionation_part
has_rt_dose
has_prescribed_radiation_dose
has_fraction_count
has_count
at_date_time
measurement_of
treated_in
rdfs:label

Printing: Datawyse | Universitaire Pers Maastricht

ISBN 978-94-6380-049-5

Clinical data science in Radiotherapy:

data extraction and quantitative analysis

DISSERTATION

to obtain the degree of Doctor at the Maastrich University,
on the authority of the Rector Magnificus Prof.dr. Rianne M. Letschert
in accordance with the decision of the Board of Deans,
to be defended in public
on the 28th of November 2018 at 10:00

by

Johan Peter Antoon van Soest

Affiliation

Department of Radiation Oncology (MAASTRO)
GROW School for Oncology and Developmental Biology
Maastricht University Medical Centre+
Maastricht, The Netherlands

Promotores

Prof. dr. ir. A.L.A.J. Dekker
Prof. dr. V. Valentini

Assessment committee

Prof. dr. D. de Ruysscher (Chair)
Prof. dr. G. Meijer (Nederlands Kanker Instituut, Universiteit Utrecht)
Prof. dr. A. Abu-Hanna (Universiteit van Amsterdam)
Prof. dr. R. Muller
dr. J. Buijsen

Contents

Chapter 1

Introduction

Prognostic and predictive models in medicine

The amount of automatically recorded data is growing rapidly [1]. The information embedded in this data is largely unstructured; hidden in free text, images, audio, video or other forms of continuous information sources (e.g. an electrocardiogram). However, we are more-and-more able to extract information from these unstructured sources, as well as recording or interpreting structured data.

The benefit of recording more data is that we can develop a more complete understanding about patients, diseases and the impact of a treatment. The downside is that our cognitive capacity is limiting the assertions and reasoning humans can perform on all the available information [1,2]. Specifically, our cognitive capacity is limited to 5 information elements, and decreasing when performing reasoning tasks using numerical values, or words with similar characters [1,3]. Given all the available data, in combination with limited human reasoning capacity, it becomes harder for humans to investigate whether specific findings or interventions are better or worse in prognostic or predictive performance. For example, does a generic categorical tumor staging still better represent the observed tumor status, in comparison to the more detailed information we can (semi-)automatically extract from images (e.g. a tumor length/volume, or more advanced numerical image characteristics)? When more detailed information is more accurate, we have to find intelligent methods to combine and represent this detailed information useful for clinical practice. This is the field of Clinical Decision Support Systems (CDSSs), where information is automatically extracted from sources, and can give an integrated view of patients while supporting the clinical workflow [4,5]. In a CDSS, clinical experts can see the (aggregated) information regarding a patient, including rule-based, statistical or advanced mathematical (e.g. machine learning) algorithms to give a prognosis of a patient. Development of such systems has evolved over time, from rule-based systems (e.g. implementing guidelines and questions) [6] to systems predicting the increase/decrease of the patient's outcome by administering a certain treatment (e.g. radiotherapy, chemotherapy, or its combination) [5]. Mind that these systems are supporting, and not replacing the clinical expert and patient's decision.

Machine Learning

One of the options to predict the patient's outcome, is to use machine learning (ML) algorithms. Machine learning is a sub-field of soft computing in Artificial Intelligence (AI); meaning that machine learning attempts to give the most accurate solution to a problem while respecting that it cannot reach the exact solution [7]. To find the most accurate solution, computers follow mathematical formulas and instructions (algorithms) to produce a new mathematical formula. By applying the learned mathematical formulas to new cases, we can give a prognosis of the outcome [8]. Learning these new mathematical formulas (or prognostic/predictive models) are in theory not limited to a

certain number of variables. These models can be used by humans and computers to make, for example, a prognosis/prediction, to group items with similar characteristics, detect anomalies, or detect patterns. However, it must be noted that the models' predictive abilities are limited to the data learned upon, thus influenced by the correctness and real-world reflection of the data used during training.

Models can be used in two ways: 1) interpreting the models, and making operational decisions based on the interpretation results 2) using the models in practice. In the latter case, the model itself has limited value, but has value when correctly integrated in practice (e.g. software and/or workflows) [8]. As a result, the user does not necessarily need to be aware of the actual algorithm, as the application where the model is being used is more important. Some examples of ubiquitous machine learning algorithms and applications are email spam-detection, facial detection when taking a picture, predictive text input (on smartphones) and digital personal assistants (showing the information you need, at the time/place you need it).

Machine learning (ML) algorithms are available in many different forms. However, we can distinguish two types of algorithms: supervised and unsupervised. Supervised algorithms use information and outcomes of historical cases to predict the outcomes of new cases. Unsupervised algorithms assess the data at hand, without any available outcome. As a consequence, unsupervised ML algorithms will give insights into the data, although these insights do not necessarily correlate to a (future) outcome.

Another aspect of ML algorithms is the complexity of the models produced. Some ML algorithms (like Support Vector Machines using a linear kernel) are interpretable by humans and can be conceptualized [9]. Other algorithms (like neural networks with hidden layers) are harder to interpret and conceptualize for humans. Hence, produced models of the latter are sometimes seen as "black-box" models: given several input variables, we cannot understand all steps taken by the produced model to reach its result [10].

Although using machine learning algorithms to produce models sounds straightforward, the produced models are optimized based on the data given to the algorithm to learn on. Therefore, they are dependent on how well data used for learning represents the treatment setting (e.g. other hospitals, future patients) in which the model is being used. For example, a model learned on the data of hospital A, might not work in hospital B. This may be due to different patient populations, treatment policy or other factors. This problem is not only limited to the level of hospitals, however may be even on higher levels (e.g. differences in guidelines among countries). Hence, produced models need evaluation and validation in *different* datasets [11]. During these evaluations, performance measures can be calculated that describe how well the models perform in this new dataset. Many different performance measures exist in ML, and are dependent on the type of algorithm learned. For supervised ML algorithms, these are based on the model-predicted outcome compared to the actual outcome [12–14].

To assess if a machine learned model is correct, and in how many cases, we therefore need to know characteristics of the dataset it was learned upon, characteristics of the dataset it was validated upon, and performance measures on both learning and validation datasets. A complete checklist for the information to report has been developed by a large consortium of academics, and is available in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [15].

The need for data

In general, most machine learning algorithms are hungry for data. The more information elements (variables) used in an ML algorithm, the more patients are needed to give a reliable estimate for future patients. For example, a causal relationship (e.g. smoking causes lung cancer) based on one patient does not hold, as we can find many patients who did smoke, but didn't have lung cancer or vice versa. Therefore, we need more patients to give a more reliable estimate of the *chance* of lung cancer for the investigated number of patients. The more people/patients involved in this estimate, the more accurate it will resemble the real world.

To give an even more accurate correlation, we could investigate if adding another variable to the relationship increases the correlation. For example, we could find that the correlation of smoking *and* age results in a more detailed view of the actual truth. However, by increasing the number of input variables (or covariates; in this case smoking status and age), we also need to increase the number of measurements (patients) to get a more reliable correlation. Hence, the more information we want to use in a machine learning algorithm (to make it more specific), the more subjects (e.g. patients) we need to represent a reliable result and to model the real-world situation. When a ML model is trained on too few subjects, the algorithm is highly susceptible to *overfitting* [13,16]. As the ML algorithm learned on a relatively small number of subjects (patients), the model is prone to optimize for these subjects. Hence, as described before, external validation (in a different dataset) is needed to test if a model has been overfitted.

Although it is generally advisable to include as many patients/subjects as possible, current research practice is limited in patient inclusion. Especially when clinical trial data is used for machine learning purposes (Chapter 11). A specific hospital can only include as many patients as are eligible for inclusion in a clinical trial, which is roughly 4-5% of all the patients seen in a hospital [17–19]. These patients mostly have a limited number of comorbidities and are not too young or old. Therefore, the number of patients in such datasets are sometimes too low, especially for a small hospital single-center clinical trial. As a consequence, we can argue for which purpose clinical trials are the gold

standard, and whether an observational analysis is sometimes a better approach to answer a scientific question [20].

To be sure to not "overfit" a model resulting from an ML algorithm, we historically use a rule-of-thumb for minimal dataset sizes introduced by Harrell et al [16]. This minimal sample size is based on the number of variables used for prediction and characteristics of the actual outcome. However, this rule-of-thumb (for every prediction variable we need at least 10 outcome events) is controversial, as it is based on only one study, and didn't consider the underlying (statistical) mechanisms of the ML algorithm and dataset characteristics [21]. Although these remarks are valid, there is no alternative criteria available to our knowledge. Hence, we will still hold to this lower boundary, until better methods are available.

To overcome the low sample size of clinical trials, it would be better to develop methods to include all patients available in a hospital for observational studies (Chapter 5). Although this sounds straightforward, several issues regarding secondary use of clinical data (Section 1, Chapters 2-5), development of a technical infrastructure, and standardization arise. Several of these issues are addressed in section 1 and 2 (Chapters 6-8).

Data standardization

As machine learning algorithms need large amounts of data, single hospital datasets are quickly becoming too small and vulnerable to generalizability issues as explained above. Hence, datasets from multiple hospitals are preferable, however introduces challenges in terms of interoperability. Terminologies, measurement methods, data storage, access methods and interpretation can widely differ among hospitals.

As explained in Chapter 3, standardization starts with the syntactical interoperability. This entails the structure for storage of information, and how to access this information. In essence, it is the syntax how our data can be retrieved for analysis purposes. For example, the definition of the same (order of) columns among all centers.

After the syntactical interoperability, we can move towards the semantic interoperability. This interoperability level (on top of the syntactical level) describes which terminologies to use and, related to these terminologies, what the definitions are for every value in every column. For example, hospital A might use the values 0 and 1 to represent respectively male and female. However, hospital B might use "M" and "F", or another language-dependent value (Chapter 4). Replacing these local values with standardized terminologies can overcome language barriers, and subsequent confusion of tongues during data analysis [22]. Hence, semantic interoperability among datasets from different hospitals is mandatory.

Although syntactic and semantic interoperability are well-known terms in clinical data and IT, ontologies are less known, or sometimes confused for standardized terminological definitions. Ontologies built on standardized terminologies provide the oppor-

tunity to describe the relationship among different concepts. For example, a patient can have one or multiple diseases, and patients can have one or multiple treatments. These treatments can have a relationship to individual diseases. Such relationships are normally only implicitly recorded in databases, due to the optimized performance of databases to address a specific use case (e.g. a medical record system). By using ontologies, these relationships can be described explicitly, and hence overcoming interpretation issues, as further explained in Chapter 4 and 6.

The use of syntactic and semantic interoperability is already ubiquitous in clinical/medical informatics. International terminologies such as the International Classification of Diseases (ICD) [23], SNOMED [24,25], LOINC [26,27] and the NCI Thesaurus [28] are commonly used. However, the ontological relationship between these terminological concepts to describe individual patients are less well defined. Therefore, we will describe the development and use case for such an ontology in radiation oncology (Chapter 6). By defining these terminologies and relationships (ontologies) using Semantic Web technologies, we are creating linked data [29,30]. Meaning that information is available on a URL (e.g. http://www.cancerdata.org/roo/) and more information can be found on this URL, where other data can be pointing to the same URL. In essence, this will create a web of data, which computers can interpret (instead of humans reading webpages). Chapter 6 shows an example of the possibilities of linked data, by linking public or external knowledge to private data and therefore enriching the knowledge we have on a specific patient.

The use of linked data does not end with clinical data stored in structured information systems. Chapter 7 will describe how imaging information can be converted into linked data, and can be connected to the linked data of Chapter 6. Subsequently, derived imaging information (e.g. measurements on images) can be converted into linked data as well, and connected to the work in Chapter 7 [31].
As mentioned in Chapter 7, naming conventions of (anatomical) structures on medical images are institute dependent, and are text fields in clinical systems and storage standards. Hence, automatic calculation of measurements on specific structures results in many errors. To overcome this specific issue, we developed software to link local naming conventions to developed standards. Chapter 8 describes this developed software for local naming conventions, and its application in calculation of image-derived information.

Cancer and Radiation Oncology

Cancer is caused by uncontrolled cell division, which invades local and surrounding tissues and can spread through the human body. It is the second leading cause of death worldwide, with 8 million deaths in 2012. Furthermore, the worldwide incidence of

cancer is estimated at 14 million cases per year, expecting to grow to 23 million cases per year in the next decade [32]. Although cancer can manifest itself in different forms, all forms of cancer share the same six acquired capabilities on a cellular level: the hallmarks of cancer as described by Hanahan and Weinberg [33]. These six capabilities are 1) stimulating cell division 2) de-activating cell division inhibitors 3) evading apoptosis (cell death) 4) self-sufficiency in growth signals 5) Insensitive to growth inhibiting signals 6) invasion in other tissues and metastases. Most cancer treatments can be related to one or more hallmarks. For example, irradiating the tumor will reduce cell division, invoke apoptosis (by killing the cell), and attempt to reduce invasion and metastases, or remove the tumor cells completely (due to apoptosis of tumor cells) [34].

By combining multiple treatment modalities (e.g. using radiation and systemic therapy, such as chemotherapy), the aim is to cover all hallmarks and hence to more effectively treat patients.

Radiation Oncology in Rectal Cancer

For locally advanced rectal cancer patients (LARC), the current treatment guideline prescribes (chemo-)radiotherapy and subsequently surgery (local excision or radical resection). However, recent insights have shown that for some patients, surgery is not needed as the tumor tissue was absent in the pathological surgery specimen [35–38]. This leads to the question whether surgery is needed for specific groups of patients, and whether we can identify these groups using non-invasive methods, at various timepoints during the treatment process [39–41]. If we can identify these groups of patients, we could tailor these patient's treatment by omitting surgery (and common consequences such as a stoma) [42]. On the other hand, by identifying patients with a worse prognosis, we could intensify (if the patient's health conditions and anatomy allows) the treatment using additional chemotherapy or other adjuvant treatments [35,43].

Measuring treatment outcome

To make sure treatments are effective, monitoring and measuring *outcomes* is imperative. These outcomes can describe the treatment effect on the tumor, or (adverse) patient effects. Treatment effects can be measured in three different categories:

1) Tumor presence and size after treatment onwards (local control)
2) Metastases occurring after treatment onwards (distant control)
3) Patient survival after treatment onwards [2].

All three categories can be measured using various methods. For example, tumor presence after treatment can be measured using medical imaging (e.g. CT or MRI), or by inspecting pathological specimen after surgery. Metastases after treatment can be detected using medical imaging, and patient survival can be measured by hospital or municipality records.

The (adverse) patient effects can be broader, not only in clinical, but also psychological, social, societal or financial perspective. From the clinical perspective, these effects are mostly – but not limited to – *toxicities* due to the received treatment [44,45]. These toxicities are adverse effects due to irradiating (at a lower intensity) healthy tissues, reducing their normal function and therefore resulting into discomfort or even severe complications. Patient effects can be recorded as clinician-reported, or patient-reported outcomes (PROs) [46,47].

Radiation oncology as an IT-supported environment

One of the treatments for oncology is radiation therapy. Using (external beam) radiation therapy, energetic particles are delivered to the tumor from multiple angles. By using these different angles, the number of particles delivered to the tumor is high, where the particles delivered to the traversed organs and tissues is lower. Hence, creating a focused target of irradiation.

To deliver these beams of radiation at the exact location, computerized information technology (IT) is used. For example, in creating a CT scan of the patient, delineating the tumor contours on this CT scan, calculating the best (optimized) treatment plan, and even the actual treatment itself (programming the linear accelerator to move in specific directions, and to deliver specific radiation intensities, for how many seconds) [48]. Hence, we can state that radiation oncology is already an IT-driven environment. This also means that many information elements (e.g. basic patient information, tumor information, treatment information, etc.) are already available in a structured format within the information systems and software needed to conduct radiation therapy [49]. When bringing this into perspective of other hospital departments, radiation oncology (and other IT-driven hospital departments, such as Radiology) already has a wealth of standardized information, and can play an exemplary role in the development of data-driven solutions to get better insights into the best treatment for specific (groups of) patients.

The main clinical examples throughout this thesis are focused on rectal cancer. Specifically, validation of existing, and development of prognostic and predictive models for patients diagnosed with locally advanced rectal cancer (LARC).

To identify these sub-groups of LARC patients, several prediction models have been built before [39,50,51]. In this thesis, these prediction models were validated in different settings: different hospitals (Chapter 9), datasets collected at later time points (Chapter 10) and in new and larger datasets including a longer follow-up period (Chapter 11).

The need for distributed data analysis

As mentioned previously, machine learning algorithms are as good as the data they reflect. The more detail used in prognostic/predictive models (using more patient characteristics as input variables), the more data we need to make robust models. In practice, it means that the number of patients needed to learn a detailed prognostic model already goes beyond the number of patients available in a single center [52]. Hence, we need to integrate routine-clinical data sources from multiple institutes to make the next step in prediction models [2,53]. This would lead to more robust prediction models (as models are learned on a larger dataset), and more clinically applicable prediction models (as we are including all clinical patients, instead of patients only included in clinical trials).

Including more centers' routine clinical data introduces issues on the organizational level as well. The more centers delivering data for analysis purposes, the more societal trending questions become apparent regarding data handling (who will maintain/own this large set of data?), trust (do I trust the data collecting institute to handle my data properly? Do I trust that any analysis won't lead to wrong/adverse perspectives?), and security (is one location containing all data an interesting source for hackers?) [54,55].

Although the need for larger datasets does not directly imply different forms of analysis, the need for better governance and control over data introduces a rethink on how we handle and analyze data. One of the possibilities is to keep the data at the collecting institute, and send the analysis algorithms around, rather than centralizing data. The only information shared is on an aggregate/statistical level per institute or subset, largely limiting the threat of re-identification of individual patients (Chapter 3). Furthermore, by not sharing the data, keeping track of data becomes more transparent, as the number of duplications is limited. This benefits data governance, including for example individuals' right to be forgotten or retractions of previously given consent. By having only one source, the process to remove or add individuals becomes better manageable.

In regards of scaling, distributed data analysis can be seen as a model for growth. Linking additional centers would become easier, instead of facilitating additional storage resources for data duplication. However, this model for growth relies on standardized data; on a syntactic, semantic and ontological level as explained above. Hence, a proposal to setup such an infrastructure, both for local institutes and to link multiple institutes, is described in Chapter 5.

The use of this infrastructure, and the future perspective of secure data/information sharing is presented in Chapter 13. This chapter introduces a technical solution for privacy-preserving analytics on data sources/institutes which hold different information elements on the same individuals.

Objective and outline of this thesis

The main objective of this thesis is to setup and develop an infrastructure for distributed machine learning, and to use components of this infrastructure for clinical data analysis; mainly on treatment effects in rectal cancer patients.

To address this aim, this thesis is divided into four sections (Table 1). The first section (Chapters 2-5) describes the need and architecture for data standardization and (optionally) privacy preserved machine learning. Section two (Chapters 6-8) describes developments in data representation and semantic interoperability in radiation oncology. Section three (Chapters 9-12) reuses components of section one and two, to apply model development and validation for clinical purposes. Finally, section four gives a future perspective on the use of the infrastructure architecture outside radiation oncology (Chapter 13); and provides a general summary and discussion on the topics covered in this thesis (Chapter 14).

Section	Chapter	Main topic
Clinical data science architecture	Chapter 2	Introduction: Big Data in Radiation Therapy
	Chapter 3	Distributed Learning in rectal cancer: high level concept
	Chapter 4	Standardized data collection in rectal cancer
	Chapter 5	Technical architecture for distributed learning
Data Representation and Interoperability	Chapter 6	Radiation Oncology Ontology and Linked Data
	Chapter 7	Data extraction and standardization in medical imaging
	Chapter 8	Terminology linking in Radiation Oncology
Clinical data analysis	Chapter 9	Rectal cancer prediction model validation in routine clinical data
	Chapter 10	Rectal cancer prediction model validation: generalizability or transferability?
	Chapter 11	Pooled analysis of trial datasets in rectal cancer
	Chapter 12	Evaluation of automatic contouring of medical images
Future Perspective and Discussion	Chapter 13	Analysis infrastructure for complementary data sources
	Chapter 14	General summary and discussion

References

1. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, *et al.* Rapid-learning system for cancer care. *J Clin Oncol* 2010;28(27):4268–4274
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* 2013;10(1):27–40. doi:10.1038/nrclinonc.2012.196
3. May CP, Hasher L, Kane MJ. The role of interference in memory span. *Mem Cognit* 1999;27(5):759–767
4. Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A Roadmap for National Action on Clinical Decision Support. *J Am Med Inform Assoc JAMIA* 2007;14(2):141–145. doi:10.1197/jamia.M2334
5. Shortliffe EH, Cimino JJ, editors. *Biomedical Informatics*. London: Springer London; 2014. doi:10.1007/978-1-4471-4474-8
6. Shortliffe EH. Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases. *Proc Annu Symp Comput Appl Med Care* 1977:66–69
7. Abu-Hanna A, Lucas PJ. Prognostic models in medicine. *Methods Inf Med-Method Inf Med* 2001;40(1):1–5
8. Witten IH, Frank E. *Data mining: practical machine learning tools and techniques*. 2nd ed. Amsterdam ; Boston, MA: Morgan Kaufman; 2005
9. Iasonos A, Schrag D, Raj GV, Panageas KS. How To Build and Interpret a Nomogram for Cancer Prognosis. *J Clin Oncol* 2008;26(8):1364–1370. doi:10.1200/JCO.2007.12.9791
10. Hart A, Wyatt J. Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. *Med Inform (Lond)* 1990;15(3):229–236. doi:10.3109/14639239009025270
11. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. *J Clin Epidemiol* 2016;69:245–247. doi:10.1016/j.jclinepi.2015.04.005
12. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, *et al.* Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med* 2013;10(2):e1001381. doi:10.1371/ journal.pmed.1001381
13. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media; 2008
14. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15:361–387
15. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015;13(1). doi:10.1186/s12916-014-0241-z
16. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984;3(2):143–152
17. Wittes RE, Friedman MA. Accrual to Clinical Trials. *J Natl Cancer Inst* 1988;80:884–885
18. Gotay CC. Accrual to cancer clinical trials: directions from the research literature. *Soc Sci Med* 1991;33(5):569–577
19. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *Jama* 2004;291(22):2720–2726
20. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. *Br J Cancer* 2014;110(3):551–555. doi:10.1038/bjc.2013.725
21. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, *et al.* No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol* 2016;16. doi:10.1186/s12874-016-0267-3
22. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med* 1998;37(4–5):394
23. World Health Organization, editor. *International statistical classification of diseases and related health problems*. 10th revision, 2nd edition. Geneva: World Health Organization; 2004

24. Rothwell DJ. SNOMED-based knowledge representation. *Methods Inf Med* 1995;34(1–2):209–213
25. Cornet R, de Keizer N. Forty years of SNOMED: a literature review. *BMC Med Inform Decis Mak* 2008;8(Suppl 1):S2. doi:10.1186/1472-6947-8-S1-S2
26. Huff SM, Rocha RA, McDonald CJ, De Moor GJ, Fiers T, Bidgood Jr WD, *et al.* Development of the logical observation identifier names and codes (LOINC) vocabulary. *J Am Med Inform Assoc* 1998;5(3):276–292
27. Bakken S, Cimino JJ, Haskell R, Kukafka R, Matsumoto C, Chan GK, *et al.* Evaluation of the clinical LOINC (Logical Observation Identifiers, Names, and Codes) semantic structure as a terminology model for standardized assessment measures. *J Am Med Inform Assoc* 2000;7(6):529–538
28. Fieschi M. NCI Thesaurus: Using Science-Based Terminology to Integrate Cancer Research Results Sherri de Coronado", Margaret W. Haberb, Nicholas Sioutosc, Mark S. Tuttle'1, Lawrence W. Wright. *Medinfo 2004: Proceedings of the 11th World Congress on Medical Informatics*, Ios Pr Inc; 2004
29. Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. *Int J Semantic Web Inf Syst* 2009;5(3):1–22
30. Speicher S, Arwe J, Malhotra A. Linked Data Platform 1.0. https://www.w3.org/TR/ldp/ [accessed March 13, 2018]
31. van Soest J, Lustberg T, Marshall MS, Dekker A. Horizontal and vertical medical data federation: Linking clinical and DICOM data using Semantic Web technologies, *SWAT4LS*. Amsterdam: 2016
32. Cancer fact sheets: All cancers excluding non-melanoma skin cancer 2016
33. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000;100(1):57–70
34. Joiner M, Kogel A van der, editors. *Basic clinical radiobiology*. 4th ed. London: Hodder Arnold; 2009
35. Maas M, Nelemans PJ, Valentini V, Das P, Rödel C, Kuo LJ, *et al.* Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. *Lancet Oncol* 2010;11(9):835–844. doi:10.1016/S1470-2045(10)70172-8
36. Habr-Gama A, Gama-Rodrigues J, São Julião GP, Proscurshim I, Sabbagh C, Lynn PB, *et al.* Local Recurrence After Complete Clinical Response and Watch and Wait in Rectal Cancer After Neoadjuvant Chemoradiation: Impact of Salvage Therapy on Local Disease Control. *Int J Radiat Oncol* 2014;88(4):822–828. doi:10.1016/j.ijrobp.2013.12.012
37. Lee JH, Jang HS, Kim JG, Lee MA, Kim DY, Kim TH, *et al.* Prediction of pathologic staging with magnetic resonance imaging after preoperative chemoradiotherapy in rectal cancer: Pooled analysis of KROG 10-01 and 11-02. *Radiother Oncol* 2014;113(1):18–23. doi:10.1016/j.radonc.2014.08.016
38. Buijsen J, van Stiphout RG, Menheere PPCA, Lammering G, Lambin P. Blood biomarkers are helpful in the prediction of response to chemoradiation in rectal cancer: A prospective, hypothesis driven study on patients with locally advanced rectal cancer. *Radiother Oncol* 2014;111(2):237–242. doi:10.1016/j.radonc.2014.03.006
39. 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* 2011;98(1):126–133. doi:10.1016/j.radonc.2010.12.002
40. Janssen MHM, Öllers MC, Riedl RG, van den Bogaard J, Buijsen J, van Stiphout RGPM, *et al.* Accurate Prediction of Pathological Rectal Tumor Response after Two Weeks of Preoperative Radiochemotherapy Using 18F-Fluorodeoxyglucose-Positron Emission Tomography-Computed Tomography Imaging. *Int J Radiat Oncol* 2010;77(2):392–399. doi:10.1016/j.ijrobp.2009.04.030
41. Janssen MHM, Aerts HJWL, Buijsen J, Lambin P, Lammering G, Öllers MC. Repeated Positron Emission Tomography-Computed Tomography and Perfusion-Computed Tomography Imaging in Rectal Cancer: Fluorodeoxyglucose Uptake Corresponds With Tumor Perfusion. *Int J Radiat Oncol* 2012;82(2):849–855. doi:10.1016/j.ijrobp.2010.10.029
42. Appelt AL, Pløen J, Harling H, Jensen FS, Jensen LH, Jørgensen JC, *et al.* High-dose chemoradiotherapy and watchful waiting for distal rectal cancer: a prospective observational study. *Lancet Oncol* 2015;16(8):919–927
43. Burbach JPM, den Harder AM, Intven M, van Vulpen M, Verkooijen HM, Reerink O. Impact of radiotherapy boost on pathological complete response in patients with locally advanced rectal cancer: A systematic review and meta-analysis. *Radiother Oncol* 2014;113(1):1–9. doi:10.1016/j.radonc.2014.08.035

44. 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* 1995;31(5):1341–1346
45. US Department of Health and Human Services. Common terminology criteria for adverse events (CTCAE) version 4.0. *Natl Inst Health Natl Cancer Inst* 2009(04)
46. Basch E, Reeve BB, Mitchell SA, Clauser SB, Minasian LM, Dueck AC, *et al.* Development of the National Cancer Institute's Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). *JNCI J Natl Cancer Inst* 2014;106(9):dju244–dju244. doi:10.1093/jnci/dju244
47. Dueck AC, Mendoza TR, Mitchell SA, Reeve BB, Castro KM, Rogak LJ, *et al.* Validity and Reliability of the US National Cancer Institute's Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). *JAMA Oncol* 2015;1(8):1051. doi:10.1001/jamaoncol.2015.2639
48. Kalet IJ, Austin-Seymour MM. The Use of Medical Images in Planning and Delivery of Radiation Therapy. *J Am Med Inform Assoc* 1997;4(5):327–339
49. Law MYY, Liu B. DICOM-RT and Its Utilization in Radiation Therapy. *RadioGraphics* 2009;29(3):655–667. doi:10.1148/rg.293075172
50. Valentini V, Van Stiphout RG, 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* 2011;29(23):3163–3172
51. Valentini V, van Stiphout RG, Lammering G, Gambacorta MA, Barba MC, Bebenek M, *et al.* Selection of appropriate end-points (pCR vs 2yDFS) for tailoring treatments with prediction models in locally advanced rectal cancer. *Radiother Oncol* 2015
52. Roelofs E, Dekker A, Meldolesi E, van Stiphout RGPM, Valentini V, Lambin P. International data-sharing for radiotherapy research: An open-source based infrastructure for multicentric clinical data mining. *Radiother Oncol* 2014;110(2):370–374. doi:10.1016/j.radonc.2013.11.001
53. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, *et al.* 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol* 2013;109(1):159–164. doi:10.1016/j.radonc.2013.07.007
54. Lynn B, Stoller M. How to stop Google and Facebook from becoming even more powerful. *The Guardian* 2017. http://www.theguardian.com/commentisfree/2017/nov/02/facebook-google-monopoly-companies [accessed March 16, 2018]
55. Martijn M, Tokmetzis D, Medendorp H. *Je hebt wél iets te verbergen: over het levensbelang van privacy*. Amsterdam: de Correspondent; 2018

Chapter 2

Big Data in Radiation Therapy: challenges and opportunities

Authors
Tim Lustberg, Johan van Soest, Arthur Jochems, Timo Deist, Yvonka van Wijk, Sean Walsh, Philippe Lambin, Andre Dekker

Adapted from
The British Journal of Radiology, 2016 October; 90: 20160689
DOI: 10.1259/bjr.20160689

Abstract

Data collected and generated by radiation oncology can be classified by the Volume, Variety, Velocity and Veracity (4Vs) of Big Data because they are spread across different care providers and not easily shared owing to patient privacy protection. The magnitude of the 4Vs is substantial in oncology, especially owing to imaging modalities and unclear data definitions. To create useful models ideally all data of all care providers are understood and learned from; however, this presents challenges in the guise of poor data quality, patient privacy concerns, geographical spread, interoperability and large volume. In radiation oncology, there are many efforts to collect data for research and innovation purposes. Clinical trials are the gold standard when proving any hypothesis that directly affects the patient. Collecting data in registries with strict predefined rules is also a common approach to find answers. A third approach is to develop data stores that can be used by modern machine learning techniques to provide new insights or answer hypotheses. We believe all three approaches have their strengths and weaknesses, but they should all strive to create Findable, Accessible, Interoperable, Reusable (FAIR) data. To learn from these data, we need distributed learning techniques, sending machine learning algorithms to FAIR data stores around the world, learning from trial data, registries and routine clinical data rather than trying to centralize all data. To improve and personalize medicine, rapid learning platforms must be able to process FAIR "Big Data" to evaluate current clinical practice and to guide further innovation.

Primarily because of the ubiquity of imaging in oncology, as well as many other diagnostic and therapeutic procedures, cancer data are firmly in the realm of "Big Data" [1]. To make an estimate, in the past 10 years, approximately 140 million patients were diagnosed with cancer in about 100,000 hospitals globally. If one assumes a data volume (depending on the hospital) of 0.1–10 Gb of data per patient, the total volume of cancer patient data in the world is estimated to be 14–1400 petabyte of data. Specifically, data in radiation oncology can be classified as "Big Data" because: the use of data-intensive imaging modalities (Volume), the imaging archives are growing rapidly (Velocity), there is an increasing amount of imaging and diagnostic modalities available (Variety), and interpretation and quality differs between care providers (Veracity). With this deluge in data, it becomes increasingly hard to translate all these data into knowledge and subsequently leverage that knowledge to guide clinical decisions [2]. The radiation oncologist is overwhelmed with scientific literature, swiftly evolving treatment techniques and the exponentially increasing amount of clinical data [2]. To provide high-quality individualized treatments, radiation oncologists need help translating all these data into knowledge that supports decision-making in routine clinical practice [3]. Collecting these data provides its own set of challenges. The data are spread over care providers around the world, difficult to share while protecting patient privacy, non-interoperable and varying in quality.

The gold standard to assess the utility of innovations that directly affect patients is clinical trials. However, clinical trials only provide information about a select patient population, which often represents only a small percentage of the actual population. Also, clinical trials provide the radiation oncologist with little information when making clinical decisions for someone who does not (exactly) fit the trial population owing to age, comorbidities etc. On the other hand, clinical trials do provide high-quality reusable data owing to the clear definitions that are provided by trial protocols. Initiatives such as IBM Watson attempt to simplify accessing knowledge garnered from scientific literature for physicians. Patient characteristics provided by the physician are used to find and retrieve relevant publications (and possible other sources), which can aid the physician in making precision decisions for that particular patient [4].

To fill the gap of evidence between clinical trials and the common patient (i.e. one who does not fit trial inclusion criteria), data registries are being created around the world [5]. In general, the goal is to register a select set of parameters for all patients treated for a certain cancer. This results in a large patient population with high-quality data [5]. However, this requires great effort from care providers to collect these data, which limits the number of elements recorded, as someone must fill in a form, digital or paper, to provide the registries with the data specified in the registry protocol. There are some early initiatives to automatically provide the registries with the data they require by data mining the Oncology Information Systems [6]. In theory, this should work well for all structured data (e.g. the fractionation schema or age) but is challenging for data

which are usually recorded in free text (e.g. smoking behavior or comorbidities). Registries in general give insights in practice but are not designed to guide decisions for individual patients. ASCO CancerLinQ is the exception; it aims to create a "super" registry with a learning approach on routine healthcare data in medical oncology [7]. Cancer screening shows promising advancements in identifying patients at high risk using data mining techniques [8]; this particular example shows the power of centralizing data.

A different approach is to use routine clinical data from around the world to transform data into knowledge. As a proof of concept, the euroCAT project created data stores at several cancer centers, which can be accessed using web technologies. The local data are mined, pseudo anonymized, translated, mapped to standard concepts and made available to trusted partners in the network. The trusted partners do not have direct access to the data, but they can send a machine learning algorithm to the different data stores to learn from the data without sharing them (i.e. knowledge sharing, not data sharing). To demonstrate the power of this technique, an existing model was improved by combining the data of five centers (www.eurocat.info). However, a lot of time and effort is still required to access and utilize all data generated in clinical routine and translate them into knowledge. The end result of distributed learning is prediction models which can support physicians when making patient-specific choices (www.predictcancer.org). A different successful data mining was started by Public Health England. Data from all linear accelerators in England were collected automatically using the Radiotherapy Dataset tools (http://www.ncin.org.uk/collecting_and_using_data/rtds). This data set was analyzed to examine the variation in given treatments for different regions of the country.

An important topic when discussing collection of healthcare data is patient privacy. All three approaches handle privacy issues differently. Registries are usually hosted in a central location by professional societies or government-related entities, which are often authorized to collect identifiable patient data and securely store them while giving researchers an anonymized view of these data. Clinical trials work with informed consent and with pseudo anonymized data. Distributed systems are privacy-by-design systems, as they simply do not allow data to leave the site where they were collected.

It should be noted that there is a perception among healthcare providers that data must be kept in isolation due to privacy issues. However, there are existing solutions for these issues. The real barrier to learning from (Big) data in healthcare is that it requires willingness, resources and expertise [9].

Healthcare data are not yet "Big" enough to apply purely data-driven machine learning approaches, and clinical expertise is needed to create useful models that make sense to clinicians. Clinical trials, clinical registries and routine clinical data all provide unique evidence, which is currently utilized separately. Combining the three evidence sources into interoperable data stores makes them complementary to each other and will enable

healthcare to move forward. However, data quality (Volume, Variety, Velocity and Veracity) and sharing issues are hindering progress. To achieve "Big Data" in healthcare, the data must be Findable, Accessible, Interoperable and Reusable. The Findable, Accessible, Interoperable and Reusable Guiding Principles [10] can be applied to achieve good data management and stewardship, which will enable knowledge discovery and innovation. Eventually, when data-driven machine learning approaches have matured, it will provide a large knowledge base and clinical trials will only be used for a small subset of studies that requires to specific setup or a trial to prove (i.e. a new experimental treatment).

IBM stated that there are numerous ways to improve healthcare using their technology and provides a conclusion of utmost importance: "Information technology cannot drive change" [4]. "Big Data" can be a powerful tool to move healthcare forward, but healthcare providers need to invest resources to make this happen. Consequently, industry leaders in radiation oncology are already exploring this horizon market in anticipation of the opportunities and challenges that "Big Data" in healthcare represents, both in terms of efficiency and efficacy. Our experience is that the technical limitations of sharing data are minimal. Practical reasons are that healthcare providers are not willing, or do not have the resources and/or knowledge to share their data. Sharing data can influence the reputation of the healthcare provider because it allows their performance to be compared with others. Furthermore, these data can be used in research that a competing institute is working on as well, possibly creating unwanted competitors. By limiting the access to data to a machine and only sharing the model learned from the data, these issues are eliminated or largely negated. Despite the conflicting interests of healthcare providers, change may be driven by pressure from external institutions (such as government and health insurance companies) to ensure that the highest standard and most cost-effective care is delivered to the patient.

Many world leaders throughout history have referenced the seventeenth century poem by John Donne – "No man is an island" – when illustrating the need for collective responsibility and action towards a brighter future for all. This maxim rings as true in healthcare as it does in all other areas of life. We believe that utilizing patient privacy-preserving distributed machine learning to translate and combine all data sources into knowledge will enable healthcare to move to individualized, high-quality, affordable and safe cancer treatments, ensuring the sustainability of healthcare. This will also allow moving further towards participative medicine with customized patient decision aids.

References

1. Zarrouk M. Delivering Excellence in Patient Care with Ready Access to Clinical Data 2012. http://www.netapp.com/us/media/wp-7169.pdf [accessed July 28, 2016]
2. Oberije C, Nalbantov G, Dekker A, Boersma L, Borger J, Reymen B, *et al.* A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: A step toward individualized care and shared decision making. *Radiother Oncol* 2014;112(1):37–43. doi:10.1016/j.radonc.2014.04.012
3. Benedict SH, Hoffman K, Martel MK, Abernethy AP, Asher AL, Capala J, *et al.* Overview of the American Society for Radiation Oncology–National Institutes of Health–American Association of Physicists in Medicine Workshop 2015: Exploring Opportunities for Radiation Oncology in the Era of Big Data. *Int J Radiat Oncol* 2016;95(3):873–879. doi:10.1016/j.ijrobp.2016.03.006
4. Kohn MS, Sun J, Knoop S, Shabo A, Carmeli B, Sow D, *et al.* IBM's Health Analytics and Clinical Decision Support: *IMIA Yearb* 2014;9(1):154–162. doi:10.15265/IY-2014-0002
5. Bilimoria KY, Stewart AK, Winchester DP, Ko CY. The National Cancer Data Base: A Powerful Initiative to Improve Cancer Care in the United States. *Ann Surg Oncol* 2008;15(3):683–690. doi:10.1245/s10434-007-9747-3
6. Efstathiou JA, Nassif DS, McNutt TR, Bogardus CB, Bosch W, Carlin J, *et al.* Practice-Based Evidence to Evidence-Based Practice: Building the National Radiation Oncology Registry. *J Oncol Pract* 2013;9(3):e90–e95. doi:10.1200/JOP.2013.001003
7. Shah A, Stewart AK, Kolacevski A, Michels D, Miller R. Building a Rapid Learning Health Care System for Oncology: Why CancerLinQ Collects Identifiable Health Information to Achieve Its Vision. *J Clin Oncol* 2016;34(7):756–763. doi:10.1200/JCO.2015.65.0598
8. Liao LJ, Chou HL, Lo WC, Wang CT, Chou HW, Chen CD, *et al.* Initial outcomes of an integrated outpatient-based screening program for oral cancers. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2015;119(1):101–106. doi:10.1016/j.oooo.2014.09.020
9. Sullivan R, Peppercorn J, Sikora K, Zalcberg J, Meropol NJ, Amir E, *et al.* Delivering affordable cancer care in high-income countries. *Lancet Oncol* 2011;12(10):933–980. doi:10.1016/S1470-2045(11)70141-3
10. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. doi:10.1038/sdata.2016.18

Chapter 3

VATE: VAlidation of high TEchnology based on large database analysis by learning machine

Authors
Elisa Meldolesi, Johan van Soest, Anna Rita Alitto, Rosa Autorino, Nicola Dinapoli, Andre Dekker, Maria Antonietta Gambacorta, Roberto Gatta, Luca Tagliaferri, Andrea Damiani, Vincenzo Valentini

Adapted from
Future Medicine, 2014 October; volume 3 issue 5
DOI: 10.2217/crc.14.34

Abstract

The interaction between implementation of new technologies and different outcomes can allow a broad range of researches to be expanded. The purpose of this paper is to introduce the VAlidation of high TEchnology based on large database analysis by learning machine (VATE) project that aims to combine new technologies with outcomes related to rectal cancer in terms of tumor control and normal tissue sparing. Using automated computer bots and the knowledge for screening data it is possible to identify the factors that can mostly influence those outcomes. Population-based observational studies resulting from the linkage of different datasets will be conducted in order to develop predictive models that allow physicians to share decision with patients into a wider concept of tailored treatment.

Introduction

Over the past decade, remarkable advances in cancer care with the adoption of newest diagnostic and treatment technologies has created new challenges [1].

The use and role of medical imaging technologies in clinical oncology has greatly expanded from a primarily diagnostic tool to award a central role in the context of individualized medicine. Multiple imaging features involving descriptors of intensity distribution, spatial relationships between the various intensity levels, texture heterogeneity patterns, descriptors of shape and the relations of the tumor with the surrounding tissues have been analyzed for their relationship with treatment outcomes or gene expressions [2]. Beside these innovations, the concomitant research progress in pathology, biologic biomarkers (e.g., KRAS, BRAF, microsatellite instability, among others [3,4]), genomics and proteomics, justifying the growing trend toward 'individualized medicine' as key predictors for different treatment options (e.g., radiotherapy dose/fractionation and/or chemotherapy [CT]). Furthermore, tremendous advances in radiation therapy technology have allowed for remarkable precision in treatment delivery and for the realization of dose escalation with a concomitant decrease in treatment-related morbidity. 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 [5,6]. Accordingly, combinations of radiation therapy with targeted biological agents have proven growing efficacy and hold promise for future advances [7,8]. Therefore, several potential treatment options for each patient have to be considered instead of the inflexible 'one size fits all similar groups' approach. We have to move toward a 'shared decision making' process where an active interaction between doctors and patients results in the determination of the best therapeutic interventions.

In this context of progressive technologies and treatment innovation, a possible answer to the increasing necessity of individualized medicine is the development of predictive models allowing decision-making physicians to deliver tailored treatment; moving from an evidence-based treatment toward a personalized medicine concept (build on an evidence base) with an essential role for decision support systems (DSS). Predictive models based on individual patient features, complement existing consensus or guidelines and allow the transition from prescription by consensus to prescription by numbers.

The goal of this paper is to emphasize the influence of the interaction between the implementation of new technologies and different outcomes in radiation therapy, and to identify the factors that influence the outcomes (e.g., local recurrence, distant metastasis or overall survival). This identification will be performed using automated computer bots and the knowledge regarding screening and analysis of collected data. This approach can easily be reused and translated for a broad range of researches due to the usage of generalizable and flexible technologies in terms of data representation and semantics. The

purpose of the VAlidation of high TEchnology based on large database analysis by learning machine (VATE) project is to address this analysis to the field of rectal cancer.

The advances in radiation oncology and the progress in individualized medicine have created, over the past decade, new challenges.

Personalized medicine is defined by the National Cancer Institute 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 tumor to help diagnose, plan treatment, find out how well treatment is working, or make a prognosis" [9].

To date, the standard efforts in the medical field and inherently also in oncology is to con- sider the outcomes of randomized clinical trials (RCTs) as having the key role in the definition of clinical guidelines, protocols and research. Besides RCTs, population-based observational studies are progressively emerging as a complementary form of research, especially to ensure that the result of clinical trials translate into tangible benefits in the general population [10]. In the past decades, the excellent internal validity of RCTs led to substantial improvements in treatment and outcome of patients with cancer. However, patients participating in RCTs are 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 with a population included in RTCs [11]. Therefore, small benefits observed in highly selected trial patients are likely to disappear when the same treatments are applied in routine practice. Furthermore, some patient groups are underrepresented in RCTs, including the elderly, those with comorbidities [12,13] and patients from underrepresented ethnic and socioeconomic backgrounds [14–16]. Given the differences between patients recruited to trials and those seen in routine practice, increased toxicity might also be expected when the results of RCTs are applied to routine practice. Considering that, observational studies are essential to identify whether practice has changed appropriately, to document harms of therapy in a wider population, in patients of different age and with different comorbidities and to determine whether patients in routine practice are reaching the expected outcomes [17–19] with the expected toxicity. However, when multiples RCTs within one tumor type and comparing over decades and worldwide different treatment agents, classes and subpopulations, are observed in a pooled analysis [20–22], the best of both research regimens is reached: high amount of patient's variability with high data quality.

Although the findings of large randomized trials have addressed important questions, practical patient care issues remain that cannot be addressed by subgroup analyses of existing trials. Consensus guidelines reports try to help clinicians, but still large areas of controversies exist [23–25].

Moving into the era of individualized medicine, it is more and more important to develop systems that allow shared decision making by the physician and the patient and choose a tailored treatment. However, the development of DSS requires large heteroge-

neous datasets [1]. Features coming from many subdisciplines like diagnostic and clinical imaging, laboratory data, treatment outcome data, biologic environment, genomics and proteomics are routinely collected in the clinical practice and analyzed by innovative 'rapid learning' research techniques. This interaction between clinical information and biomedical informatics driven research allows for extraction of knowledge of the masses for the benefit of the individual [26]. From the mathematical point of view, the development of predictive models requires a large amount of data to provide sufficient statistical power to act as an efficient and reliable predictive tool. Furthermore, a secondary dataset is needed for validation of the models, preferably by external (from a different institution) datasets [27]. Only after external validation a prediction model can be implemented as an acceptable decision support tool. Therefore, development of prediction models brings up some stringent and challenging demands on the quality as well as the quantity of the data. Hence, the necessity to create large databases realized by sharing and combining multiple datasets which are often horizontally (patients scattered across institutes) and vertically partitioned (features on one patient in separate data silos).

Considering that data could be shared among different departments of a single hospital or among institutes on a regional, national and international level, integration of information is a big challenge for data-sharing initiatives.

Furthermore, for some tumor types that most benefit from a radiation therapy treatment like rectal cancer, the integration of radiotherapy information and clinical data could also be highly significant for the development and validation of multifactorial prediction models. The process based on large databases approach, requires setting a flexible strategy for data collection, data mining and outcome reporting, which is quite different from the fixed design of a prospective randomized controlled trial. It is necessary to collect data without knowing beforehand what the relevant features will be and it implies to use tools to add new variables to the large databases in an ad-hoc manner. From a structural point of view, it is, first of all, necessary to proceed with a standardization of any considered variable in order to universally define data and procedures allowing automatic and consistent upload from any exploitable data source. This standardization process is obtained through the creation of an ontology, a terminological system where all the information, related in this case to rectal cancer, are specified and organized in a well-defined data collection model. At the same time, tailored data mining techniques should be selected to exploit the heterogeneity of the collected data, to get evidence from differences while accounting for confounding factors. Finally, an appropriate DSS should be finalized to provide practical support to clinical choices.

In this context the VATE project is growing. The term VATE derives from the Latin 'vates', in other words, prophet, fortune teller. The project, related to rectal cancer, was conceived from the idea that by combining the implementation of new technologies with clinical outcomes, it could be possible to give an answer to multiple controversial questions in radiotherapy. The development of predictive measures represents an important practical tool to support clinical choices besides the already existing consensus guidelines.

All the steps needed to define and organize relevant data, to link it between different Institutions (Figure 1) and to develop a predictive model with specific reference to the VATE project (Figure 2), are described in the following paragraphs.

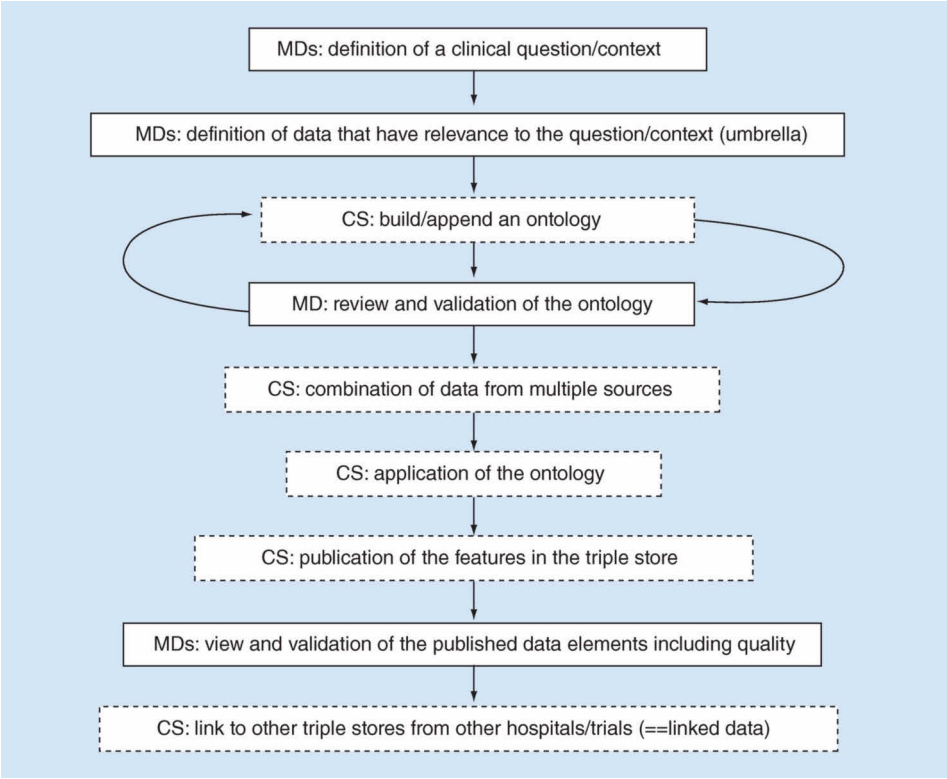


Figure 1 - Steps needed to link data. CS: Computer Science; MD: Medical Doctor

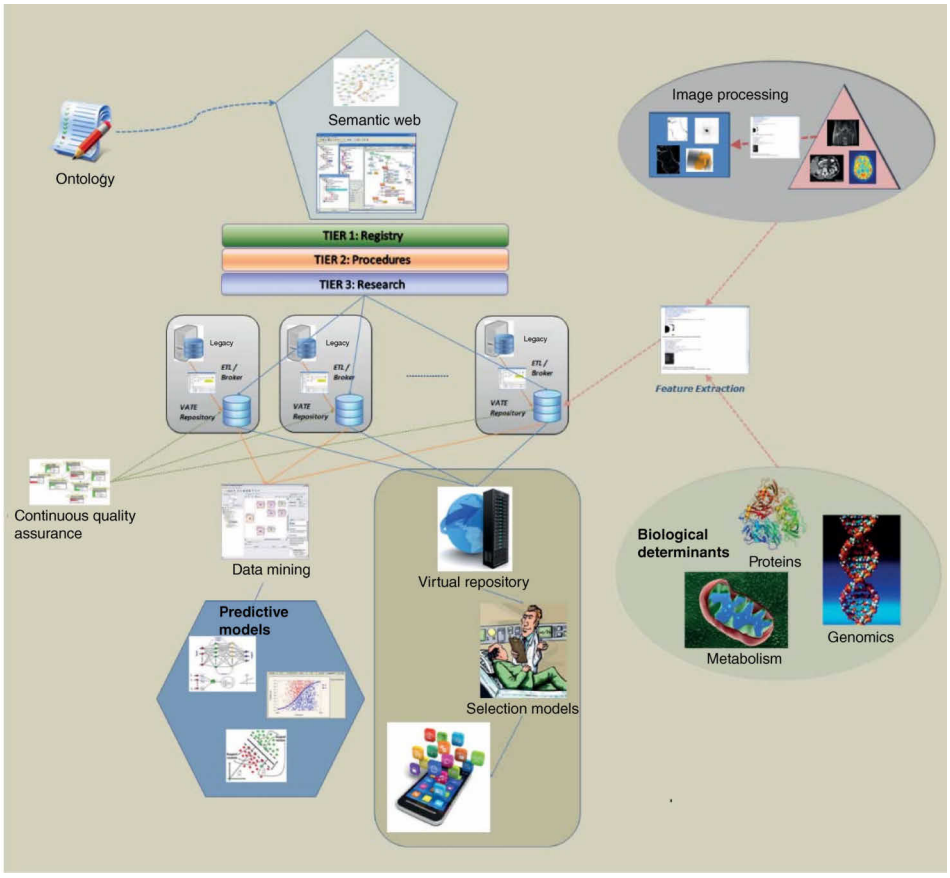


Figure 2 - VAlidation of high TEchnology project infrastructure

Standardized data collection & ontology

The process of standardization of the data collection will take advantage of an ontology. 'Ontology' is a compound word, composed of onto-, from the Greek ὄντος (òntos) which is the present participle of the verb εἰμί (eimi), in other words, 'to be, I am', and λόγία (lògia), in other words, 'science, study and theory'. An ontology formally represents knowledge as a set of concepts within a domain, and the relationships between those concepts. In practice, an ontology is (or can reuse) a classification system where each variable, in this case related to the domain of rectal cancer, can be represented using uniform and unambiguous definitions. Next to variable definitions, it can define relationships between variables. As these relationships can address variables defining space (e.g., relationships between institutional and standard terminologies) and time (e.g., versions of classifications), ontologies can enhance the understanding of a datasets. Eventually, better and unambiguous understanding leads to an approach where rectal cancer research data could be made available without differences in interpretation; for now, and the future. This kind of data collection model must be open also to the necessity of extending the number of collectable variables over time to be able to comprehend all the clinical, treatment and technical advances.

A team of rectal cancer specialists with previous experience on handling prospective trials, data collection and contributing to the oncology registries, selected features that they thought to have a relation to the outcome of rectal cancer patients. Features were organized according to three different tiers, with increasing granularity of data, to be able to answer different clinical questions. The first and most general level, the registry level, includes all the minimal information used for rudimentary epidemiological analysis only. The second level, the Procedures level, includes treatment information, related toxicities and the evaluation of outcomes in term of disease-free survival and acute and late toxicities. The last tier, the research level, includes clinical and imaging information used for in-depth, advanced research projects only. All the information needs to be converted to be semantically interoperable. Semantic interoperability (SI) is important in the VATE project, as it is the 'glue' between different datasets. SI means that data definitions and meaning of values are equal between different institutions and/or datasets. After making all datasets semantically interoperable, we are able to analyze and perform data mining on multiple datasets without compromising the outcome due to differences in definitions of variables between datasets. To achieve SI, we used generally available and accepted coding systems (e.g., National Cancer Institute [NCI] Thesaurus, Common Terminology Criteria for Adverse Events [CTCAE], and Systematized Nomenclature of Medicine – Clinical Terms [SNOMED]) to represent different datasets in a uniform manner. All needed concepts from the different coding systems are represented in our ontology. Relationships between concepts were determined after several architectural meetings and are still under active development.

In detail, the registry level includes general patients' information starting from the location and/or type of the tumor according the ICD-10 classification; the treating hospital, age at the time of first rectal cancer diagnosis, gender and overall survival information of the patient.

The procedure level defines that all tumor and treatment information with related toxicities is recorded. The definition of the treatments' intent is stated. General patient characteristics are more detailed than the registry level and include height, weight, BMI, ethnicity, performance status, comorbidities and results of blood and serum test. Tumor characteristics are recorded at the diagnosis, after neo- adjuvant treatment (if any) and after surgery. Tumor features include imaging characteristics used to stage the tumor such as its size, clinical TNM classification (cTNM; according to guidelines of American Joint Commission on Cancer) [28], Mesorectal Fascia status, presence/absence of extramesorectal lymph nodes, stage group [28], tumor location and methods of investigation, histological type (according to the International Classification of Disease for Oncology [ICD-O] classification [29]) and grade [29]. Intestinal margin, Tumor Regression Grading (TRG) score and residual tumor (R) are registered after surgery. In the procedure level all the rectal cancer treatment characteristics are recorded including (neo)adjuvant chemoradiotherapy information and surgery data. CT start and end date, CT agent, prescribed and delivered cycles and related toxicities are recorded. Radiotherapy (RT) on primary tumor (T), lymph nodes (N) and metastasis (M) information are described. Start RT date, total prescribed and delivered dose, fraction dose, fraction schedule, unexpected interruption days, RT technique, treatment position, immobilization devices, image-guided radiation therapy technique and frequency, intrafractional motion management devices, gross tumor volume (GTV) - Clinical Target Volume (CTV) and CTV-Pathological Target Volume (PTV) margins and related toxicities are recorded. Surgical intervention is described. Date of surgery, type of local and regional sur- gery, surgery technique and related toxicities are recorded. Type of stoma, stomas' placement and removal date, metastasis location – if any – and metastasectomy are registered. Toxicity is recorded according to the CTCAE v3.0 or 4.0 classifications [30,31].

Finally, several outcome features are reported at this level. Imaging used for follow-up, date of follow-up, date of local recurrence – if any – date and location of metastasis (M+) – if any – and biopsy confirmation for local recurrence and/or M+ are recorded.

The third and most detailed level, the research level, includes clinical and imaging information used for advanced research projects such as radiomics [32]. Concomitant medication, pre- existing general and rectal quality of life (QoL; EORTC QoL questionnaire [QLQ]-C30, EORTC QLQ-C29 and EQ-5D-5L) and tumor features including tumor markers are recorded. Diagnostic imaging (computed tomography, PET and/or magnetic resonance [MR]) for initial diagnosis and re-evaluation after neo- adjuvant treatment – if any – are uploaded. Radiotherapy treatment planning information is stated. Planning computed tomography imaging, RTSTRUCT, RTPLAN, RTDOSE, RT algorithm, dose to

OAR, QC in vivo dosimetry and QC patient specific pretreatment QA are recorded for radiotherapy treatment on T, N and M+. Follow-up imaging information and FUP-general and rectal QoL evaluation are reported.

In the VATE project, a well-structured and continuously updated ontology available to all data providers is able to reduce data entropy. Using an ad-hoc manual data input interfaces and/or automated remapping applications (brokers) the ontology can be used to enforce SIO so that data can be shared between applications.

Data warehouses and connected medical software have to 'understand' what the exchanged data means in order to process it. Machines however are only able to deal with data and generate information by using terminologies, definitions and ontologies. Therefore, data has to be converted to information by putting it in context, generating knowledge by making sense of the information and comparing all the knowledge in the field to find the best path to take in the decision [33].

Finally, the usage of data warehouses containing data following terminological standards leads, in the VATE project, to the development of important tools (e.g., predictive models, treatment planning integrated software) able to provide practical support to clinical choices. By integrating radiotherapy treatment information (using a specific VATE application in the planning workstation) with clinical, biological and laboratory data, it will be possible to choose the best treatment option, the most convenient dose distribution or fractionation to be used for each individual patient treated for rectal cancer (Figure 3).

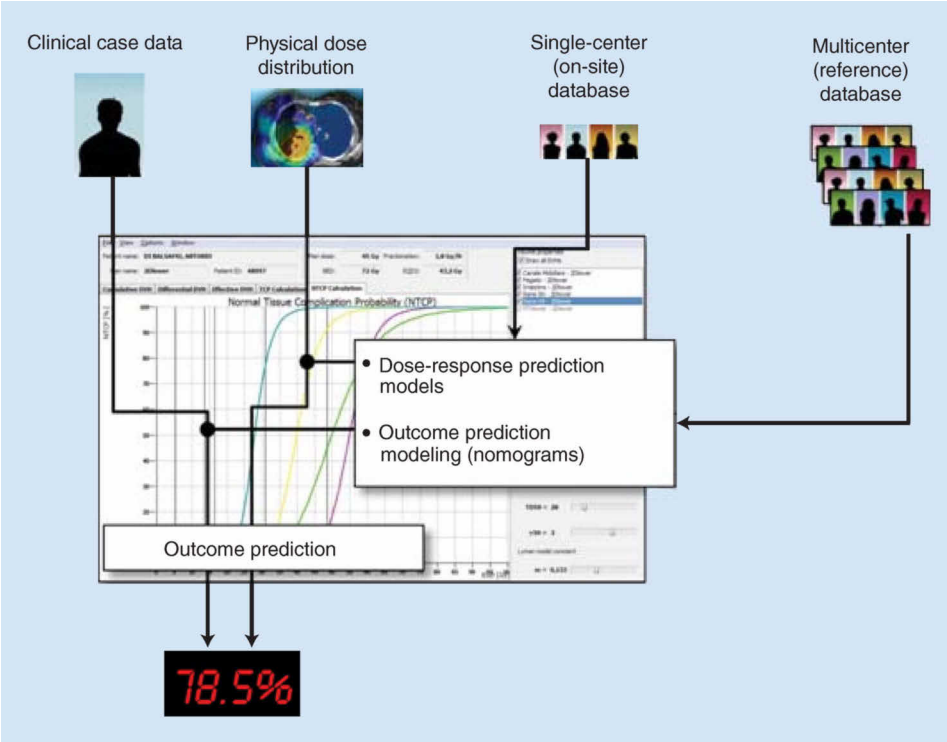


Figure 3 - Example of VAlidation of high TEchnology database connection

Sharing data: data warehouse & Semantic Web approaches

The shift toward better electronic capture of patient records is a great opportunity to increase the portability and accessibility of patient information, with the promise to make data capture and processing easier and more reliable than in an ordinary hand-written paper medical record process. However, major challenges still exist when data needs to be shared among different departments of a single hospital or among institutes on a regional, national and international level. Besides overcoming legal and political barriers, SI is a vital requirement [34]. In general, SI is the ability of any communicating entity (not only computers) to share unambiguous meaning. For a computer network, this can be described as the ability of exchanging data reducing at a minimum (and eliminating in an ideal case) the entropy introduced in the information flow.

There are two basic approaches to use the ontology to share data. A proven but cumber- some method is the data warehouse; a novel and more flexible method is to use Semantic Web technology.

An example of data sharing infrastructure based on a data warehousing architecture was implemented between the Policlinico Universitario Agostino Gemelli in Rome, Italy (Gemelli) and the MAASTRO Clinic in Maastricht, The Netherlands in mid-2010 [35]. This infrastructure (Figure 4) has been used to share rectal cancer data by pooling two existing datasets from both institutes in order to facilitate research projects such as the Thunder clinical trial [36] and 'knowledge engineering' research [37,38]. Four categories of information were analyzed: clinical (e.g., demographics, TNM-stage, date of diagnosis and histopathology, among others); outcome (e.g., survival, local control and toxicity); imaging (e.g., diagnostic and follow-up PET, computed tomography and MR imaging) and treatment data from radiotherapy planning and delivery (e.g., delineation, planning-CT, dose matrix, beam setup, prescribed dose and fractions, cone beam CTs, Orthogonal EPID imaging and delivered fractions) and from non-radiotherapy treatments (e.g., surgery and CT). SNOMED Clinical Terms were used as the ontology base to convert Italian to English standard terms creating a new database (DB), the Ontology DB, in which local terms were mapped to the SNOMED concept. Throughout a de-identification process (Key DB) and synchronization mechanism clinical and imaging data were linked together, and a research database and a research picture archiving and communication system were created. Finally, a virtual private network connection was established to retrieve data from the research DB and research picture archiving and communication system. Using data mining and machine learning techniques, predictive models were developed [39].

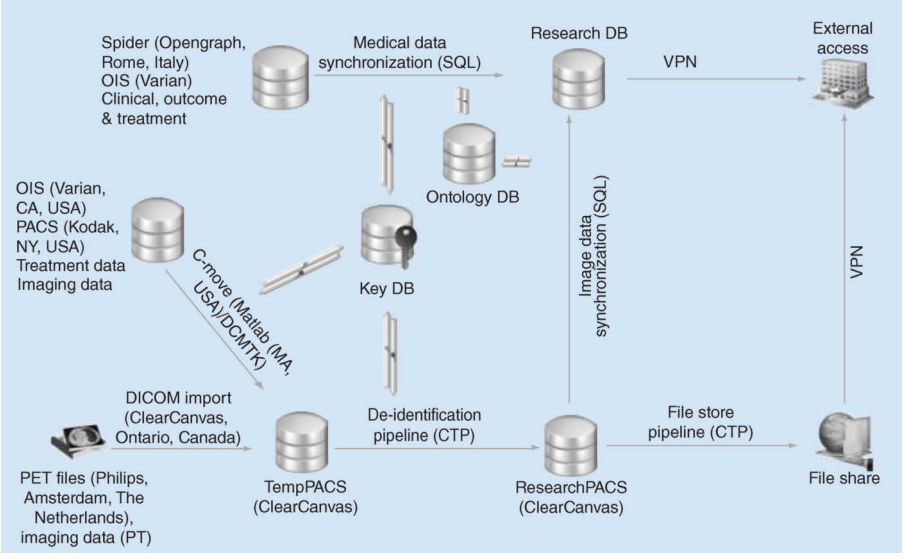


Figure 4 - Overview of data sources, flow and external access.
CTP: Clinical Trial Processor; DB: Database; DCMTK: Digital Imaging and Communications in Medicine Toolkit; DICOM: Digital Imaging and Communications in Medicine; OIS: Oncology Information System; PACS: Picture Archiving and communication System; SQL: Structured Query Language; VPN: Virtual Private Network.

The data warehouse approach results in a semantic interoperable dataset, which can be interpreted by applications. The main drawback of this solution is that a data warehouse is not very flexible and scalable for global implementation. If new data sources are added, specific data elements are added or if the data model is changed the interface to the data should be changed. Unfortunately, this flexibility is essential for research, as research directions and thus the questions may change. If the data warehouse is changed, either the application to the data no longer works or everyone needs to continuously upgrade and update their local data warehouse to keep in line. This is undoable on a global scale. Fortunately, the current World Wide Web has proven to be scalable and very flexible; the Semantic Web is an extension of the World Wide Web ideas to expose semantic interoperable data.

For the Semantic Web technology, data is fundamentally represented using the Resource Description Framework (RDF) [40]. Data is rep- resented in one table with three columns (subject, predicate and object). This RDF storage can be queried and modified using the SPARQL Protocol and RDF Query Language (SPARQL) [41]. In a typical setting, a SPARQL endpoint is placed on top of the RDF storage, which enables querying the RDF storage. SPARQL queries can query one or multiple (linked) triplets (one row of subject, predicate, object), based on the search criteria, and present the query results in an n-by-m table (where n is the number of variables queried, and m is the number of rows returned). By representing the ontology using RDF (and subsequently Resource Description Framework Schema and Web Ontology Language [OWL] implementations of RDF), the Semantic Web technology enables a semantically interoperable representation of data elements defined in the umbrella protocol (Table 1).

Table 1 - Examples of "semantic" triple representation

Subject	Predicate	Object example (URI)	Reference
Patient	Has been diagnosed with	Malignant neoplasm of rectum (http:// purl.bioontology.org/ontology/ICD10/C19)	ICD-10
Patient	Has biological sex	Male (http://ncicb.nci.nih.gov/xml/owl/ EVS/Thesaurus.owl#C20197) Female (http://ncicb.nci.nih.gov/	NCI Thesaurus
Disease	Has stage finding	T1 stage finding (http://ncicb.nci.nih.gov/ xml/owl/EVS/Thesaurus.owl#C48720)	NCI Thesaurus

ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems; NCI: National Cancer Institute; URI: Uniform resource identifier.
Reproduced with permission from [42]

Flexibility is an unavoidable requirement of the VATE project and finds in the Semantic Web technology the most efficient data sharing infrastructure. The power of the Semantic Web technology is the extremely flexible RDF representation but also the federated nature of the web where data and knowledge can reside anywhere and can be queried SPARQL [41]. As RDF contains a meta-structure (one table with three columns), it is independent of domain- specific structures, which enables flexibility and federation.

As an example, the CTCAE system (grades and definition) is available in a linked data representation, meaning that this knowledge can be reused without the need to re-represent this knowledge. The strategy taken in our protocol is to add SPARQL end points to local data sources (ranging from excel sheets or XML files to relational databases) as early as possible. This allows for internal (institutional) federated SPARQL queries to gather data for analysis. If this is not possible, an alternative strategy is to extract the data from the source (using 'classical' Extraction, Transformation, Loading [ETL] tools) and store the data in an RDF store. A further strategy is that all knowledge and data that can be made public (e.g., clinical trial data) or is already public can be accessed through a public SPARQL endpoint. Data that cannot be made public (e.g., routine clinical care data) will be available through private SPARQL end points.

Depending on the local regulations these private end points will be only available from inside the hospitals, or to authenticated and authorized users outside the hospital.

An example of a query using the Semantic Web platform [43], where values stored in the 'sexCode' and 'stageCode' fields follow a publicly encoded ontology (containing a thesaurus) established by the National Cancer Institute [44] has been created.

Finally, the possibility to share information between different institutions, as in the VATE project, allows getting data from the multicenter (reference) database. Therefore, the choice of the best treatment protocol for each single rectal cancer patient in terms of tumor control and normal tissue sparing could be simplified by the possibility to also use data stored on other radiotherapy treatment planning workstations (Figure 2).

Data quality issues & proposed mitigation

The infrastructure described above ensures that an application can get access to data at various institutes without the need for query rewriting or knowing the specifics of the underlying source systems. But while this infrastructure generalizes data extraction it does not guarantee the quality of the data from which application have to learn. To prevent a 'garbage in, garbage out' classification of the produced knowledge, the data quality must be known and if possible improved. The best way to improve the data is using the umbrella protocol to ensure that data is prospectively collected in the most consistent and best possible way. However, because resources are limited to be complete and correct in data collection and/or to reevaluate historical data, data quality issues will always be present. The expected data quality issues and the measures to mitigate them in the VATE project are described in Table 2.

Table 2 - VATE project: expected data quality issues and measures of mitigation

Problem	Example problem	Mitigation	Example mitigation
Completely missing data	Hospital A does not have a PET scanner, so all PET derived features are missing. Hospital B does have and use a PET scanner in all rectal cancer patients	Impute based on populations from other centers and what is known for the patient	Suppose a (probabilistic) relation between tumor size and SUV is learned from Hospital B, then the tumor size of Hospital A can be used to infer SUV max in Hospital A even if they don't have a PET scanner and are using the same scan protocols
Randomly missing data	Physician in Hospital A forgets to note the TNM stage of the patient	Because data are missing randomly, the percentage of missing data is generally low and samples are large, machine learning techniques will be unaffected by these errors	Do nothing
Biased data: continuous	A PET scanner is calibrated differently in hospital A than in hospital B, so the SUV values are different. Hospital B uses the SOP of the umbrella protocol	Assuming patients are similar a conversion is possibly between two distributions	Determine the distribution of SUV values in hospital A and B and derive a conversion function from SUVs in hospital A to hospital B
Biased data: categorical	CTCAE v3 was used, but after data X CTCAE v4 was used to score toxicities	Impute the new score from the old score, if possible	A (probabilistic/deterministic) conversion between the two CTC systems is possible
Random errors	In hospital A, a physician has noted an incorrect stage on an individual patient	Because errors are random, the percentage errors will be low and samples are large, the effect when using machine learning will be low	Do nothing
Biased missing data	In hospital A, severe toxicities are noted but mild toxicities are not. In hospital B toxicities are always noted		Compare occurrence of toxicities in hospital A with hospital B. Detect too low, unexplained mild toxicities in hospital A. Infer a probability of mild toxicity for patients of hospital A based on the distribution of hospital B

CTCAE: Common Terminology Criteria for Adverse Events; SOP: Standard Operating Procedure; SUV: Standardized uptake value; TNM: TNM Classification of Malignant Tumours

Many of the mitigation strategies shown in Table 2 assume that there is a hospital (or maybe a clinical trial) that does have high quality data which can be used to detect and correct any quality issues in the low-quality hospital. As the volume of data increases, these mitigations are expected to become easier and better.

Several strategies are employed to detect data quality problems. First and foremost is the domain knowledge from the physicians, physicists etc. that are involved in the project, they can often name many of the data quality issues that can be expected before they are observed. Another strategy is to include the treating hospital as a feature during learning. If the machine learning algorithm selects the treating hospital as predictive for outcome, then data quality is suspected.

In the VATE project the imputation method used to leverage the data from the high-quality source will be based on Bayesian statistics (e.g., Bayesian networks) as these have shown to be the best for imputation [37,45].

Statistical analysis

The strategy to collect data in a standard and consistent manner and to analyze them properly for decision support is named 'Umbrella protocol' [46]. The main difference between the development of a regular protocol and an umbrella protocol is the necessity, in the former to predefine the hypotheses investigated. In an Umbrella protocol, multiple hypotheses are considered but none are predefined.

The aim of the VATE project is not only to develop, validate and improve prediction models for overall survival or for acute and late radiation-induced side effects for rectal cancer patients, but also to use the prediction models to better inform patients on the risks (acute and late toxicity) and benefits (overall survival) of the treatment; going toward an individualized treatment. Furthermore, the realization of software (VATE software) usable during planning procedures, able to provide to clinicians reliable models to predict the outcome both for tumor control and the probability to develop complications, will provide the possibility to optimize internal guidelines for choosing treatment protocols and evaluate case by case the best technology, dose distribution or fractionation to be used for an individual patient's treatment plan. Finally, emerging radiation delivery techniques or other new diagnostic or treatment options can be investigated to assess their added value using the distributed machine learning (and data mining) strategy. Eventually, this strategy would allow us to compare the outcome of new treatment options clinically introduced, with the current standard (and accompanying retrospective data) in terms of radiation-induced toxicity, patient-rated symptoms, QoL and overall survival.

To learn from the collected data, a distributed learning approach will be taken in the VATE project. 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 algo-

rithm is split up into two parts: One master application which is installed on a central server (called the gate- way) and coordinates the learning between the hospitals. 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 gateway 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 [47].

The information exchanged between gateway and local nodes is limited to aggregated values (e.g., parameter weights, general statistics and coefficients) and contains no patient data. All traffic between gateway and local nodes is man- aged, monitored and audited by the infrastructure. An entire learning run is an iterative process that usually requires many cycles (~500) until the master determines that the learning process has been completed (see Figure 5).

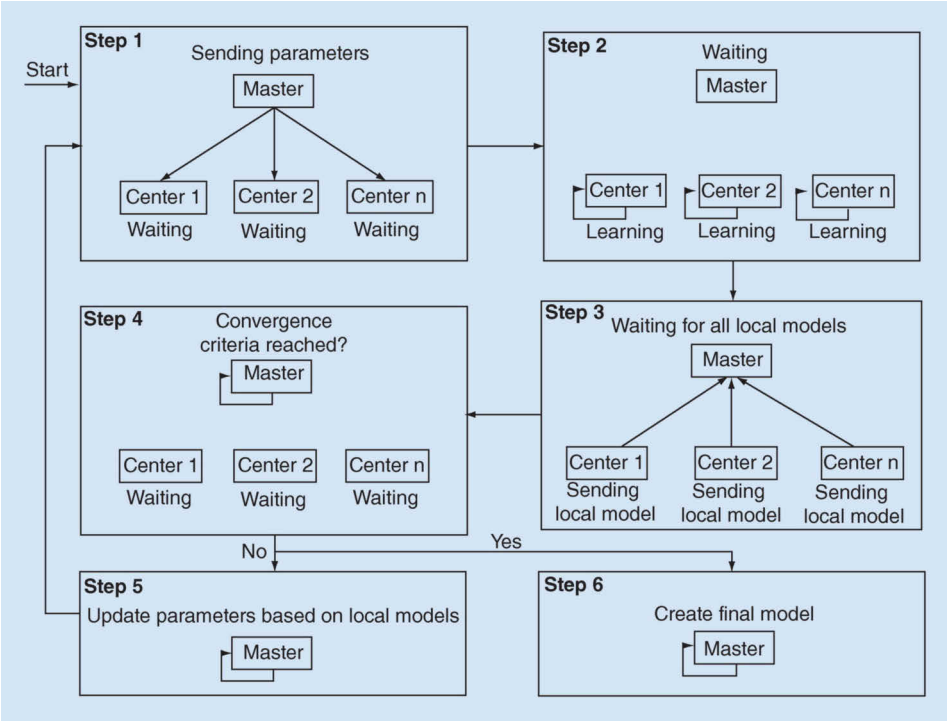


Figure 5 - Distributed machine learning flow

The width of the dataset used in the VATE project must be considered as the first pitfall to address correct statistical and data analysis tools to retrieve reliable results and models. Indeed, the generic definition of a 'large database' links itself with many technical issues that have to be correctly implemented in order to get an effective analysis of the data. Each model, anyhow defined, must undergo to a strict evaluation process mainly based on internal and external validation [48] in order to become a reliable tool to be used in clinical contexts.

The methodological process to learn, in other words, to go from data to useful decision sup- port is depicted in Figure 6 [1]. From the hypothesis, experts determine which features should be included in the learning effort. In the preprocessing step data quality is improved by imputation for missing data and outlier and bias detection and correction. Especially Bayesian approaches will be used as these are considered the best method [45]. 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 [37,49], Support Vector Machines [50] or Cox regression [22]. The final model can be presented to the end-user in a variety of ways such as nomograms [22] or via interactive websites such as PredictCancer [39].

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 of the receiver operating characteristic. The c-statistic is comparable to the area under the curve for dichotomous outcomes but can also be used for Cox regression analysis. A graphical assessment of calibration will be done by plotting the expected versus the observed outcome. In addition, the Hosmer–Lemeshow test will be used, to score the calibration.

The clinical value of the models will be assessed using decision curve analysis [51,52]. 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 cutoff point (for a clinical decision).

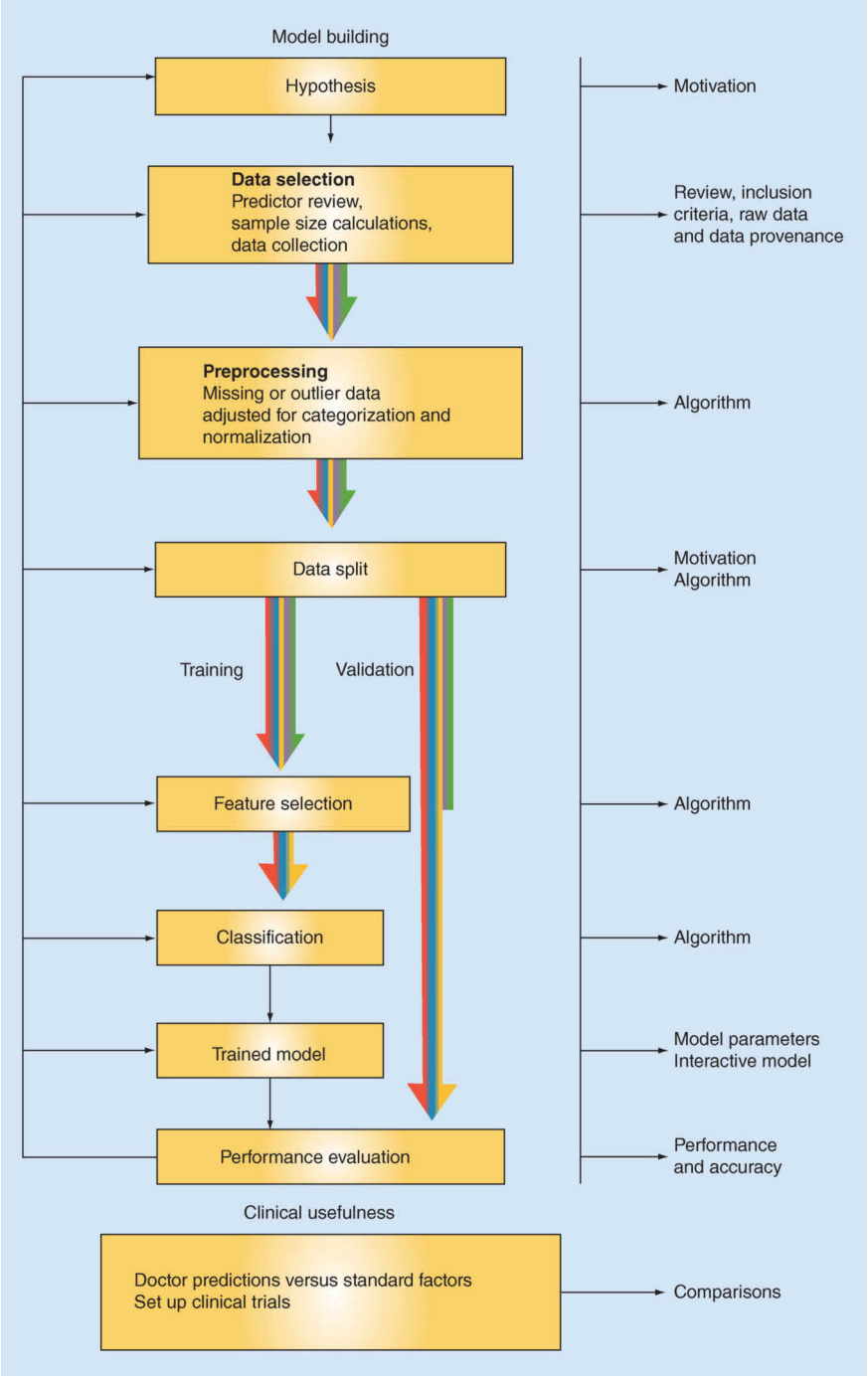


Figure 6 - Schematic overview of methodological process in clinical decision support system development. Reproduced with permission from [1]

Privacy protection of patients

To ensure the privacy of individual patients, all data will be pseudonymized or anonymized. Within each institute participating to the VATE project consortium, a local data manager will be responsible for pseudo/anonymizing the patient data. All data, originating from local repositories at the institute's site, is routed to a local repository. During transfer, all data will pass software to anonymize/pseudonymize all identifiable and traceable information. When anonymizing data, information (not needed for research) is removed and will not enter the local repository; while pseudonymization replaces data with other values, based on a list of original and replaced values. As a consequence, from the local SPARQL endpoint, which is queried by other research group members from inside or outside the institute during normal activity, it is impossible to reconcile clinical information to the relevant patient.

A second layer of privacy protection is that 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 values totally unrelated to specific patients are exchanged between nodes and their master [47,53,54]. Although this protocol does not require an intervention and the data is fully de- identified, internal review board ethics approval is recommended before implementing a local node.

Conclusion & future perspective

The interaction between the implementation of new technologies and different selected outcomes through the usage of automated computer bots and the knowledge for screening the collected data can allow a broad range of research to be expanded, due to the very generalizable and flexible technology utilized.

In the field of oncology, the possibility to predict the outcome for a certain patient in combination with a specific treatment with more accuracy will lead to more specific risk groups and thus better treatment decisions for individual patients. It will also stimulate research focused on specific risk groups, trying to find new treatment options or other combinations of treatment options for these subgroups. Furthermore, 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 resulting from large randomized trials that commonly include patients specifically selected to the research point. Complementary information coming from the combination of RCT and population-based observational studies will be able to find out new drugs and treatment strategies, and to review the efficiency of different approaches in large patients' populations.

The development of measures that allow the decision-making physician to deliver tailored treatment is particularly important as the oncology profession moves into the era of individualized medicine.

Clinicians are now facing two new challenges: one is the trend toward 'individualized medicine' trying to consider several potential options for each patient in place of inflexible 'one size fits all similar groups' approach. Second there is a move toward 'shared decision making', where doctors and patients actively discuss and decide on therapeutic interventions.

Therefore, predictive models, based on individual patient features, which complement existing consensus or guidelines allow us to move from prescription by consensus to prescription by numbers. The end result is a more patient specific selection from the treatment options menu and a possibility to share decisions with patients based on an objective evaluation of risks and benefits [55]. All of this comes with the 'overload' of information for physicians and patients. Therefore, it is important that a physician properly informs the patient on his or her options; patients might take the risk of specific inconveniences over a less aggressive treatment, or vice versa.

The development and validation of advanced predictive models based on large data collections is a necessary step toward the construction of a generation of software applications aimed to add a new dimension of knowledge during the planning phase. These tools will give clinicians a better view of the probability associated with the specific patient, both in terms of tumor control and of the development of complications. This will eventually lead to an optimization of the internal guidelines for choosing treatment protocols and evaluate case by case the best technology, dose distribution or fractionation to be used for the single, individualized treatment.

Finally, considering the important role that predictive models could play in the clinical practice, clinicians must be aware of the limits and confidence intervals of these prediction models. They need to be internally validated taking into account the quality of the collected data. An external validation of models is also essential to support general applicability of the prediction model. Therefore, structural collaboration between different groups is crucial to generate enough anonymized large databases from patients included or not in clinical trials.

Prediction of outcome, in order to choose the optimal treatment, is complicated because of the very complex, dynamic nature of cancer and organs at risk. Therefore, treatment can only become more personalized if accurate, science- based decision aids are developed, that can aid in clinical decision making in daily practice.

The VATE project is trying to expand on global scale the outcomes that have been reached with previous studies [22,50]. This preliminary work is daily being used via predictcancer.org by physicians and patients. The clinical impact of the development and validation of advanced predictive models will evolve, over the next few years, to the possibility of having decision supporting tools that can help doctors in the daily practice.

The possibility to have a better view of the probability associated with the specific patient, will eventually lead to an optimization of the internal guidelines for choosing treatment protocols and evaluate case by case the best treatment strategy. In case of radiotherapy it could also be helpful in choosing the best technology, dose distribution or fractionation to be used for the single, individualized treatment.

References

1. 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* 2013;10(1):27–40. doi:10.1038/nrclinonc.2012.196
2. 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* 2012;48(4):441–446. doi:10.1016/j.ejca.2011.11.036
3. Yamane LS, Scapulatempo-Neto C, Alvarenga L, Oliveira CZ, Berardinelli GN, Almodova E, *et al.* KRAS and BRAF mutations and MSI status in precursor lesions of colorectal cancer detected by colonoscopy. *Oncol Rep* 2014;32(4):1419–1426. doi:10.3892/or.2014.3338
4. Lin CC, Lin JK, Lin TC, Chen WS, Yang SH, Wang HS, *et al.* The prognostic role of microsatellite instability, codon-specific *KRAS* , and *BRAF* mutations in colon cancer: Prognostic Factors in Colon Cancer. *J Surg Oncol* 2014;110(4):451–457. doi:10.1002/jso.23675
5. Bentzen SM. Preventing or reducing late side effects of radiation therapy: radiobiology meets molecular pathology. *Nat Rev Cancer* 2006;6(9):702–713. doi:10.1038/nrc1950
6. Fowler JF. 21 years of Biologically Effective Dose. *Br J Radiol* 2010;83(991):554–568. doi:10.1259/bjr/31372149
7. 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* 2013;4(3):21
8. Bonner JA, 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* 2010;11(1):21–28. doi:10.1016/S1470-2045(09)70311-0
9. NCI Dictionary of Cancer Terms: Personalized Medicine
10. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. *Br J Cancer* 2014;110(3):551–555. doi:10.1038/bjc.2013.725
11. Zietman AL. Falsification, Fabrication, and Plagiarism: The Unholy Trinity of Scientific Writing. *Int J Radiat Oncol* 2013;87(2):225–227. doi:10.1016/j.ijrobp.2013.07.004
12. Tyldesley S, Zhang-Salomons J, Groome PA, Zhou S, Schulze K, Paszat LF, *et al.* Association between age and the utilization of radiotherapy in Ontario. *Int J Radiat Oncol* 2000;47(2):469–480. doi:10.1016/S0360-3016(00)00440-5
13. Faivre J, Lemmens VEPP, Quipourt V, Bouvier AM. Management and survival of colorectal cancer in the elderly in population-based studies. *Eur J Cancer* 2007;43(15):2279–2284. doi:10.1016/j.ejca.2007.08.008
14. Bach PB, Cramer LD, Warren JL, Begg CB. Racial Differences in the Treatment of Early-Stage Lung Cancer. *N Engl J Med* 1999;341(16):1198–1205. doi:10.1056/NEJM199910143411606
15. Boyd C, Zhang-Salomons JY, Groome PA, Mackillop WJ. Associations Between Community Income and Cancer Survival in Ontario, Canada, and the United States. *J Clin Oncol* 1999;17(7):2244–2244. doi:10.1200/JCO.1999.17.7.2244
16. 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* 2005;23(27):6639–6646. doi:10.1200/JCO.2005.12.633
17. 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* 2007;25(17):2383–2388. doi:10.1200/JCO.2006.09.1926
18. Booth CM. Evaluating Patient-Centered Outcomes in the Randomized Controlled Trial and Beyond: Informing the Future with Lessons from the Past. *Clin Cancer Res* 2010;16(24):5963–5971. doi:10.1158/1078-0432.CCR-10-1962

19. 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* 2012;30(21):2624–2634. doi:10.1200/JCO.2011.41.1140
20. Cheung WY, Shi Q, O'Connell M, Cassidy J, Blanke CD, Kerr DJ, *et al.* The Predictive and Prognostic Value of Sex in Early-Stage Colon Cancer: A Pooled Analysis of 33,345 Patients from the ACCENT Database. *Clin Colorectal Cancer* 2013;12(3):179–187. doi:10.1016/j.clcc.2013.04.004
21. Lieu CH, Renfro LA, de Gramont A, Meyers JP, Maughan TS, Seymour MT, *et al.* Association of Age With Survival in Patients With Metastatic Colorectal Cancer: Analysis From the ARCAD Clinical Trials Program. *J Clin Oncol* 2014;32(27):2975–2982. doi:10.1200/JCO.2013.54.9329
22. Valentini V, Van Stiphout RG, 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* 2011;29(23):3163–3172
23. Valentini V, Aristei C, Glimelius B, Minsky BD, Beets-Tan R, Borras JM, *et al.* Multidisciplinary Rectal Cancer Management: 2nd European Rectal Cancer Consensus Conference (EURECA-CC2). *Radiother Oncol* 2009;92(2):148–163. doi:10.1016/j.radonc.2009.06.027
24. Haustermans K, Debucquoy A, Lambrecht M. The ESTRO Breur Lecture 2010: Toward a tailored patient approach in rectal cancer. *Radiother Oncol* 2011;100(1):15–21. doi:10.1016/j.radonc.2011.05.024
25. van den Bogaard J, Janssen MHM, Janssens G, Buijsen J, Reniers B, Lambin P, *et al.* Residual metabolic tumor activity after chemo-radiotherapy is mainly located in initially high FDG uptake areas in rectal cancer. *Radiother Oncol* 2011;99(2):137–141. doi:10.1016/j.radonc.2011.04.004
26. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, *et al.* 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol* 2013;109(1):159–164. doi:10.1016/j.radonc.2013.07.007
27. Vickers AJ. Prediction models: revolutionary in principle, but do they do more good than harm? *J Clin Oncol* 2011;29(22):2951–2952
28. AJCC 7th Edition - Colon and Rectum Cancer Staging 2007. https://cancerstaging.org/references-tools/quickreferences/Documents/ColonMedium.pdf [accessed May 10, 2018]
29. Fritz AG, editor. *International classification of diseases for oncology: ICD-O*. Third edition, First revision. Geneva: World Health Organization; 2013
30. US Department of Health and Human Services. Common Terminology Criteria for Adverse Events V3.0 (CTCAE). *Natl Inst Health Natl Cancer Inst* (03)
31. US Department of Health and Human Services. Common terminology criteria for adverse events (CTCAE) version 4.0. *Natl Inst Health Natl Cancer Inst* 2009(04)
32. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, *et al.* Radiomics: the process and the challenges. *Magn Reson Imaging* 2012;30(9):1234–1248. doi:10.1016/j.mri.2012.06.010
33. Valentini V, Schmoll HJ, Velde CJH van de, editors. *Multidisciplinary Management of Rectal Cancer: Questions and Answers*. Berlin Heidelberg: Springer-Verlag; 2012
34. Lewalle P, Rodrigues JM, Zanstra P, Ustun B, Kalra D, SURJAN G, *et al.* A Deployment and Research Roadmap for Semantic Interoperability: the EU SemanticHEALTH project:6
35. Roelofs E, Dekker A, Meldolesi E, van Stiphout RGPM, Valentini V, Lambin P. International data-sharing for radiotherapy research: An open-source based infrastructure for multicentric clinical data mining. *Radiother Oncol* 2014;110(2):370–374. doi:10.1016/j.radonc.2013.11.001
36. Maastricht Radiation Oncology. Validation of a Predictive Model After Complete Response in Rectal Cancer (Thunder). Identifier: NCT00969657. *ClinicalTrialsGov*. https://clinicaltrials.gov/ct2/show/NCT00969657 [accessed May 11, 2018]
37. 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: Bayesian network for survival prediction in lung cancer. *Med Phys* 2010;37(4):1401–1407. doi:10.1118/1.3352709

38. Dehing-Oberije C, Yu S, De Ruysscher D, Meersschout S, Van Beek K, Lievens Y, *et al.* Development and External Validation of Prognostic Model for 2-Year Survival of Non–Small-Cell Lung Cancer Patients Treated With Chemoradiotherapy. *Int J Radiat Oncol* 2009;74(2):355–362. doi:10.1016/j.ijrobp.2008.08.052
39. MAASTRO Clinic. Cancer Prediction Models by PredictCancer.org. http://predictcancer.org/Main.php?page=Home [accessed May 11, 2018]
40. Brickley D, R.V. G. RDF Schema 1.1. *W3C Recomm* 2014
41. Ariane AK, Rémy C, Giovanni M, Christel D, Jean C, Marie-Christine J. An Ontological Approach for the Exploitation of Clinical Data. *Stud Health Technol Inform* 2013:142–146. doi:10.3233/978-1-61499-289-9-142
42. 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* 2014;112(1):59–62
43. Semantic Web example query and result. https://tinyurl.com/VateSPARQL [accessed May 16, 2013]
44. NCI Center for Biomedical Informatics and Information Technology | National Cancer Institute. https://cbiit.cancer.gov/ [accessed May 11, 2018]
45. Luta G, Ford MB, Bondy M, Shields PG, Stamey JD. Bayesian sensitivity analysis methods to evaluate bias due to misclassification and missing data using informative priors and external validation data. *Cancer Epidemiol* 2013;37(2):121–126. doi:10.1016/j.canep.2012.11.006
46. Marsolo K. Approaches to Facilitate Institutional Review Board Approval of Multicenter Research Studies: *Med Care* 2012;50:S77–S81. doi:10.1097/MLR.0b013e31825a76eb
47. Boyd S. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found Trends® Mach Learn* 2010;3(1):1–122. doi:10.1561/2200000016
48. Harrell FE. Resampling, Validating, Describing, and Simplifying the Model. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*, New York, NY: Springer New York; 2001. doi:10.1007/978-1-4757-3462-1_5
49. 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* 2011;56(6):1635–1651. doi:10.1088/0031-9155/56/6/008
50. 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* 2011;98(1):126–133. doi:10.1016/j.radonc.2010.12.002
51. 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* 2008;8(1). doi:10.1186/1472-6947-8-53
52. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making* 2006;26(6):565–574. doi:10.1177/0272989X06295361
53. 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* 2012;19(5):758–764. doi:10.1136/amiajnl-2012-000862
54. Liu K, Kargupta H, Ryan J. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Trans Knowl Data Eng* 2006;18(1):92–106. doi:10.1109/TKDE.2006.14
55. Valentini V, Lambin P, Myerson RJ. Is it time for tailored treatment of rectal cancer? From prescribing by consensus to prescribing by numbers. *Radiother Oncol* 2012;102(1):1–3. doi:10.1016/j.radonc.2011.12.001

Chapter 4

An umbrella protocol for standardized data collection in rectal cancer: A prospective uniform naming and procedure convention to support personalized medicine

Authors
Elisa Meldolesi, Johan van Soest, Nicola Dinapoli, Andre Dekker, Andrea Damiani, Maria Antonietta Gambacorta, Vincenzo Valentini

Adapted from
Radiotherapy and Oncology, July 2014, Volume 112, Issue 1, pages 59 – 62
DOI: 10.1016/j.radonc.2014.04.008

Abstract

Predictive models allow treating physicians to deliver tailored treatment moving from prescription by consensus to prescription by numbers. The main features of an umbrella protocol for standardizing data and procedures to create a consistent dataset useful to obtain a trustful analysis for a Decision Support System for rectal cancer are reported.

Introduction

Over the past decade, we have witnessed remarkable advances in cancer care, with many new diagnostic methods and treatment modalities becoming available [1,2], including advances in radiation oncology. To date, in the medical field and inherently also in oncology, clinical practice is based on evidence-based guidelines and protocols as results of the outcome of randomized clinical trials [3–6] in which great effort is taken to limit the variability in the study population so that any benefit of one arm over the other can be detected. As a result, the presented evidence is often valid for only a subgroup of patients and trial results are quickly outdated due to advances in technology and often of insufficient quality [7].

The development of measures that allow treating physicians to deliver tailored treatment, leads the transformation from a population based treatment (where "one size fits all") toward a personalized medicine concept with an essential role of Decision Support Systems [8,9]. Unlike evidence from trials, these systems require large heterogeneous datasets not only to create a prediction model with a sufficient statistical power but also to validate it, preferably by external datasets [10]. Only after external validation a prediction model can be implemented as an acceptable decision support tool. Therefore, development of prediction models brings up some stringent and challenging demands on the quality as well as the quantity of the data. Hence, the need exists to create large databases realized by sharing and combining multiple datasets which are often horizontally (patients scattered across institutes) and vertically (features on one patient in separate data silos) partitioned. Finding a way to aggregate this routinely collected patient data and applying innovative "rapid-learning" research techniques allows extraction of knowledge of the masses for the benefit of the individual [11].

The implementation of a program to develop and validate multi-factorial prediction models for different treatment outcomes based on large datasets requires setting a flexible strategy for data collection, data mining and outcome reporting, which is quite different from the fixed design of a prospective randomized controlled trial. It is necessary to collect data without knowing beforehand what the relevant features will be. Therefore, it implies the need of flexible data storage applications which support ad-hoc addition of relevant features. It is also fundamental to codify each variable in a standardized way to allow automatic and consistent upload from any exploitable data source. Furthermore, tailored data mining techniques should be selected to exploit the heterogeneity of the collected data, to get evidence from patient differences in characteristics and outcomes, while accounting for confounding factors. Finally, appropriate Decision Support System should be finalized, containing one or more prediction models based on the data mining outcomes, to provide a practical support to clinical choices which often require a balanced decision between conflicting outcomes. The strategy to collect data in a standard and consistent manner and to analyze them properly for decision support is called an 'Umbrella protocol' [12]. In this manuscript an umbrella protocol for rectal cancer is reported.

Materials and methods

Standardized data collection and ontology

The standardized data collection (SDC) procedure will take advantage of an ontology that formally represents knowledge as a set of concepts within a domain, and the relationships between those concepts. When defined correctly, ontologies also define the semantics of different concepts. Ontologies can enclose (or reuse) a terminological system [13], which enable computers to ''understand'' the meaning of data (semantics) [14]. Therefore, ontologies, comprising relations of concepts and a terminological system defining concepts, can convert data into information. Furthermore, as defined by de Keizer et al. [13], ontologies can be used by different partners/institutes to define consensus regarding concepts (and the structure of data) within a specific domain. Within the VATE project (VAlidation of High TEchnology based on large database analysis by learning machine) between the Policlinico Universitario Agostino Gemelli in Rome, Italy and the MAASTRO Clinic in Maastricht, the Netherlands, at the time of this publication, a team of rectal cancer specialists selected more than 200 features which could have a relation to the outcome of rectal cancer patients. These variables were organized according three different tiers, with an increasing number of features, to obtain a well-defined data collection model. The first and most general level, the Registry level, includes the minimal information (age, gender, ethnicity e.g.) used for epidemiological analysis only. The second level, the Procedures level, includes treatment information and related toxicities, and the evaluation of outcome in terms of disease free survival and acute and late toxicities. The highest level, the Research level, includes clinical and imaging information used for in-depth, advanced research projects only. The development of a three tiers Ontology arose from the necessity to give all the willing research Institutions (e.g. Cancer Registry, Research Units etc.) the possibility to contribute to the research according to their available data corresponding level. In our ontology (http://www.github.com/RadiationOncologyOntology/ROO), which is part of the umbrella protocol, we preferably used concepts from existing and mature terminological systems (e.g. NCI Thesaurus, CTCAE, SNOMED-CT) and created semantic links between those concepts (Table 1). This initial ontology design is different from the CDISC SHARE initiative [15], where the primary goal is to create a repository containing the similarities between concepts from different terminological systems. Our approach is more comparable to the CDISC BRIDG initiative [16], where existing conceptual UML models are represented in ontologies for clinical research. However, in our situation, there is no widespread available UML model, specific for clinical routine radiation oncology.

A well-structured and continuously updated ontology available to all data providers is able to reduce data entropy. There are two basic approaches to use the ontology to share data. A proven but cumbersome method is the datawarehouse, where data are stored in a relational database structure. A relatively novel and more flexible method is

to use Semantic Web technology, where anyone can say anything about anything (e.g. anyone can contribute/publish data about any topic).

Table 1 - Examples of "semantic" triple representation

Subject	Predicate	Object example (URI)	Reference
Patient	Has been diagnosed with	Malignant neoplasm of rectum (http://purl.bioontology.org/ontology/ICD10/C19)	ICD-10
Patient	Has biological sex	Male (http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C20197) Female (http://ncicb.nci.nih.gov/	NCI Thesaurus
Disease	Has stage finding	T1 stage finding (http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C48720)	NCI Thesaurus

ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems; NCI: National Cancer Institute; URI: Uniform resource identifier.

Sharing data: datawarehouse and Semantic Web approaches

The shift toward better electronic capture of patient records is a great opportunity to increase the portability and accessibility of patient information, however major challenges still exist when data need to be shared between institutes on a regional, national, and international level. Besides overcoming legal and political barriers, semantic interoperability (SIO) is a vital requirement. In general, SIO is the ability of any communicating entity (not only computers) to share unambiguous meaning. For two computers, this is the ability to exchange data without loss of meaning by the receiving system. An example of data sharing infrastructure based on datawarehousing architecture was implemented between the Policlinico Universitario Agostino Gemelli in Rome, Italy (Gemelli) and MAASTRO Clinic in Maastricht, the Netherlands (MAASTRO) in mid-2010 [17]. The main drawback of this approach is that a datawarehouse is not very flexible and scalable in regard to changing ontologies or federation of multiple datawarehouses. If the datawarehouse structure is changed, the application querying the warehouse needs to be updated. When using a federation of warehouses, all warehouses need to update their structure, resulting in an impairment of the ad-hoc update process. Flexibility is essential for research as the research direction and thus the questions (e.g. regarding data) may change. The Semantic Web facilitates this flexibility by using the Resource Description Framework (RDF) as the basis to store data. In short, RDF represents data in the form of triples (subject-predicate-object, e.g. Patient1-hasSex-Female). The definitions of subjects, predicates objects are defined in an ontology. 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 [18]. As an example, the CTCAE system [19] is available on the Semantic Web, meaning that this knowledge can be re-used without the need to re-represent this knowledge and struc-

ture in a local database. Within the biological domain of life sciences, semantic web approaches are important drivers behind knowledge management [20] and to integrate knowledge and data [21], but in the medical domain Semantic Web approaches have only recently been shown to have value, like the secondary use of electronic health records for colorectal cancer screening [22] and to mine for drug-drug interactions [23].

The Semantic Web strategy taken in our umbrella protocol is to add local SPARQL endpoints to local data resources (ranging from excel sheets to relational databases to XML files). This enables internal federated SPARQL queries to combine data from multiple local data resources. If creating local endpoints is not directly possible (e.g. due to the data format), an alternative strategy is to extract the data from the local resource (using 'classical' ETL tools) and store it in a (local) triple-store. Afterward, internal federated SPARQL queries can create a public SPARQL endpoint to publish datasets (e.g. clinical trial datasets). Data that cannot be made public (e.g. routine clinical care data) will be available through private SPARQL endpoints. Depending on the local regulations these private endpoints will be available from inside the hospitals or to authorized and authenticated users outside the hospital.

The following link provides an example of a query using the semantic web platform (http://tinyurl.com/VateSPARQL), where values stored in the 'sexCode' and 'stageCode' fields are defined in the NCI Thesaurus.

Privacy protection patients

To ensure the privacy of individual patients, all data will be anonymized (or pseudonymized). All data, originating from local repositories at the institute's site, are routed to a local repository in non-anonymized form, while unique identifiers that could point directly or indirectly to individual patients are removed or pseudonymized. As a consequence, from the local SPARQL endpoint, which is queried by other research group members from in- or outside the institute during normal activity, it is impossible to reconcile data in the SPARQL endpoint to the actual patient.

A second layer of privacy protection is that data do not need to leave the institute in the process of machine learning a predictive model by using information from one or more institutes. This principle is called distributed learning, and only requires statistical indexes totally unrelated to specific patients to be exchanged between learning bots and their master [24–26]. An example of such a distributed system is a parallel system were local applications learn a model from local data. These local models are brought to a central location which combines them into one consensus model. This is then sent out again to all local applications which adjust the model based on the local data. The new local models are sent again to the central location and a new iteration is run and so on. The adjustment which the local applications can make is more restricted after every iteration thus driving consensus between the centers. Once a pre-defined convergence criterion is met the

learning ends and the final consensus model is published to all participating centers [27]. An efficient method to do is the "alternating direction method of multipliers" [25].

Results

An 'umbrella protocol' strategy has been used to standardize data and procedures to create a consistent dataset useful to obtain a trustful analysis for the Decision Support System. An ontology, a well-defined data collection model, able to collect, standardize and organize features related with rectal cancer patients, has been created. The three data-storing levels have been used to classify all the information to easily address the query depending on the different analyses requested (epidemiological, toxicity, outcomes e.g.).

Discussion

The possibility to predict the outcome for a certain patient in combination with a specific treatment with more accuracy, will lead to better identification of risk groups and thus better treatment decisions in individual patients, but it will also stimulate research focused on specific risk groups which try to find new treatment options or other combinations of treatment options for these subgroups. These treatments 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 resulting from large randomized trials that only take into account relative indiscriminate features such as tumor stage and the physical condition of a patient. These guidelines are developed for groups of patients and thus lead to over-treatment in some patients and inadequate therapy in others, resulting in major expenses for individuals and society. Therefore, predictive models, based on individual patient features, which complement existing consensus or guidelines allow us to move from prescription by consensus to prescription by numbers. To accomplish this goal, we also need an infrastructure to support the retrieval of patient features, with respect to their semantics. We decided to use the Semantic Web for data representation and retrieval. Although we are not the first to use Semantic Web technology in the health sciences [28–30], we are within the domain of radiation oncology. The resulting predictive models, based on patient features, enable a more patient specific selection from the treatment options menu and a possibility to share decisions with patients based on an objective evaluation of risks and benefits [9]. Finally, considering the important role that predictive models could play in the clinical practice, clinicians must be aware of the limits of these prediction models. They need to be internally validated taking into account the quality of the collected data. An external validation of models is also essential to support general applicability of the prediction model. There-

fore, structural collaboration between different groups is crucial to generate enough anonymized large databases from patients included or not in clinical trials.

Ultimately this work should lead to clinically relevant Decision Support Systems (DSS) for rectal cancer. Underlying the DSS are the data and prediction models that were learned from the data. These should be presented completely transparent to the user so that model provenance (from which data and how the model was learned) and the match between the characteristics of the learning data sample and the new patient can be evaluated. Not only will this transparency increase trust but will also allow users to contribute data where they feel the proof database is sparse.

We reported the main features of an umbrella protocol for sharing data with the end goal to develop a Decision Support System for rectal cancer.

References

1. 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* 2013;10(1):27–40. doi:10.1038/nrclinonc.2012.196
2. Beets-Tan RGH, Beets GL. MRI for assessing and predicting response to neoadjuvant treatment in rectal cancer. *Nat Rev Gastroenterol Hepatol* 2014;11(8):480–488. doi:10.1038/nrgastro.2014.41
3. Glimelius B. Multidisciplinary treatment of patients with rectal cancer: Development during the past decades and plans for the future. *Ups J Med Sci* 2012;117(2):225–236. doi:10.3109/03009734.2012.658974
4. Fichera A, Allaix ME. Paradigm-Shifting New Evidence for Treatment of Rectal Cancer. *J Gastrointest Surg* 2014;18(2):391–397. doi:10.1007/s11605-013-2297-z
5. Quirke P, West NP, Nagtegaal ID. EURECCA consensus conference highlights about colorectal cancer clinical management: the pathologists expert review. *Virchows Arch* 2014;464(2):129–134. doi:10.1007/s00428-013-1534-x
6. Tudyka V, Blomqvist L, Beets-Tan RGH, Boelens PG, Valentini V, van de Velde CJ, *et al.* EURECCA consensus conference highlights about colon & rectal cancer multidisciplinary management: The radiology experts review. *Eur J Surg Oncol EJSO* 2014;40(4):469–475. doi:10.1016/j.ejso.2013.10.029
7. Zietman AL. Falsification, Fabrication, and Plagiarism: The Unholy Trinity of Scientific Writing. *Int J Radiat Oncol* 2013;87(2):225–227. doi:10.1016/j.ijrobp.2013.07.004
8. Wibe A, Law WL, Fazio V, Delaney CP. Tailored rectal cancer treatment – a time for implementing contemporary prognostic factors? *Colorectal Dis* 15(11):1333–1342. doi:10.1111/codi.12317
9. Valentini V, Lambin P, Myerson RJ. Is it time for tailored treatment of rectal cancer? From prescribing by consensus to prescribing by numbers. *Radiother Oncol* 2012;102(1):1–3. doi:10.1016/j.radonc.2011.12.001
10. Vickers AJ. Prediction models: revolutionary in principle, but do they do more good than harm? *J Clin Oncol* 2011;29(22):2951–2952
11. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, *et al.* 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol* 2013;109(1):159–164. doi:10.1016/j.radonc.2013.07.007
12. Marsolo K. Approaches to Facilitate Institutional Review Board Approval of Multicenter Research Studies: *Med Care* 2012;50:S77–S81. doi:10.1097/MLR.0b013e31825a76eb
13. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems I: Terminology and typology. *Methods Inf Med* 2000;39(1):16–21
14. Valentini V, Schmoll HJ, Velde CJH van de, editors. *Multidisciplinary Management of Rectal Cancer: Questions and Answers*. Berlin Heidelberg: Springer-Verlag; 2012
15. crodgers. CDISC SHARE. *CDISC* 2016. https://www.cdisc.org/standards/share [accessed June 6, 2018]
16. Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG Project: A Technical Report. *J Am Med Inform Assoc* 2008;15(2):130–137. doi:10.1197/jamia.M2556
17. Roelofs E, Dekker A, Meldolesi E, van Stiphout RGPM, Valentini V, Lambin P. International data-sharing for radiotherapy research: An open-source based infrastructure for multicentric clinical data mining. *Radiother Oncol* 2014;110(2):370–374. doi:10.1016/j.radonc.2013.11.001
18. Ariane AK, Rémy C, Giovanni M, Christel D, Jean C, Marie-Christine J. An Ontological Approach for the Exploitation of Clinical Data. *Stud Health Technol Inform* 2013:142–146. doi:10.3233/978-1-61499-289-9-142
19. US Department of Health and Human Services. Common terminology criteria for adverse events (CTCAE) version 4.0. *Natl Inst Health Natl Cancer Inst* 2009(04)
20. Antezana E, Kuiper M, Mironov V. Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief Bioinform* 2009;10(4):392–407. doi:10.1093/bib/bbp024
21. Holford ME, McCusker JP, Cheung KH, Krauthammer M. A semantic web framework to integrate cancer omics data with biological knowledge. *BMC Bioinformatics* 2012;13(Suppl 1):S10. doi:10.1186/1471-2105-13-S1-S10

22. Fernández-Breis JT, Maldonado JA, Marcos M, Legaz-García M del C, Moner D, Torres-Sospedra J, *et al.* Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. *J Am Med Inform Assoc JAMIA* 2013;20(e2):e288–e296. doi:10.1136/amiajnl-2013-001923
23. Pathak J, Kiefer RC, Chute CG. Using Linked Data for Mining Drug-Drug Interactions in Electronic Health Records. *Stud Health Technol Inform* 2013:682–686. doi:10.3233/978-1-61499-289-9-682
24. 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* 2012;19(5):758–764. doi:10.1136/amiajnl-2012-000862
25. Boyd S. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found Trends® Mach Learn* 2010;3(1):1–122. doi:10.1561/2200000016
26. Liu K, Kargupta H, Ryan J. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Trans Knowl Data Eng* 2006;18(1):92–106. doi:10.1109/TKDE.2006.14
27. Dekker A, Nalbantov G, Oberije C, Wiessler W, Eble M, Dries W, *et al.* Multi-centric learning with a federated IT infrastructure: application to 2-year lung-cancer survival prediction. *2nd ESTRO FORUM*, *ESTRO FORUM*. Geneva, Switzerland: Elsevier; 2013
28. Marshall MS, Boyce R, Deus HF, Zhao J, Willighagen EL, Samwald M, *et al.* Emerging practices for mapping and linking life sciences data using RDF — A case series. *Web Semant Sci Serv Agents World Wide Web* 2012;14:2–13. doi:10.1016/j.websem.2012.02.003
29. Pathak J, Kiefer RC, Chute CG. The linked clinical data project: applying semantic web technologies for clinical and translational research using electronic medical records, ACM Press; 2012. doi:10.1145/2166896.2166920
30. Rubin DL, Napel S. Imaging Informatics: Toward Capturing and Processing Semantic Information in Radiology Images. *Yearb Med Inform* 2010;19(01):34–42. doi:10.1055/s-0038-1638686

Chapter 5

Application of Machine Learning for Multicenter Learning

Authors
Johan van Soest, Andre Dekker, Erik Roelofs, Georgi Nalbantov

Adapted from
Machine Learning in Radiation Oncology: Theory and Applications, Springer International Publishing, 2015, page 71-97
Book Editors: Issam El Naqa, Ruijiang Li, Martin Murphy
ISBN: 978-3-319-18304-6
DOI: 10.1007/978-3-319-18305-3_6

Abstract

Advancements in radiation oncology are driving more specific, and thus improved, treatment opportunities. This creates challenges on the assessment of treatment options, as more information is needed to make an informed decision. One of the methods is to use machine-learning techniques to develop predictive models. Although prediction models, embedded in clinical decision support systems (CDSSs), are the foreseen solution, developing/training such prediction models requires large amounts of detailed patient information to reach decisive power. The number of patients needed to train a reliable prediction model rapidly outgrows the numbers available in a single institution, hence the need for multicenter machine learning. To be able to learn over multiple centers, several infrastructural prerequisites need to be addressed. First, data needs to be extracted from multiple source systems and represented using standardized terminologies, preferably including the semantics (the actual description) of the represented data. For research and model training purposes, this means that value representations (e.g. “m” or “f” indicating gender) need to be converted into standardized terms (the NCI Thesaurus codes C20197 or C16576, respectively), and that patient-identifiable information (e.g. name, institutional ID, address, etc.) needs to be removed or changed in a non-identifiable way. If datasets from different institutions use the same standardized terminology and data structure, data can be merged. Finally, after merging, prediction models can be learned on the complete dataset, in this chapter known as centralized learning.

Introduction

Technical advancements in the fields of physics, radiobiology, and engineering (and indirectly chemistry) are the main drivers for better, and thus more specific, treatment opportunities in radiation oncology. These advancements largely influence treatment methods, especially in regard to treatment planning (IGRT, IMRT, VMAT) and radiation techniques used.

In the current era of *evidence-based medicine* (EBM), all of these advancements need to be validated to be sure whether a specific treatment (plan) is better than the current standard (e.g., in regard to possible patient outcome). However, we also observe that new treatment options do not necessarily improve the outcome for an entire population but might only work for specific groups of patients. The standardized treatment (according to the current guidelines) might be too intense for specific groups of patients (resulting into higher toxicities and/or other radiation-induced complications) or could result in undertreatment of patients. At this point, it becomes interesting to apply machine learning to retrospectively identify prognostic factors (e.g., risk factors) and to develop predictive models to classify patients in distinct groups [1]. These groups can then be used to alter treatment options, e.g., to intensify or temper treatment.

The more subgroups we can identify, the better we can optimize treatment for individual patients, leading towards the next era, called *individualized medicine* (IM). This also imposes challenges on patient subgroup discovery and development of prognostic models as done for many years. Only several large institutions (in terms of patient turnover per year) can perform fine-grained subgroup analysis, as we need a fair number of patients with and without a specific outcome to test hypotheses regarding new treatment options for specific subgroups. Only with these large numbers of patients can we translate results of *individualized medicine* [2] into clinical practice by means of clinical decision support systems (CDSS) [3]. Therefore, we need to collaborate in radiation oncology research and share data to perform machine learning on larger, multicenter datasets.

In this chapter, we will explain the current possibilities of machine learning in a multicenter setting. We will start with the prerequisites and infrastructure fundamentally needed for multicenter machine learning. Afterwards, we will describe the concept of centralized and distributed machine learning, including the benefits and challenges. Finally, we will describe several applications/ initiatives related to multicenter machine learning and conclude with a summary of this chapter.

Prerequisites

When performing multicenter machine learning, several prerequisites are needed to be addressed before starting the machine learning process. In this paragraph, we will describe the topics of data extraction and representation, network infrastructures, and privacy preservation.

Data Extraction

Within radiation oncology, data extraction for machine learning is a labour-intensive task, as many data silos exist where data resides. In general, we need to connect to different data sources, extract data from these sources using local querying dialects, and afterwards store the extracted data in a central storage. These steps need to be performed for different information systems used in radiation oncology. We will describe the most common systems in this paragraph. First, we need to include the electronic medical record (EMR), where general patient characteristics are stored (e.g., age, gender, and diagnostic, geographical, and follow-up information such as complication and quality of life scores). Second, medical images (for diagnostic, treatment, and validation purposes) are stored in a picture archiving and communication system (PACS). Although images cannot be used directly in predictive model training, extracted information from these images can be used. Third, treatment planning-related information (e.g., radiation plan information regarding beams and dose) needs to be incorporated, as the treatment planning system (TPS) stores information in its own database, as well as in the PACS. Fourth, the record and verify system (R&V) holds information regarding the planned treatment (e.g., dose, fractionation, beams) and the actual delivery. This information is also needed during machine learning, e.g. to determine structural differences in the planned and delivered treatment.

Other systems (e.g., sources containing biological data) may apply in specific or future settings; however, we've specified only the general sources of information used for machine learning. Regarding multicenter machine learning, this data extraction leads to the first challenge as different institutions have different systems (in terms of manufacturers and products) in place. All these systems may store data differently, which requires a customized approach for data extraction for every participating institution.

ETL Tooling and Data Warehousing

To (continuously) extract data and store it in a central location, one could consider the use of extraction, transformation, and load (ETL) tooling. This tooling can extract data from different sources (different systems), reconcile data belonging to one patient (transformation), and store the data in a central database: the data warehouse (DWH). This could be useful for large-scale machine learning and research institutions with many smaller-sized trials. As shown by Roelofs et al. [4], implementing a data ware-

house can significantly reduce the data collection time, in comparison to manual data extraction and collection. In regard to multicentered settings, this also reduces the number of systems/databases a user/researcher has to include in the data request/retrieval process, thus reducing the time to merge all different datasets. Furthermore, as data are extracted and inserted into the DWH, it should be known what the data represents. The ETL process should therefore be well documented regarding queries, transformations, and the meaning of the stored data in the DWH. In comparison to the DWH, directly querying the source system for research purposes has several disadvantages. These disadvantages are mainly on the topics of query and data validity and query load on production/source systems. When a DWH is in place, query validity should not be an issue (as the data is checked before being incorporated in the DWH). Furthermore, query load issues should be mitigated, as the DWH should run on a different database/server as the production/source systems, and therefore cannot affect clinical operations.

Image Biomarker Extraction

The intrinsic information of images (not just the readily available metadata) needs to be extracted from the actual image slices. Extraction of image "features" is not a standard functionality of a PACS; however, features may sporadically be available as TPS systems may store additional information in the metadata of the DICOM images. If features are stored in the metadata, these values are needed to be validated, especially in a multicenter setting where different sites may use different TPS systems, which could implement different algorithms to calculate these features.

When there are no (or only a small number of) features already available, every site in the multicenter setting needs to implement a feature extraction pipeline which calculates variables based on the images available in the local PACS. As the local PACS stores CT and/or PET images, delineated contours (RTSTRUCT), planned (RTPLAN), and delivered (RTDOSE) dose information, the number of features to extract becomes larger. For example, we can extract information regarding the tumor volume, maximum diameter, specific points of the dose-volume histogram (DVH) for target volumes or organs at risk, tumor activity/metabolism, and differences between planned versus delivered dose. Furthermore, *radiomic* analysis on these images produces more than 200 features, based on more advanced image processing algorithms (by calculating intensity distribution metrics based on, e.g., Fourier transformations and wavelets) [5]. Several of these features are potential imaging biomarkers: features which have prognostic and predictive value in terms of patient outcome or tumor response.

Preferably, this feature extraction pipeline should use common communication protocols, such as DICOM (to receive images) and SQL (to send extracted features to a local database). This increases the possibility to reuse this pipeline in all submitting centres and increases the homogeneity of applications and calculation algorithms used by different centres. Eventually, using equal feature extraction pipelines should result in easi-

er comparison of features/variables between centres. Although we can generalize the applications and algorithms used, including scanning and reconstruction parameters, there is still a large variability at the input of this feature extraction pipeline: differences between delineations of different centres. As shown in literature, differences in delineations may occur between individuals, even within one site [6]. These differences in delineations could result in different outcomes after feature extraction. Especially when two different structures (e.g., rectum and bladder) are close to each other, for example, it might be possible that the delineating individual accidentally delineates the bladder wall as part of the rectum. This results in a higher SUV-mean/max and therefore could compromise the prognostic value of the extracted features.

Based on the examples of delineation differences and calculation applications/ algorithms used, it is important to specify the provenance of a specific variable: how did we acquire/extract this information (and which algorithms did we use)? And what are the sources used to extract the information? We will elaborate on these questions in the next paragraph.

Data Representation and Semantic Interoperability

To be able to exchange data between participating sites, all sites need to be *syntactically interoperable*. This means that they have to agree which (technical) protocol they use to transfer data; implying that data representation should be equal among participating sites.

Next to standardization of syntactical interoperability, *semantic interoperability* needs to be in place. We will use the definition of Valentini et al. [7] to describe semantic interoperability: "The ability of any communicating entity (not only computers) to share unambiguous meaning. For computers, this is the ability to exchange information and have that information properly interpreted by the receiving system in the same sense as intended by the transmitting system." In general, this means that the receiver cannot interpret information differently, as the sender uses unambiguous terms to describe that information. Therefore, we need to use terminological systems which are known by both sender and receiver. As defined by De Keizer et al. [8], a terminological system can be a thesaurus, classification, vocabulary, nomenclature, or coding system. A terminological system may pertain to more than one of these systems. For example, ICD-10 [9] is a coding system and vocabulary (as the term is accompanied by a definition); another example is the National Cancer Institute's Thesaurus (NCIT) [10], which (in addition to a vocabulary) also contains a list of synonyms or other relationships. Finally, multiple terminological systems can be embedded in an ontology, where concepts from terminological systems are reused and relations of concepts in a specific domain are described. Furthermore, an ontology can be used as a consensus model to represent data within a specific domain (e.g., radiation oncology) between different participating sites [8].

Relational Databases and Ontologies

In regard to multicenter learning, we need to make sure every participating site uses the same database structure to be able to uniformly query (or federate) the data warehouse (DWH) database. This database structure can be derived by creating a so-called entity-relationship (ER) model, based on the ontology; however, it needs to be adhered by all centres. An example to derive this ER model is the normalized universal approach described by Gali et al. [11]. Next to this database structure, it is important to use the same database system, as different database systems/vendors have different dialects. To mitigate differences in database systems/ vendors, it is also possible to use automatic conversion libraries such as Hibernate (http://hibernate.org/), although these systems add another layer of complexity when performing queries and/or data federation.

When adhering to an ontology, values from local systems need to be replaced with standardized values from terminological systems as defined in the ontology. For example, the property biological sex containing the text "male" or "female" needs to be replaced by NCI Thesaurus code C20197 or C16576, respectively. Another participating site may use 0 and 1 or "m" and "f"; however, within the DWH database, all sites should use the NCI Thesaurus codes for semantic interoperability. This conversion of values is typically done in the *transform* step of the ETL process. Therefore, the ETL process needs to be tailored per participating centre.

Although data representation is possible within relational databases, it is cumbersome to maintain in a multicenter machine learning setting. As new results give new insights into biological concepts and relationships, the need for extra variables is rapidly growing. Given this fact, it is inevitable that a multicenter network for machine learning will have substantial downtime. For example, when a new concept is added to the ontology, every participating site needs to update their ETL system and DWH database structure, to become up to date with the new ontology version. This may take some time, as administrators of the ETL and DWH system need to validate whether this change is valid and does not compromise patient deidentification. If one of the queried columns is not available, the Relational Database Management System (RDBMS) will result an error rather than an empty result set. Therefore, it might be that the whole federation/distributed querying system may not work (if proper error handling is not in place). In this example, we used the addition of a column, a relatively easy task which occurs frequently. However, the more complex the changes in the ontology and database structure, the more time and effort it will take to get the network up and running again.

Semantic Web, RDF, and Linked Data

One of the solutions to cope with rapidly changing ontologies in a multicenter setting is to move from relational databases to Semantic Web technologies [12,13]. In this paragraph, we will only discuss the *Resource Description Framework* (RDF), *linked data*, and the *SPARQL protocol and RDF query language* (SPARQL) as a subset of Semantic Web technologies.

Resource Description Framework

RDF is a standard, recommended by the World Wide Web Consortium (W3C) [14], and can be seen as a flexible alternative for the relational database. Where "traditional" relational databases store their data in a structure of tables and columns, the RDF specifies only one table with three columns named *subject, predicate,* and *object.* Each row in this single table repository is called a triple, as it only has three cells. Due to this basic difference in structure, the concept of data representation is also different. Because of this fixed table structure, the ontology becomes more important and serves as a data model consensus between centres.

As an example, we have an ontology describing patients and their first name, last name, biological sex, and age. Figure 1 shows the visual representation of this ontology. The RDF triples based on this ontology are represented in Table 1.

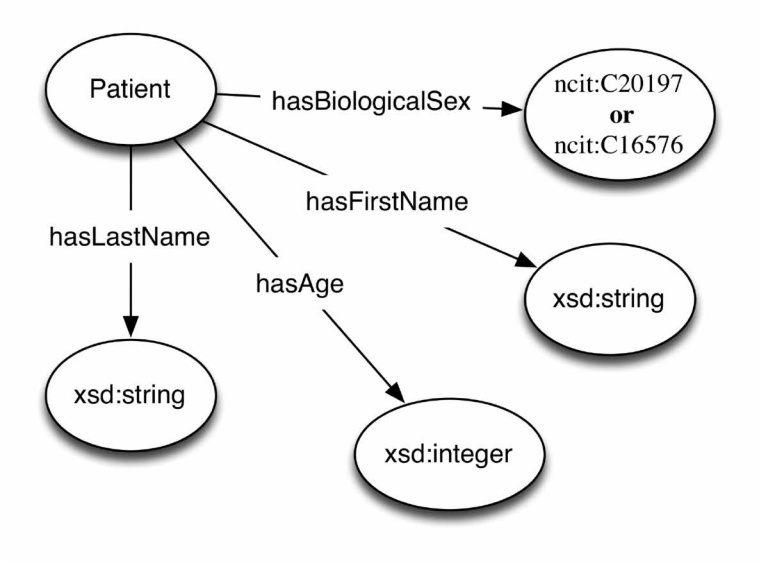


Figure 1 - Visual representation of the sample ontology

Table 1 - RDF representation of a patient based on the ontology of Figure 1

Subject	Predicate	Object
mySet:patient1001	rdf:type	ncit:C16960
mySet:patient1001	myOntology:hasFirstName	"John"^^xsd:string
mySet:patient1001	myOntology:hasLastName	"Doe"^^xsd:string
mySet:patient1001	myOntology:hasBiologicalSex	ncit:C20197
mySet:patient1001	myOntology:hasAge	"67"^^xsd:integer

Unique Resource Identifiers and Linked Data

To assure semantic interoperability, we will use the concept of unique resource identifiers (URIs), which is incorporated in the RDF specification. The RDF specification states that all resources (concepts and predicates) need to have a URI, which can be a unique resource locator (URL; e.g., http://www.mydomain.org/ontology#hasFirstName) or a unique resource name (URN; e.g., myOntology:hasFirstName). This means that someone needs to own a domain name (e.g., mydomain.org) and is administrator of this domain. If this is the case, he or she can make unique URLs for this domain, for example, to create a unique URI for patient 1001 (e.g., http://www.mydomain.org/ rdf#patient1001). If the domain administrator assigns a specific sub-path of the domain to a dataset (called *a namespace*), for example, http://www.mydomain.org/rdf#, then this sub-path can also be substituted by a *prefix*, for example, "mySet". This namespace can then be used to shorten the notation of a unique patient, as shown in Table 1. This concept of unique resources also holds for ontologies, where in Table 1 the prefix "myOntology" can be used to define the namespace http://www.mydomain.org/ ontology# and the prefix "ncit" refers to the unique location of the NCI thesaurus. As everyone should use the same, unique namespaces, the use of URIs enforces semantic interoperability. Therefore, semantic interoperability is enforced within the Resource Description Framework.

Next to the enforcement of semantic interoperability, the use of URIs has a second benefit, namely, the possibility of linked data. As every resource has its unique URI, an RDF store at site A may point to a resource at site B by using the URI of the resource at point B [15]. For example, if a patient underwent a diagnostic scan at hospital A and was treated in clinic B, then clinic B can specify the treatment and link it to the patient resource with the unique URI used in hospital A.

Querying using SPARQL

We have described how data can be represented in RDF, and how URIs enforce semantic interoperability and linked data. But how can we retrieve this data from an RDF store? To query these RDF stores, the W3C has adopted the *SPARQL protocol and RDF query language* (SPARQL) [16]. Most RDF stores have integrated a SPARQL endpoint in their RDF store. A SPARQL endpoint is the public interface to receive SPARQL queries and return a result table, all using the HTTP protocol. In contrast to SQL queries, SPARQL queries do not search tables due to the underlying RDF store structure. SPARQL queries perform pattern matching on the triples in the triple store, where variables can be used to retrieve unknown values or to dynamically link values. For example, the query in Listing 1 will try to retrieve the first name, last name, and age for all patients. We will shortly describe the lines in this query example.

```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2  PREFIX ontology: <http://www.mydomain.org/ontology#>
3  PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
4
5  SELECT ?patient ?firstName ?lastName ?age
6  WHERE {
7     ?patient rdf:type ncit:C16960 .
8     OPTIONAL { ?patient ontology:hasFirstName ?firstName . }
9     OPTIONAL { ?patient ontology:hasLastName ?lastName . }
10    OPTIONAL { ?patient ontology:hasAge ?age . }
11 }
```

Listing 1 - Basic SPARQL query retrieving patient resources, related first and last names, and age of patient data stored in an RDF store, based on the ontology defined in Figure 1.

On line 1–3, the shorthand (prefix) notations for URL locations are defined. Line 5 defines the variables retrieved from the pattern matching; these variables have to start with a question mark. Lines 6–11 define the actual pattern searched for. As shown in Listing 1, our basic pattern is to retrieve all patient resources which have a predicate called "rdf:type," which refers to the terminological code of a patient, defined in the NCI Thesaurus (using the prefix "ncit:," which is replaced by the full URL at line 3). Afterwards, we extend our pattern match by including extra properties for every resource linked to the patient resource. If the linked resources of the patient variable have a predicate matching to our specified property (in our ontology), then the variable firstName, lastName, or age will be filled with the found value. If not found, then the query will return the patient resource URI; however, the variables firstName, lastName, or age are not filled in (due to the "OPTIONAL" keyword).

Next to querying one RDF store, a SPARQL query can also be federated to multiple stores. This is an advantage in regard to multicenter learning, as a single query can retrieve data from multiple sources. Due to the structure of RDF stores, data residing in geographically separated RDF stores can easily be merged, as the data structure is the same for all stores (1 table; 3 columns) and all RDF stores should use URIs. Federation can be done both horizontally (different patients in different RDF stores) or vertically (information of a single patient stored in multiple RDF stores). An application of horizontal federation in SPARQL queries is shown in Listing 2; an application of vertical federation is shown in Listing 3.

```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2  PREFIX ontology: <http://www.mydomain.org/ontology#>
3  PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
4
5  SELECT ?patient ?firstName ?lastName ?age
6  WHERE {
7          SERVICE <http://endpoint1.mydomain.org/> {
8                  ?patient rdf:type ncit:C16960 .
9                  OPTIONAL { ?patient ontology:hasFirstName ?firstName . }
10                 OPTIONAL { ?patient ontology:hasLastName ?lastName . }
11                 OPTIONAL { ?patient ontology:hasAge ?age . }
12         }
13
14         SERVICE <http://endpoint2.mydomain.org/> {
15                 ?patient rdf:type ncit:C16960 .
16                 OPTIONAL { ?patient ontology:hasFirstName ?firstName . }
17                 OPTIONAL { ?patient ontology:hasLastName ?lastName . }
18                 OPTIONAL { ?patient ontology:hasAge ?age . }
19         }
20 }
```

Listing 2 - An example of horizontal federation in a SPARQL query

```
1  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2  PREFIX ontology: <http://www.mydomain.org/ontology#>
3  PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
4
5  SELECT ?patient ?firstName ?lastName ?age
6  WHERE {
7          SERVICE <http://endpoint1.mydomain.org/> {
8                  ?patient rdf:type ncit:C16960 .
9                  OPTIONAL { ?patient ontology:hasFirstName ?firstName . }
10                 OPTIONAL { ?patient ontology:hasLastName ?lastName . }
11         }
12
13         SERVICE <http://endpoint2.mydomain.org/> {
14                 OPTIONAL { ?patient ontology:hasAge ?age . }
15         }
16 }
```

Listing 3 - An example of vertical federation in a SPARQL query

In these examples, we will use the "SERVICE" command of SPARQL to identify the execution of a subquery (or pattern match) on a different SPARQL endpoint. In Listing 3, we used the exact same pattern query in both services/subqueries (line 7–19). Both subqueries are sent to the respective endpoints, and the subquery results are merged at the federation endpoint. Finally, the requested variables are returned to the requesting application or user. In Listing 3, both services have different patterns to match. The first service (line 7–11) searches for all patients and their first/last name on SPARQL

endpoint 1. The second service (line 13–15) will reuse the patient resources found in endpoint 1 and tries to find patterns matching the hasAge predicate for these given patient resources. When found, it will use the object linked to the hasAge predicate (in this case a literal of type integer) and store it in the variable "?age". Finally, the query engine will return the output as one table (using the variables of line 5 as columns), including information retrieved from both endpoints.

In this paragraph, we have presented an alternative to the widely known relational databases to represent and retrieve data. The use of Semantic Web technologies, and especially RDF, has several advantages over relational databases. Especially the meta-structure of RDF (independent of the modelled domain) and the use of URIs are useful with regard to a flexible storage solution while inherently adopting semantic interoperability and linked data.

On the other hand, using Semantic Web technology has some downsides when used in multicenter machine learning. The main downside is that local institute staff needs to be introduced to Semantic Web technologies, in order to maintain these data repositories and endpoints. Furthermore, development in the field of RDF stores/ repositories is an ongoing process and is not yet comparable to relational databases in terms of reliability and performance, especially in daily clinical practice. On the contrary, for research projects (where uptime is less critical), the Semantic Web is more favorable because of its flexibility in storage and data structures.

Network Infrastructure

Up until now, we only described how to extract information from multiple sources (databases, image archives) and to apply standardized terminological systems on the data extracted from these sources. Furthermore, we have described how to represent data using the relational database and semantic web technology. In this paragraph, we will combine the topics of the previous paragraphs and explain how we can use them together. First, we will describe the institutional infrastructure, after which we will describe the multicenter infrastructure.

Institutional Infrastructure

In this paragraph, we will describe several approaches to represent a single point of access for the outside world (e.g., participating sites in the multicenter machine learning setting). We will discuss five different approaches, namely:

- Traditional ETL and DWH
- Traditional ETL and DWH with an RDF store
- Traditional ETL and DWH with a virtual RDF store
- Virtual RDF store per institute
- Virtual RDF store per source and institute

Traditional ETL and DWH

In the approach using relational databases, records from different source systems (e.g., EMR, PACS, TPS, and R&V) are merged using an ETL tool and converted into the requested data formats following standards used by all collaborating sites (Figure 2). The merged and transformed data are being saved in the DWH database. This database will afterwards be queried when requesting data for machine learning purposes. Therefore, this database needs to be compliant to the ontological structure (among all participating centers). When the ontology is altered, all participating centers need to update the DWH database structure, as well as the transform and/or storage scripts in the ETL tooling.

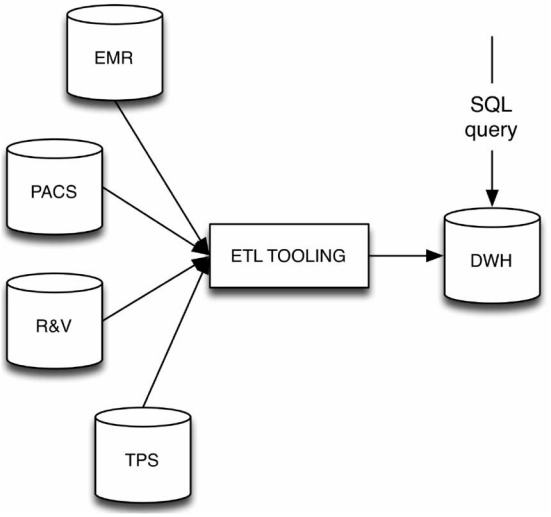


Figure 2 - Infrastructure of the traditional ETL and DWH approach

Traditional ETL and DWH with an RDF store

This approach uses an RDF store on top of the traditional ETL and DWH approach (Figure 3). It enables the possibility to create an institutional DWH instead of a DWH dedicated for the study. Afterwards, the "Database to RDF" conversion application reads the DWH database and transforms the data it into triples, taking into account a given ontology. This RDF store will afterwards be queried when requesting data for machine learning purposes. Only the "Database to RDF" application needs to follow the rules and data structure defined in the ontology. When the ontology is altered (e.g., adding an extra data element), only this database-to-RDF application needs to be altered (when the information is already available in the DWH). Updating the RDF store is done by clearing and repopulation and is performed at specific time intervals.

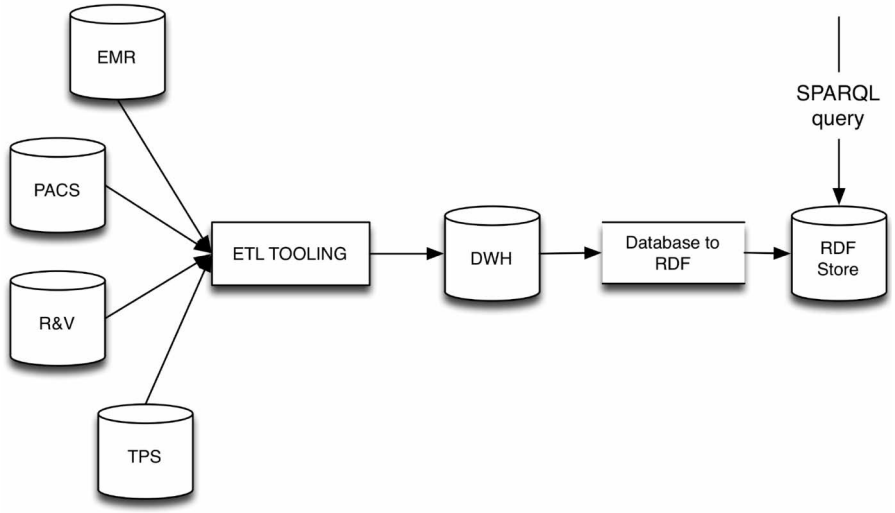


Figure 3 - Infrastructure of the approach using a traditional DWH with an RDF store

Traditional ETL and DWH with a Virtual RDF Store

This approach uses only the database-to-RDF conversion application on top of the traditional ETL and DWH approach (Figure 4). This approach is almost equal to the physical RDF store approach (Figure 3); however, it has one difference in converting data from relational databases to RDF.

In this case, the "Database to RDF" application acts as a SPARQL endpoint, accepting SPARQL queries and returning the result of these queries. There is no data stored, as there is no RDF store, only a SPARQL endpoint. When performing a SPARQL query, the database-to-RDF application will transform SPARQL queries into SQL queries and executes these SQL queries on the DWH. In regard to maintenance, this option holds the same requirements as using the physical RDF store. The only difference is the absence of an intermediate RDF store, resulting in real-time results of the data available in the DWH.

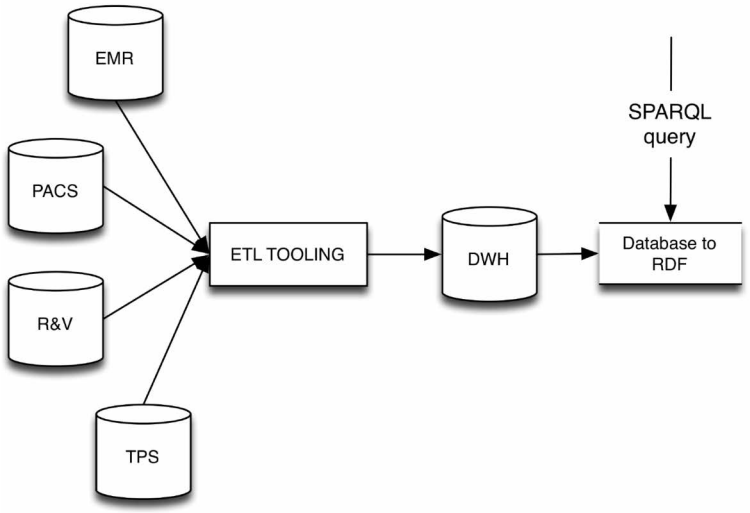


Figure 4 - Infrastructure of the approach using a traditional ETL and DWH with virtual RDF store

Virtual RDF Store per Institute

As the DWH usually is not a real-time representation of the clinically available data, this approach removes the DWH and directly queries the source systems. In this approach, the database-to-RDF application is functioning as a SPARQL endpoint without an RDF store and converts SPARQL queries into SQL queries for the different source systems (Figure 5). It therefore creates challenges for the database-to-RDF application, as it needs to transformation data (to convert local terms to standardized terms), which was previously done by the ETL tooling. If multiple source systems are involved, the database-to-RDF application merges the results from all sources and presents them as a SPARQL query result. The main benefit of this approach is that we can query for real-time data, rather than have to wait before the data is added to the DWH. Furthermore, data redundancy of the intermediate storage (the DWH) is not needed, reducing the need for storage resources. However, the main disadvantage is with regard to performance, as data and queries are transformed on the fly.

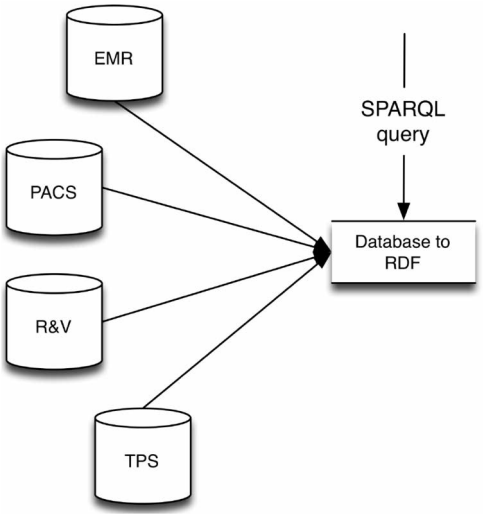


Figure 5 - Infrastructure using only a virtual RDF store

Virtual RDF Store per Source and Institute

This approach is almost similar to the "Virtual RDF store per institute" approach, however, with differences in data transformation and federation (Figure 6). First, every local data source will get a SPARQL endpoint, using, for example, the database-to-RDF application. This application will convert the data from the source system into RDF, compliant with the ontology used in the multicenter setting. Afterwards, the central federation endpoint will be used to merge all triples from all database-to-RDF applications/sources (vertical federation). In this setting, one SPARQL query will be sent to the federation endpoint. This federation endpoint will split the SPARQL query into several sub-SPARQL queries and execute these SPARQL queries on the SPARQL endpoints placed on top of the data sources. Afterwards, the federation endpoint will merge the results and return the merged result set to the application/user performing the query. The benefit of this approach is the distribution of computational resources to reduce the query execution time. The drawback is that n+1 applications (where n is the number of database-to-RDF applications) need to be maintained and updated when the ontology changes.

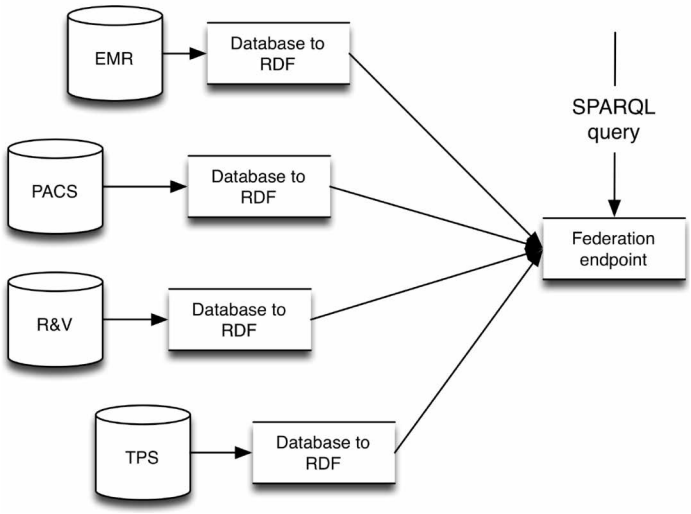


Figure 6 - Infrastructure using a virtual RDF store per source and institute

Multicenter Infrastructure

In the previous paragraph, we described the institutional infrastructure options to create one façade or data query endpoint for every center. It depends on whether we are using centralized or distributed machine learning and whether we need an additional computation unit (e.g., a dedicated or virtual server) in each center. Both distributed and centralized approaches can be implemented using relational databases or Semantic Web technology; however, the decision regarding data representation techniques needs to be made upfront and accepted by all participating centers. In this paragraph, we will first describe the centralized machine learning infrastructure and afterwards move towards the distributed infrastructure.

Centralized Multicenter Infrastructure

The general overview for the centralized multicenter infrastructure is shown in Figure 7. The participating sites are displayed as a data store, as we do not need to know what the institutional infrastructure looks like. This approach gives participating centers the opportunity to establish the institutional infrastructure according to local policies. Additional to all institutional entry points, a central machine learning server (performing the computations) and a central federation point need to be set up. The central federation point will perform the horizontal federation between participating centers. To ensure privacy, the data stores of the participating centers may limit external access by only allowing access from the central federation point. The central machine learning server will accept and execute algorithms (including queries to execute on the central federation point). After the algorithm has finished, it will return the outcome of the computation to the external source which sent the job (algorithm+query).

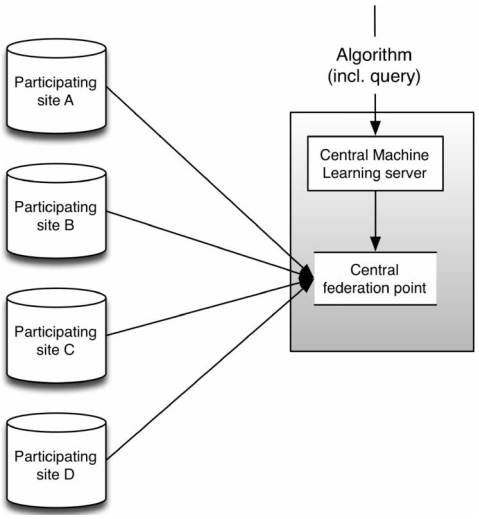


Figure 7 - Centralized multicenter infrastructure

Distributed Multicenter Infrastructure

The distributed multicenter infrastructure is different from the centralized version with respect to computational locations. As shown in Figure 8, the central federation point has been removed, and local computation units (machine learning slaves/ agents) have been introduced. In this infrastructural setting, the central machine learning server (master server) is a coordinating server. When a job (algorithm+query) is submitted to the central ML master, the algorithm is being split into smaller sub-algorithms. These sub-algorithms and queries are packed into subjobs and sent towards the local computation units. They will query the local endpoint and execute the sub-algorithm. After finishing the sub-algorithm, the results are sent back to the central ML master, which gathers the results from all local endpoints. The central master will then determine whether it will perform a new sub-job on all endpoints or aggregate values and sends the final (aggregated) result back to the job-submitter.

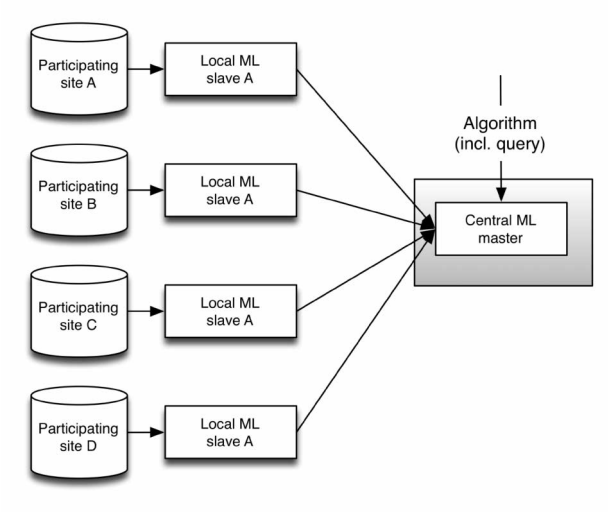


Figure 8 - Distributed multicenter infrastructure

Privacy Preservation

For both distributed and centralized multicenter infrastructures, privacy preservation is a major topic to consider. If correctly implemented, the distributed multicenter infrastructure is generally more secure as the results of the algorithm (e.g., a predictive model) are transferred instead of the source data. However, this does not mean that the issues concerning privacy preservation are solved. For example, it is still possible to retrieve metadata about a dataset of one patient. In this section, we will address several options for privacy preservation, ranging from pseudonymization to irreversibly modifying the original datasets. Despite of all the options described below, we must state that, in our opinion, there is no standard method to ensure privacy preservation. The researcher/designer of the infrastructure will always have to find a balance between the loss of information and the anonymity of participating patients.

Pseudonymization

The first option for privacy preservation is bidirectional pseudonymization of patient identifiers, for example, replacing patient names and hospital's patient identification numbers by study-specific alternatives. This can be achieved by maintaining a two-column table, where one column contains the patients' identification number and the second column contains the study identification number for this patient. Variations to this concept may apply, for example, using an extra column to maintain the study where this mapping applies to. Typically, the pseudonymization of hospital to study identification numbers is done during the transform part of the ETL process. Other patient identifying information (e.g., first and last names) can be replaced by the same study ID or may not be incorporated and thus removed during the ETL process.

The second option is to use an unidirectional pseudonymization algorithm, for example, by hashing patient identifiers (e.g., using an SHA-{1–3} algorithm). This hash should be unidirectional, meaning that the pseudonymized patient identifiers cannot be reversed to the original identifiers. Unidirectional pseudonymization might be more appropriate than bidirectional pseudonymization, however might introduce problems when study data are needed to be linked to the actual patients. For example, when study results show a worse outcome for specific patients and when it is immoral to withhold this information to these patients.

Data Obfuscation

When using strict inclusion criteria with rare variables, it might be that the resulting dataset is very small and patients might become identifiable by combination. For example, if only two patients match some inclusion criteria and the biological sex (which is a requested variable) is different in both patients, we can identify these patients when querying local source systems. This issue holds for both the centralized and distributed multicenter infrastructures. To reduce the chance of compromising the anonymity of patients, Murphy and Chueh [17] introduced a method for data obfuscation where (especially in the case of a small number of events/patients) results are obfuscated by returning a random value within a specific range based on the actual value. This method does not circumvent the problem completely, as someone with bad intentions is able to approximate the original value by sending the same request multiple times. To circumvent these actions, Murphy and Chueh proposed to implement an audit system, where performing the same query multiple times within a specific time span will result in a request denial. In this way, the system returns a value not completely representing the actual value however returns a value within a tolerable margin (when not exceeding the maximum number of requests).

Data Perturbation

The downside on obfuscation is that it does change the distance (e.g., Euclidian distance) between points (e.g., patients or observations) in a *k*-dimensional space, where every dimension may be a specific variable in the dataset. As the distance changes, it may influence the prediction model training algorithm and train a model that does not represent the actual data and values. This can lead to problems during validation, especially when the validation data is obfuscated, however in another way (due to the randomness in the obfuscation algorithm). Therefore, transformation of data might be a solution, as the whole dataset is transformed while maintaining the distance between points. As shown by Liu et al. [18], this transformation is still not good enough for privacy preservation, as the original data can be derived using independent component analysis (ICA) or overcomplete ICA. To overcome this issue, Liu et al. advise to use their random projection-based multiplicative perturbation (RPBMP) method, which reduces the number of dimensions and transforms the dataset while maintaining statistical

information regarding the distance between variables. Using this method, it should not be possible to retrieve the original values and would therefore obstruct the possibility to match variables to individual patients. This RPBMP method is afterwards reused by Yu et al. [19], where they explored differences in dimension reduction options and applied it to a non-small cell lung cancer (NSCLC) dataset. Data perturbation and dimension reduction are potential solutions to preserve privacy in a multicenter setting, although they could lead to issues when performing a risk analysis (identifying variables which influence a specific outcome). The risk analysis then can only determine which *compressed* dimensions are of influence; however, it cannot determine which biological (or source) variables/features are responsible for this influence in patient outcome.

Centralized and Distributed Machine Learning

When the prerequisites regarding semantic interoperability, data structure, infrastructure, and privacy preservation are in place, we can start performing machine learning. In this section, we merely touch upon centralized machine learning in favour of describing distributed machine learning approaches in full, which are considered superior for future, large-scope implementations.

Centralized Machine Learning

As described above, the centralized approach only needs one machine learning unit (Figure 7). In this case, the machine learning system will query and retrieve data from the federation data store, irrespectively of knowing where the actual data comes from (except when provenance variables are included in this dataset). As the retrieved dataset is not different in comparison to traditional machine learning approaches, we can use standard machine learning toolboxes such as Weka [20], RapidMiner [21] or others [22]. The disadvantage is that data, with/without privacy preservation in place, is transferred to a central location at time of machine learning algorithm execution. This might contradict the policy of centers regarding data sharing.

Distributed Machine Learning

The major difference between distributed and centralized machine learning is the transfer of data versus the transfer of training models. In the centralized approach, data is transferred to the machine learning system, whereas in the distributed approach the data stays within the institute. Rather than requesting a dataset, the distributed approach dispatches a sub-process of the machine learning algorithm towards the institutional machine learning unit and returns the result of this sub-process. In this setting, the amount of data per transfer diminishes; however, the data transfer frequency increases. A thorough explanation how distributed machine learning algorithms work is

given by Boyd et al. [23] and Wu et al. [24]. Another approach is the MapReduce concept, developed by Dean and Ghemawat [25], to implement the distributed machine learning concept. This is similar to the rationale used by Wu et al. [24].

Linear Regression Implementation

The MapReduce concept can be explained by using the linear regression algorithm. Before we explain the MapReduce concept, we will first explain the intuition of the linear regression algorithm, using the nonstandard gradient descent approach. This approach is different from the closed-form solution; however, it enables the possibility for distributed multicenter learning, as we will show afterwards. A trained univariate linear regression classifier can be expressed using the function:

$$f(x) = \alpha + \beta x \tag{1}$$

To learn the α and β parameters of this model, the linear regression training algorithm can be described as shown in Algorithm 1.

1: **procedure** TRAINLINEARREGRESSION (x, y)
2: $\alpha = random()$
3: $\beta = random()$
4: $J = cost(\alpha, \beta, x, y)$
5: $J'_{\alpha} = costDerivativeAlpha(\alpha, \beta, x, y)$
6: $J'_{\beta} = costDerivativeBeta(\alpha, \beta, x, y)$
7: **while** $J' \leq 0$ **do**
8: $\alpha = updateVariable(\alpha, J'_{\alpha})$
9: $\beta = updateVariable(\beta, J'_{\beta})$
10: $J = cost(\alpha, \beta, x, y)$
11: $J'_{\alpha} = costDerivativeAlpha(\alpha, \beta, x, y)$
12: $J'_{\beta} = costDerivativeBeta(\alpha, \beta, x, y)$
13: **end while**
14: **return** α, β
15: **end procedure**

Algorithm 1 - Linear Regression Training

The input parameters for Algorithm 1 are *x* and *y*, respectively, determining the prediction values and outcomes for which we are training this univariate linear model. On line 2 and 3, initial α and β values are randomly chosen. Afterwards (line 4), the cost (a measure of distance between the calculated outcome, and the actual outcome) for the randomly chosen α and β is calculated using the following function:

$$J(\alpha,\beta) = \frac{1}{2m} * \sum_{i=1}^{n} \left(f\left(x^{(i)}\right) - y^{(i)} \right)^2 = \frac{1}{2m} * \sum_{i=1}^{n} \left(\left(\alpha + \beta x^{(i)}\right) - y^{(i)} \right)^2 \quad (2)$$

In this function, the variable *n* represents the number of observations used for training. For every observation, both *x* and *y* need to be available. After calculating the cost, we also need the partial derivatives of the cost function. We can write both partial derivatives as

$$J'_{\alpha} = \frac{\partial}{\partial \alpha} J(\alpha,\beta) = \frac{1}{m} * \sum_{i=1}^{n} \left(\left(\alpha + \beta x^{(i)}\right) - y^{(i)} \right) \quad (3)$$

$$J'_{\beta} = \frac{\partial}{\partial \beta} J(\alpha,\beta) = \frac{1}{m} * \sum_{i=1}^{n} \left(\left(\alpha + \beta x^{(i)}\right) - y^{(i)} \right) x^{(i)} \quad (4)$$

After calculation of these parameters, we can enter the main loop of Algorithm 1. In this loop, we'll first update the *α* and *β* variables (line 8 and 9) using the following functions:

$$\alpha = \alpha * c * J'_{\alpha} \quad (5)$$

$$\beta = \beta * c * J'_{\beta} \quad (6)$$

In these functions, the variable *c* determines the gradient descent rate. After calculating the new *α* and *β*, the algorithm continues by calculating the cost function and the partial derivatives of the cost function again (lines 10–12). Afterwards, it will start a new iteration of this loop (line 7), calculating the new *α* and *β*, calculating the cost, and calculating the partial derivatives (lines 8–12). This process is repeated until the algorithm reaches one of several termination criteria. Most preferably, the partial derivatives should converge close to 0, as this would indicate that an optimum has been reached. Alternative termination criteria are the number of iterations (as the algorithm does not converge, e.g., due to a large gradient descent rate *c*) or no significant changes in the calculated cost of the last m iterations. Finally, the algorithm will return both *α* and *β* as the outcome of the training algorithm.

Cost function and MapReduce

As shown in Algorithm 6.1, the calculation of the cost and/or the partial derivatives is the only function in the algorithm where the original data is needed. These functions

also consume most of the computational resources and are positively correlated to the number of observations and variables incorporated in the regression model. To reduce the computation time, we can split the original data and refactor the cost and partial derivative functions to multiple machines and/or processing units. When reducing the number of summations (see equations 2, 3, and 4) for every processing unit, we can reduce the overall time to calculate the cost and/or partial derivatives.

This distribution of processing power can be achieved using the MapReduce concept. In the previous example, we can implement the MapReduce concept as shown in Algorithm 2.

```
1:  procedure CALCULATEMAPREDUCECOST(α, β, x, y)
2:      for processing unit u in processingUnits do
3:          u.startCalculatingSquaredDistance(α, β, x, y)
4:      end for
5:
6:      wait()
7:
8:      distanceSq = 0
9:      m = 0
10:     for processing unit u in processingUnits do
11:         result ← u.retrieveSquaredDistances(α, β, x, y)
12:         distanceSq = distanceSq + result[0]
13:         m = m + result[1]
14:     end for
15: return calculateCost(distanceSq, m)
16: end procedure
```

Algorithm 2 - MapReduce implementation of Cost Function

This algorithm first starts with the "Map" part and subsequently performs the "Reduce" part of the MapReduce concept. At first, all registered processing units are invoked to start the following summation calculation (line 2–4):

$$D = \sum_{i=s}^{n} \left(\left(\alpha + \beta x^{(i)} \right) - y^{(i)} \right)^2 \tag{7}$$

Note that this summation calculation only calculates the summation over a specific subset (from observation *s* to *n*). The result of this squared difference summation is afterwards temporarily stored or directly returned to the initiator.

After the initiator has waited until all processing units have finished their calculation, the "Reduce" part of the algorithm comes in. The initiator (and therefore also the "Reducer") sums all squared distances and the number of individuals processed used for this calculation (line 8–14). Afterwards, the initiator calculates the total cost using the equation:

$$J = \frac{1}{2m} * D \tag{8}$$

In this equation, we use the summed squared distances from all processing units (variable *D*) and the number of observations from all processing units (variable *m*), resulting in a result equal to Eq. 2 (the non-parallelized cost function).

To calculate the partial derivatives of the cost function, we could reuse Algorithm 2 and modify equations 7 and 8 to use the summed squared distances and numbers of observations.

MapReduce, Distributed, and Multicenter Machine Learning

As shown above, we are able to distribute the resource-intensive part of the linear regression training algorithm over multiple processing units. As previously stated, processing units can be multiple CPUs or multiple computers. In the latter situation, we need to include the data in the invocation of the summed squared distances calculation or have the data already available at all processing computers (e.g., by mirroring files/databases and/or using network drives). This concept of distributing the computation over multiple machines and mirroring the data is typically done in grid computing. To reduce the data needed to be mirrored, one can consider splitting the original data over the involved computers. As the squared difference summation is aggregated during the "Reduce" part of the algorithm, there is no need to have the complete dataset on all computers.

The absence of complete datasets on all involved machines creates the opportunity for distributed multicenter machine learning. When implemented as in Algorithm 2, the "Reducer" does not need to know the size of the complete dataset at forehand (variable *m*). All processing units send back two variables to the Reducer: the summation of the squared distance and the number of observations. This preserves privacy by not sending over the actual patient information, only the aggregated results, while still being able to perform machine learning over large datasets in different institutes. Therefore, we can perform distributed multicenter learning using the MapReduce concept, where the original data does not leave the centers.

Training, Testing and Validation

Now we defined a concept of distributed multicenter machine learning (for training purposes), we can perform testing and validation on this infrastructure. To perform testing and/or validation, the easiest approach is to use one participating institute as the testing dataset and one participating institutes' dataset as the validation dataset. For example, if we have five participating institutes $I \in i_{1-5}$, we can use the first three institutes (i_{1-3}) to execute the distributed training. Afterwards, the master node can send the trained model to i_4 for testing purposes, resulting in performance metrics of

the prediction model, for example, determining discriminative and/or accuracy (e.g., c-index, Brier score, Hosmer-Lemeshow test) of the trained model. After model development has finished, an external validation can be performed by sending the model to i_5, which calculates and returns the performance of the trained prediction model.

A second, more elaborate, option is to distribute the testing and validation steps over all participating institutes I. In this case, training would be done on 60 % of all subjects and testing and validation on, respectively, 20 and 20 % of all remaining subjects. If done correctly, assignment into the training, test, or validation set should be determined before starting the model training; however, it should be remembered during the whole process. This adds extra complexity to the computation units within the institutes. These units have to remember for which distributed learning algorithm the computational instruction is and determine which dataset to use. For the testing phase, this approach might be better, as the training and testing datasets are homogeneous. For external validation, it raises the discussion whether an external validation set should be from a completely different center.

Finally, we can perform a *k*-fold cross-validation, where the number of participating centers can determine the number of folds, where we would use $I_{learn} \subset I$, with $i_n \notin I_{learn}$ as the training dataset. In this case, i_n will be the validation set for a specific fold.

Applications

In the previous paragraphs, we defined the prerequisites and described how to perform distributed machine learning. In this paragraph, we will discuss several initiatives and applications of multicentre learning. It is not mandatory that all applications use the complete set of prerequisites described previously in this chapter.

I2B2

The Informatics for Integrating Biology and the Bedside (I2B2; http://www.i2b2.org) project aims at integrating data from different biomedical disciplines and delivering this data to researchers. The project delivers tools to translate genomic and biologic findings to clinical findings (e.g., diseases or disorders). To be able to achieve this *translational medicine* approach, institutional data sources are federated in the I2B2 DWH using ETL tooling. The DWH database structure, called the Clinical Research Chart (CRC), is generic for medical purposes, as it does not define specific data fields. The database structure is basically a “star schema” where only patient information and observations are stored [26]. To describe all information in an observation-centered storage, local terminologies, or standardized terminological systems, are needed to define different types of observations. Afterwards, researchers can query/request data. When a specific dataset has been queried, this dataset can be stored in a separate database, using the same

CRC database structure. In this separate database, researchers can clean/modify the dataset to their needs and execute machine learning algorithms on this dataset.

In regard to multicenter machine learning, I2B2 supports merging multiple research databases using the Shared Health Research Information Network (SHRINE) tool [27], resulting in a federated research database of multiple institute research databases. Therefore, it enables the opportunity for centralized multicenter learning. In this approach, the terminology to define observations can be aligned when merging databases or can be kept separate [28]. In the latter approach, the researcher has to put in more effort in data alignment during the analysis, which is not favourable as it is prone to causing mistakes in the analysis.

EuroCAT

The Euregional Computer-Aided Theragnostics (EuroCAT; http://www.eurocat.info) project aims at reuse of clinical data for research purposes and to improve the speed and quality of clinical research. The project uses a distributed learning approach as described before, targeted at prediction models for lung cancer. To be able to perform this distributed learning approach, a so-called umbrella protocol was developed by the participating partners. This protocol describes the standardized data collection, including the variables to record (and terminological systems to use), questionnaires, and informed consent document templates. The first version of the EuroCAT system used a DWH and ETL infrastructure at the local institutes. Afterwards, the DWH was replaced by an RDF store. The EuroCAT system has shown that distributed multicenter machine learning works and produces the same results as centralized learning when implemented correctly [29]. Furthermore, the project has shown that distributed multicenter learning does improve the robustness of prediction models when validating on an external dataset [30].

VATE

The VATE ("VAlidation of High TEchnology based on large database analysis by learning machine") project shares the aim of the EuroCAT project. The major difference is that this project is based on open standards (regarding IT infrastructure) and uses Semantic Web technologies (e.g., RDF and ontologies) as a basis for data representation. Prior to this project, the involved institutes had developed a data infrastructure for research purposes using open standards [31]. Equal to the EuroCAT project, the VATE project has developed an umbrella protocol for rectal cancer [32]. Different from the EuroCAT project, the variables to record are classified into several levels regarding the completeness of datasets and are maintained in a publicly available ontology (http://www.github.com/RadiationOncologyOntology/ROO). The rationale behind these rankings and this public umbrella protocol is that everyone who has data regarding rectal cancer patients can join this linked data network when the data is specified according to the

ontological rules, irrespective to the number of available variables. Due to the chosen aim of training a Bayesian Network for rectal cancer on the VATE infrastructure, missing data could be imputed or ignored during training, as shown by Jayasurya et al. [33].

PCORnet

The Patient-Centered Outcomes Research Network (PCORnet) is a program aiming at building a national research network linking datasets from clinical production systems from multiple centers, using a standardized data platform [34]. The program comprises 11 clinical data research networks (CDRN) and 18 patient-powered research networks (PPRN). The aim of the CDRNs is comparable to the previously described EuroCAT and VATE projects. The PPRN projects aim at the empowerment of patients. In these PPRNs, patients would supply the data instead of retrieving data from clinical systems. Therefore, the gathered data and research questions addressed by these projects are different from the CDRN projects [35]. The first (short-term) aims for the program are to build and implement the network in all the CDRNs and PPRNs and include one million patients in 18 months after the start of the project. Long-term aims are to perform (distributed) machine learning on the network.

Summary

In this chapter, we have seen that multicenter machine learning is possible for both a centralized and distributed approach. To be able to set up a multicenter machine learning environment, several biomedical informatics-related issues need to be addressed. The most important issue is semantic interoperability among participating centers. If the participating centers cannot agree on definitions, how do we know whether all data are equally formatted? Second, the infrastructure (both institutional and central) needs to be implemented, together with the chosen data representation. The choice for an infrastructure comes with the choice of a centralized or distributed approach. Third, privacy preservation needs to be addressed and may influence the choice for a centralized or distributed approach and the preservation measures implemented (e.g., uni- versus bidirectional pseudonymization or data perturbation versus transformation). When all prerequisites are met, the actual machine learning can be performed. In this part, a centralized approach should not be different from traditional machine learning. The distributed machine learning approach needs some modifications to traditional machine learning algorithms, as local outcomes need to be aggregated and combined at a central location. Therefore, in distributed machine learning, traditional algorithms need to be split into two parts: a central node performing the general algorithm and institutional nodes performing delegated tasks requested by the central node. Finally, we have shown that distributed machine learning is possible in practice. Showing sever-

al projects and/or initiatives where data from different locations are used to develop prediction models.

In general, we have shown that distributed machine learning is not only a task for the "traditional" machine learning expert (which is already not the case in healthcare and radiation oncology); however, it also needs other disciplines, such as expertise from the fields of terminology/ontology development, network/infrastructure, and security/privacy.

References

1. 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* 2013;10(1):27–40. doi:10.1038/nrclinonc.2012.196
2. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, *et al.* Rapid-learning system for cancer care. *J Clin Oncol* 2010;28(27):4268–4274
3. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, *et al.* 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol* 2013;109(1):159–164. doi:10.1016/j.radonc.2013.07.007
4. Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. *Radiother Oncol* 2013;108(1):174–179. doi:10.1016/j.radonc.2012.09.019
5. 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* 2012;48(4):441–446. doi:10.1016/j.ejca.2011.11.036
6. Leijenaar RTH, Carvalho S, Velazquez ER, Elmpt WJC van, Parmar C, Hoekstra OS, *et al.* Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability. *Acta Oncol* 2013;52(7):1391–1397. doi:10.3109/0284186X.2013.812798
7. Valentini V, Schmoll HJ, Velde CJH van de, editors. *Multidisciplinary Management of Rectal Cancer: Questions and Answers*. Berlin Heidelberg: Springer-Verlag; 2012
8. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems I: Terminology and typology. *Methods Inf Med* 2000;39(1):16–21
9. World Health Organization, editor. *International statistical classification of diseases and related health problems*. 10th revision, 2nd edition. Geneva: World Health Organization; 2004
10. Sioutos N, Coronado S de, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform* 2007;40(1):30–43. doi:10.1016/j.jbi.2006.02.013
11. Gali A, Chen C, Claypool K, Uceda-Sosa R. From Ontology to Relational Databases. In: Wang S, Tanaka K, Zhou S, Ling TW, Guan J, Yang D qing, et al., editors. *Conceptual Modeling for Advanced Application Domains*, vol. 3289, Springer Berlin Heidelberg; 2004
12. Allemang D, Hendler J. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. 1 edition. Amsterdam ; Boston: Morgan Kaufmann; 2008
13. Berners-Lee T, Hendler J, Lassila O. The semantic web. *Sci Am* 2001;284(5):28–37
14. Brickley D, R.V. G. RDF Schema 1.1. *W3C Recomm* 2014
15. Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. *Int J Semantic Web Inf Syst* 2009;5(3):1–22
16. Prud'Hommeaux E, Seaborne A. SPARQL query language for RDF. *W3C Recomm* 2008;15
17. Murphy SN, Chueh HC. A security architecture for query tools used to access large biomedical databases. *Proc AMIA Symp* 2002:552–556
18. Liu K, Kargupta H, Ryan J. Random Projection-based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining:23
19. Yu S, Fung G, Rosales R, Krishnan S, Rao RB, Dehing-Oberije C, *et al.* Privacy-preserving cox regression for survival analysis, ACM Press; 2008. doi:10.1145/1401890.1402013
20. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. *ACM SIGKDD Explor Newsl* 2009;11(1):10. doi:10.1145/1656274.1656278
21. Hofmann M, Klinkenberg R, editors. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. CRC Press; 2013
22. Ramamohan Y, Vasantharao K, Chakravarti CK, Ratnam ASK. A Study of Data Mining Tools in Knowledge Discovery Process. *Int J Soft Comput Eng* 2012;2(3):4

23. Boyd S. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found Trends® Mach Learn* 2010;3(1):1–122. doi:10.1561/2200000016
24. 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* 2012;19(5):758–764. doi:10.1136/amiajnl-2012-000862
25. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM* 2008;51(1):107. doi:10.1145/1327452.1327492
26. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, *et al.* Architecture of the Open-source Clinical Research Chart from Informatics for Integrating Biology and the Bedside. *AMIA Annu Symp Proc* 2007;2007:548–552
27. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, *et al.* The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. *J Am Med Inform Assoc* 2009;16(5):624–630. doi:10.1197/jamia.M3191
28. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc JAMIA* 2010;17(2):124–130. doi:10.1136/jamia.2009.000893
29. Wiessler W, Dekker A, Nalbantov G, Oberije C, Eble M, Dries W, *et al.* Privacy-preserving, multi-centric machine learning across institutions and countries: does it work?, Geneva: 2013
30. Dekker A, Nalbantov G, Oberije C, Wiessler W, Eble M, Dries W, *et al.* Multi-centric learning with a federated IT infrastructure: application to 2-year lung-cancer survival prediction. *2nd ESTRO FORUM*, *ESTRO FORUM*. Geneva, Switzerland: Elsevier; 2013
31. Roelofs E, Dekker A, Meldolesi E, van Stiphout RGPM, Valentini V, Lambin P. International data-sharing for radiotherapy research: An open-source based infrastructure for multicentric clinical data mining. *Radiother Oncol* 2014;110(2):370–374. doi:10.1016/j.radonc.2013.11.001
32. 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* 2014;112(1):59–62
33. 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: Bayesian network for survival prediction in lung cancer. *Med Phys* 2010;37(4):1401–1407. doi:10.1118/1.3352709
34. Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR. The Greater Plains Collaborative: a PCORnet Clinical Research Data Network. *J Am Med Inform Assoc* 2014;21(4):637–641. doi:10.1136/amiajnl-2014-002756
35. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. *J Am Med Inform Assoc* 2014;21(4):578–582. doi:10.1136/amiajnl-2014-002747

Chapter 6

The Radiation Oncology Ontology (ROO): publishing linked data in radiation oncology using Semantic Web and Ontology techniques

Authors
Alberto Traverso*, Johan van Soest*, Leonard Wee, Andre Dekker
* These authors contributed equally to the manuscript

Adapted from
Medical Physics, published online ahead of print 24 August 2018
DOI: 10.1002/mp.12879

Abstract

Personalized medicine is expected to yield improved health outcomes. Data mining over massive volumes of patients' clinical data is an appealing, low-cost and non-invasive approach towards personalization. Machine learning algorithms could be trained over clinical 'big data' to build prediction models for personalized therapy. To reach this goal, a scalable "big data" architecture for the medical domain becomes essential, based on data standardization to transform clinical data into FAIR (Findable, Accessible, Interoperable and Reusable) data. Using Ontologies and Semantic Web technologies, we attempt to reach mentioned goal.

We developed an ontology to be used in the field of radiation oncology to map clinical data from relational databases. We combined ontology with semantic Web techniques to publish mapped data and easily query them using SPARQL.

The Radiation Oncology Ontology (ROO) contains 1183 classes and 211 properties between classes to represent clinical data (and their relationships) in the radiation oncology domain following FAIR principles. We combined the ontology with Semantic Web technologies showing how to efficiently and easily integrate and query data from different (relational database) sources without a priori knowledge of their structures.

When clinical FAIR data sources are combined (linked data) using mentioned technologies, new relationships between entities are created and discovered, representing a dynamic body of knowledge that is continuously accessible and increasing.

Introduction

Motivation

Data driven medicine has the potential to yield improved health outcomes [1] and is an integral component of value-based healthcare [2]. One of the biggest challenges for data driven medicine is to access and analyze clinical data with machine learning techniques to predict clinical outcomes combining all available information. Subsequently, these developed machine learning techniques can be used to build decision support systems for clinicians.

The current obstacle to be addressed is the availability of outcome information (e.g. tumor control and treatment related toxicity) that must be provided to "train" the machine learning models. Many models have been built based on data from clinical trials. However, clinical trials recruit only a small part of the presenting cases, therefore questions about applicability to under-represented patient sub-groups persist. In contrast, clinical data derived from routine care is known to have data quality issues (e.g. a high rate of missing values).

To overcome the potential sensitivity to missing values as well as to provide sufficient training samples for machine learning, a scalable "big data" architecture for the medical domain becomes essential. For such a scalable architecture, data standardization is imperative. In particular, clinical data should be transformed following FAIR (Findable, Accessible, Interoperable and Reusable) principles [3]. To make healthcare data FAIR, Ontologies and Semantic Web technologies play a key role, and hence will address the issue of semantic interoperability. In [4] the authors exploited the possibility to use ontological technologies to enable semantic interoperability with data coming from multi-center post genomics clinical trials. In [5] the authors focused on applying Semantic Web technologies to the medical imaging domain, developing an ontology for medical image annotations. In [6] the authors investigated the possibility to use Semantic Web technology to store and represent metadata from DICOM image files. Both the studies showed the potential of ontologies technologies in allowing medical data interoperability.

However, the usage of ontologies and Semantic Web technologies applied to the field of radiation oncology are limited. In [7] the authors converted clinical data of prostate cancer patients from a local database using a dedicated ontology, but they did not exploit the possibility to merge different data sets from different diseases combining different ontologies. In [8] the authors stressed the concept of standardization of collected data (in rectal cancer) using ontological techniques to allow machine learning algorithms to build clinical prediction models.

In addition, they strongly suggested using Semantic Web technologies in order to allow data sharing while respecting the privacy protection of individual patients. Finally, ontologies and Semantic Web techniques represent the required infrastructure for distributed learning [9]. Compared to traditional centralized learning approach, in dis-

tributed learning clinical data do not leave the hospital, but after being transformed into FAIR, they are queried during the training of the model, while the model is 'learned' from different centers.

Conversely, when looking at the radiation oncology domain, we could not find any study aiming at: 1) developing and validating a broader ontology to be used in the radiation oncology domain; 2) combining ontology and Semantic Web techniques to transform different clinical databases into FAIR and linked data.

The role of the ROO is to provide a detailed and broad coverage of main concepts used in the radiation oncology domain such as: classification of neoplasms, patients' demographic characteristics; as well as clinical information like tumor's classification or treatment. The ontology is strongly focused on re-using published ontologies and / or terminologies. The added value of the ROO is to re-used published ontologies / terminologies by defining new predicates, which establish relations between imported concepts. Combining different terminologies and expanding relationships between them is the path to guarantee the largest coverage.

The ROO allows transforming unstructured clinical data from following FAIR principles. In particular, data will become:

- Findable (F): each data entity and their properties (F2), translated into universally concept via the ROO will have a globally unique identifier (F1) and will be indexed on the Web (F3). Metadata will include specification of the data identifier (F4)
- Accessible (A): data will be retrievable by mean of RDF triples and querable using a universal language (A1). A permanent de-centralized storage point will be permanently available (A2), even when the original database could not be anymore.
- Interoperable (I): data are represented by universally adopted RDF language (I1). Queries rely on concept from imported ontologies / vocabularies that follow FAIR principles (I2).
- Re-usable (R): several attributes specific data properties and the relations between different concepts via ROO predicates (R1)

In this paper, we: 1) developed a broad ontology to cover the domain of radiation oncology; 2) combined ontology and semantic web techniques to transform clinical data from different disconnected databases into FAIR and linked data, allowing the discovery of new relationships.

Terminologies, Vocabularies and Ontologies

Before going into the details of ontologies' structure and properties we provide the reader with some fundamentals regarding: terminologies, thesauri, vocabularies, and ontologies. Usually, a terminological system [10] is an umbrella terms including the notions of: terminologies, thesauri, vocabularies, and ontologies. Complexity increases from terminologies to ontologies:

- Terminology: a list of term referring to concept within a particular domain. For example, in the radiation oncology domain, concepts such as 'patient' or 'disease'. The terminology can be seen as a list of concepts, but without providing any definition or introducing any structures / relations between the terms
- Thesaurus: a thesaurus is a terminology, where concepts are indexed according to a certain rule (usually alphabetically). Example of a thesaurus is the International Classification of Diseases (ICD), which includes generic-related diagnostic terms (terminology), order alphabetically (thesaurus)
- Vocabulary: in a vocabulary, indexed concepts are accompanied by a definition

Conversely, an ontology is an explicit formal specification of the terms in the domain and relations among them [11] expressed in machine-readable language; therefore, they can be processed automatically. An ontology adds more complexity than a dictionary, since it explicitly defines the relationship, i.e. predicates, between unique entities. Classes (i.e. concepts), sub-classes, and predicates between concepts represent an ontology. Inference rules (also called automated reasoning) in ontologies supply further knowledge, since (new) relationships between concepts, which can be discovered, since not formally defined a priori. An ontology is commonly used to model consensus in understanding a domain between different partners (e.g. different medical centers). Major advantages of ontologies are: a) sharing common understanding; b) re-using of domain knowledge, analyzing domain knowledge, and c) inferring new knowledge starting from relationship between defined concepts. The standard for developing ontologies is the Web Ontology Language (OWL) as recommended by the W3C (World Wide Web Consortium) to represent ontologies [12].

Semantic Web Technologies

Semantic Web is not a separate Web, but an extension of the current one, in which computers primarily interpret the data instead of humans. The current web provides rich-media content (e.g. written text, images, video's,) which is not easy to interpret for computers. In the Semantic Web extension, the information is represented in well-defined structures and semantics in order to enable automated processing of the contents by computers [13]. Hence, it can function as a computer representation of already available web-content, next to the human-readable web content.

For the Semantic Web to function, computers must have access to structured collections of information. The basic building blocks are therefore provided by the Resource Description Framework (RDF) and the "SPARQL Protocol And RDF Query Language" (shorthand: SPARQL). Both RDF and SPARQL build on the existing web components of URIs and HTTP. URIs are the links to the actual resources and can be represented as URLs (e.g. http://mydomain.com/rdf/patient/12345). These URIs are used to represent nodes (resources) and arcs (predicates) in the RDF graph. HTTP is used to publish RDF information on the web or to perform SPARQL queries on RDF stores. These RDF stores

(also called SPARQL endpoints) are webpages which can be queried using the HTTP protocol. Most of these stores/endpoints also have human-readable web interfaces.

By using RDF as a universal graph data structure, the Semantic Web relies on ontologies to give domain-specific structure and interpretation to the represented data. In these ontologies, hierarchies of concepts can be defined, as well as relationships between certain concepts; all written in RDF. It is common practice to add human-readable attributes to the URIs, as it enables the creation of human-readable views on an RDF endpoint. By creating instances of concepts defined in the ontology, users can create graphs of data for representing real-life concepts (e.g. "http://mydomain.com/rdf/patient/12345 rdf:type http://mydomain.com/ontology/patient") where the resource 12345 is an instance of the class patient).

In addition, ontologies can describe inferencing rules which are interpretable by inferencing-enabled RDF stores. In these stores, it is possible to query or show the inferenced information, which is not hard-coded (or materialized) in the RDF store. Hence, updating inferencing rules in the ontology would enable users to query or show additional information without updating the RDF store itself.

This allows to uncover additional relationships in the actual data, and accommodates searches on different levels of data (e.g., patients are persons; therefore, searching for persons will include all patients in the database).

Materials and Methods

Clinical database

We used a clinical database of oncological patients with a diagnosed rectal cancer from the THUNDER trial [14]. The goal of the trial was to develop a prediction model of rectal tumor response after chemo-radiotherapy that might be helpful in individualizing treatment strategies, i.e., selecting patients who need less invasive surgery or another radiotherapy strategy instead of resection. The database includes 80 patients and contains a diverse range information, combining demographic and clinical outcomes. Due to its heterogeneous nature, it represents a good validation for the ROO. The ROO was applied to convert each value in the database, mapping them through the concepts available in the ontology. Relations between individuals were mapped using a graph structure. The graph output was then transformed into RDF triples, published on a dedicated end-point and queryable, in line with FAIR principles.

Radiation Oncology Ontology (ROO) development

We developed a Radiation Oncology Ontology (ROO). The ontology was designed using the editor tool Protégé [15] and publically published at the NCBO BioPortal (https://bioportal.bioontology.org/ontologies/ROO). The ROO adheres to the Ontology

Web Language (OWL) 2 Query Language (QL) profile (http://www.w3.org/TR/owl2-profiles/). The ontology provides basic concepts, relationships, and properties / attributes for radiation oncology.

The ontology was built following this procedure: 1) we identified variables of interest by collecting concepts and their definitions within the ontology using different datasets coming from several institutions belonging to the Euro CAT projects [16]. Due to its multi-center nature, we could allow a broad coverage for different diseases with the aim of making the ontology as much detailed as possible; 2) we published and make publicly available on BioPortal several versions of the ontology during its development. This choice allowed users downloading, using and testing our ontology.

In addition, a dedicated section on GitHub permitted users highlighting inconsistencies and / or requiring enhancements. In this way, our ontology became a dynamic body of knowledge with the aim of guaranteeing the broadest possible coverage for the radiation oncology domain.

The high-level structure of the ROO is based on the Unified Medical Language System (UMLS) Semantic Network by the Semantic Types (classes) ontology (http://bioportal.bioontology.org/ontologies/STY/?p=summary) and the assertion of the Semantic Relations (properties) as specified by the UMLS (https://uts.nlm.nih.gov/).

The ROO re-uses as much as possible entities from other ontologies such as the National Cancer Institute (NCIT) Thesaurus or the International Classification of Disease (ICD) ontologies. The ROO makes only use of ontologies published at NCBO's BioPortal and provided without any restrictions. Common re-used ontologies were: NCIT (National Cancer Institute Thesaurus); Units of Measurement Ontology (UO); Foundational Model of Anatomy (FMA); Semantic Types Ontology (STY); Semantic DICOM Ontology (SEDI), and International Classification of Diseases, Version 10 (ICD10). The ROO uses the original Unique Resource Identifiers (URIs) for these imported entities: an example is the concept of lung cancer that is inherited using the concept code C34 from the ICD ontology.

Using the ontology

Mapping between database schemas and the ontology

One of the most important test to validate the ontology is guarantying that every element in a clinical relational database and its properties can be fully mapped with respectively the concepts and predicates in the ontology. The basic idea of the mapping process is linking each component (row, columns, and values) of the database to its corresponding component (concept, property, relationship) of the ROO. The preliminary step is to identify a correspondence between the columns in the relational database and the ontology entities. A sketch representation of the mapping procedure is shown in Figure 1.

At the top, the hierarchical structure of the ROO is presented in the rectangle A. Hierarchical Relationships ('is subclass of') between classes, are expressed by dotted ar-

rows. These relationships between more general classes (parents) and more specific classes (children) represent the ontology backbone since they allow properties inheritance. ROO concepts are expressed inside blue squares. Relationships between concepts (predicates) are expressed with arrows: they connect classes between each other. For example, patient and gender classes are connected by the property 'has_gender'.

A sample table of one of the datasets is shown in the rectangle B. This table contains information about patient demographics (e.g. sex, age) as well as diagnosis (e.g. survival, tumor staging). The mappings are built between the table columns and the concepts in the ROO (shown as bold dotted double-headed arrows in the figure). For example, the column "Sex" is mapped to the concept gender (ncit:C17357) in the ROO. The link between a patient and the gender is made by the property "has_gender".

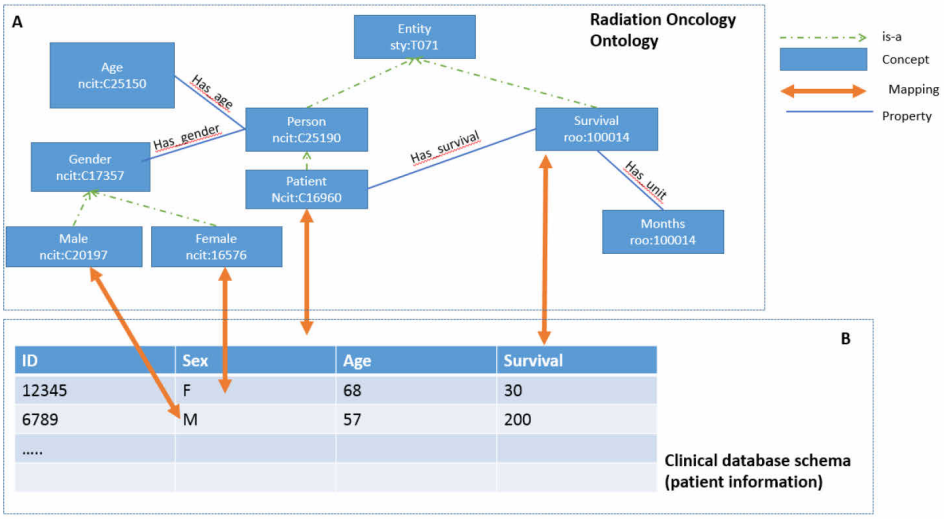


Figure 1 - overview of the ROO structure and the relational database. The hierarchical structure of the ROO is presented in the rectangle A. Hierarchical Relationships ('is subclass of') between classes, are expressed by dotted arrows. Mapping is performed to columns and values in a relational database (rectangle B).

Several languages and software tools are available to perform the mapping procedure from relational databases to RDF triples [17]. We performed the mapping between the clinical data and the ontology using the D2RQ mapping language. D2RQ mapping language is a declarative language for mapping relational database schemas to RDF vocabularies and OWL ontologies. The language is read and interpreted by the D2RQ platform, which is written in Java and open-source available. We decided to use D2RQ because it represents one of the most common tools for database transformation from relational database to network structures [18]. In addition, the language is modular, easily allowing to link entities from the database to concepts and properties in ROO. The mapping defines a virtual RDF graph that contains instructions how to connect and map the information from the relational database. This is similar to the concept of views in SQL data-

bases, except that the virtual data structure is an RDF graph instead of a virtual relational table. The mapping file, written in turtle (.ttl) syntax, contains the mapping between the database schema and the concepts defined in the ontology. The turtle syntax is the format for expressing data as RDF triples, then queryable using a dedicated language.

An example of the mapping file is shown in Figure 2. The mappings between table columns and their corresponding concepts are created using the command d2r:ClassMap. The mapping between the table columns to their corresponding properties is performed by using the command d2rq:PropertyBridge. In addition, in the mapping script each entity is associated with a Unique Resource Identifier (URI) to facilitate publishing on the Semantic Web and data linking. In the example, the entity patient is mapped to the concept C16960 from the NCIT. The bridge between a patient and his / her gender is mapped through the predicate 100018 ("has_gender") from the ROO.

```
# PATIENT TABLE

map:patient a d2rq:ClassMap;
        d2rq:dataStorage map:database;
        d2rq:class ncit:C16960; #patient
        d2rq:uriPattern
"patient_@@derived_multidelineations.identifier@@";
        .

# PROPERTY BRIDGE FOR PATIENT

map:patient_label a d2rq:PropertyBridge;
                d2rq:BelongsToClassMap map:patient;
                d2rq:property rdfs:label;
                d2rq:column
"derived_multidelineations.collection_identifier";
                d2rq:datatype xsd:String;
                .
```

```
#CREATE NOW GENDER OBJECT

map:patient_gender_obj a d2rq:ClassMap;
        d2rq:dataStorage map:database;
        d2rq:uriPattern
"gender_@@derived_multidelineations.identifier@";
        d2rq:condition "derived.gender IS NOT NULL";
        .

# Link to the gender object of the NCI thesaurus

map:patient_gender_uri_obj a d2rq:PropertyBridge;
        d2rq:belongsToClassMap map:patient;
        d2rq:property roo:100018; #has_gender
        d2rq:refersToClassMap map:patient_gender_obj;
        .
```

Figure 2 - example of the D2RQ mapping script

Publishing and querying

The mapped data, transformed into URIs, are then stored in a RDF store, which is web-enabled (HTTP) and can be queried using SPARQL. Making these RDF stores web-enabled means that it is available internally or externally on a specific network, in the same way as webpages are. This does not per definition mean that data is publically available, only that existing web techniques are used to represent semantically interoperable data. In our work, we used Blazegraph (www.blazegraph.com) as our RDF store (or SPARQL endpoint).

Results

Radiation Oncology Ontology (ROO)

The ROO contains 1183 classes, with an average number of four children per class; two classes have more than 25 children. The classes cover the most common concepts in radiation oncology including cancer diseases, cancer-staging systems, and oncology treatments. Besides the classes, 211 predicates are introduced to express relationships between different classes. We divided the properties into five big categories:

- concenptually_related_to
- functionally_related_to
- physically_related_to
- spatially_related_to
- temporally_related_to

Examples of mentioned categories are respectively:

- diagnosed_by
- has_result
- connected_to
- has_location
- follows

All entities and predicates in the ROO have a URI, which can be resolvable as a link, hosted on www.cancerdata.org. A web RDF viewer allows the users inspecting a concept by typing on an internet browser the address www.cancerdata.org/roo/[CODE], where CODE is the code of the ontology entity. For example, the user will type http://www.cancerdata.org /roo/100287 for the predicate "has_pathological_stage". In addition, the users are able to transverse the full tree of the ontology through the Web RDF viewer. The latest version of the ontology has been published on BioPortal, totally Open Source and available for the user to download. The ROO is available is the most common format, including OWL, which can be opened by the users using the software Protégé'.

Using the ontology

Mapping between database schemas and ontologies

A wiki page on how to perform the mapping between relational database schemas and the ontology is publicly available on the GitHub (https://github.com/jvsoest/Data-Integration-Tutorial/wiki/conversionClinicalData). The users can follow the guide to convert part of the Thunder dataset into RDF triples with the ROO using the example scripts provided.

Query formulation

After having mapped the data, it is possible to query them using SPARQL language. Users could query the data without having any prior knowledge of the relational database, since SPARQL queries are based on universal concepts defined by the ontology.

Following the example in the Wiki (https://github.com/jvsoest/Data-Integration-Tutorial/wiki/queryClinicalData), let's suppose we want to search all the patients with rectal cancer and retrieve following information: age at diagnosis, ECOG (Eastern Cooperative Oncology Group) performance status score, clinical TNM stage, pathological TNM status, and prescribed dose in Gray.

The example query is available at https://github.com/jvsoest/Data-Integration-Tutorial/blob/master/queries/queryClinicalData.sparql and it shown in Figure 3a. The system returns all the patients and display the results in the SPARQL result window on the web browser. Each object shown is associated with an URI, universally and unambiguously defining it when published on the Web. Furthermore, all triple patterns to find a certain variable are grouped in curly brackets. This creates the opportunity to make some variables optional or to specify some filters. For example, we could have asked for patients with an age at diagnosis below a certain value, by modifying the original query with a filter (highlighted in blue in Figure 3b).

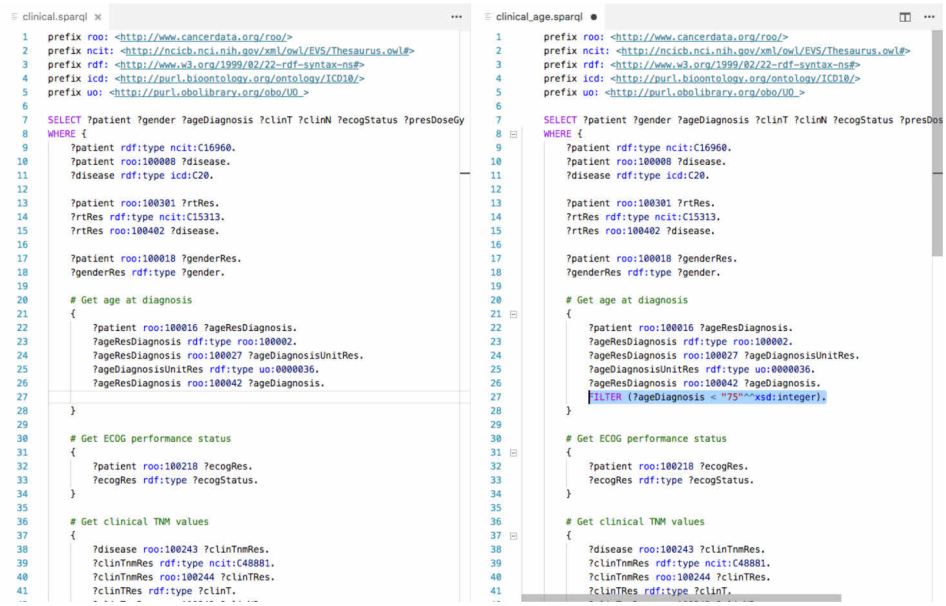


Figure 3 - a) on the left, example query without any filter; b) on the right, example query introducing a filter on the diagnosis age

Finally, this example query can be used directly in programming languages / statistical languages to request valid data matrices. For example, in R using the SPARQL package, or in any other language using a Representational State Transfer (REST) interfacing package. Results from query shown in Figure 3, where compared with data available in the original database to verify the correctness of the mapping. Data comparison and visualization was performed, and no differences were found when comparing the information available in the database with respect to the one available as SPARQL que-

ries. The advantage with respect to a standard excel file, is that RDF data could be queried without any knowledge of the original data structures, by mean of SPARQL queries based on universal concepts defined by the ROO.

Combining different databases: linked data

One of the biggest benefits of Semantic Web and ontology technologies is the possibility to query different databases and make connections within them. For example, in radiation oncology it can be interesting for clinicians to investigate some properties (e.g. survival) of patients: 1) with a certain disease AND b) treated according to a pre-determined protocol AND c) associating the publications of the clinical trial related to the protocol.

Performing such a query using traditional relational databases is a real issue, since it not only requires combining different databases, but also a prior knowledge of their schemas. We solved the problem using ontologies and Semantic Web technologies.

Query formulation

In particular, we made use of Bio2RDF: an open-source project that uses Semantic Web technologies to build and provide the largest network of linked data for the life sciences [19]. It contains among others the RDF versions of ClinicalTrials.gov and PubMed.

In our query, the first part is equal to the query used to retrieve clinical data (see section “Query formulation”). This query retrieves the patients available in the RDF store, and characteristics of these patients (e.g. age, gender, ECOG performance status), their disease (e.g. tumor classification), and the prescribed treatment. Based on this information, we linked the patients to matching treatment protocols, as we defined the protocols and linked them to the correct ClinicalTrials.gov entry in Bio2RDF.

Afterwards, the query contains a section to query the ClinicalTrials.gov linked data representation from Bio2RDF, and a URL generation for a PubMed query. To link the clinical information to public ClinicalTrials.gov (CTgov) information, we used the prescribed treatment variable (containing a unique URL) which was available on both internal (clinical) data, and the Bio2RDF CTgov linked data. From the CTgov linked data, we queried in which trials the same treatment protocol URLs were used. From this relation, we could retrieve information regarding the specific clinical trials, such as the CTgov identifier, the time period when the trial was conducted, which institutes were involved, and trial contact persons. Based on the CTgov identifier, we generated a link to the related manuscripts which have been indexed in PubMed.

The full query to run this linked data example is available at https://gist.github.com/jvsoest/eb015abfb0efd5c669fd36915ce2487d. For example purposes, this query can be executed at http://sparql.cancerdata.org/

Discussion

Rationale for the ROO

Patients' demographics and clinical information are important for radiation oncology prediction / modelling studies. In particular, it is of interest of the radiation oncology community to explore the maximum amount of available clinical data to improve semantic interoperability during patient referral, and for models aiming at predicting outcomes such as overall survival or toxicities after a treatment.

To reach this goal, data integration from different sources (internal / external relational databases) becomes a key factor, since most of the data are usually located in different relational databases.

Since relational databases can present different structures, querying them to access information without having a prior knowledge of the structures becomes a real issue. To tackle this issue, there is the need to transform clinical data following FAIR principles [3]. Ontologies and Semantic Web technologies could represent the right choice to achieve this goal.

We developed the Radiation Oncology Ontology (ROO) with the aim to provide an ontology of use within the radiation oncology field to be used to transform clinical data following FAIR principles.

Advantages of ontologies and Semantic Web data integration compared to relational databases

As presented in previous sections, the ROO has been used to transform clinical traditional database schemas into graph databases relying on ontologies. There are some differences between graph and database schemas. First, ontologies represent a domain on knowledge. Conversely, database schemas are conceived for (and linked to) particular applications, making their structures very diversified and difficult to be made interoperable. In fact, only users knowing the schemas structure (usually the owner of the data) can easily access them.

On the contrary, ontologies transform data into universally concepts that can be queried by the users using SPARQL, without knowing the structures of the data themselves. In fact, data are transformed on universal concepts defined by the ontology itself, and available using URIs (and URLs).

The usage of ontologies adds to transforming data from database schemas into FAIR data. An ontology, combined with Semantic Web technologies, is a stable conceptual interface on top of the relational database system. In fact, it can be scaled for data integration among multiple domains.

Individual database schemas are mapped to the concepts of the ontology and it is relatively easy to integrate new database systems (when mapped / converted into Semantic Web data). The only modification required would be to update the mapping file.

Overall, ontologies increase the semantic interoperability of already available data sources. This outcome has a direct impact on several clinical applications. In particular, it represents the underlying infrastructure for developing multi-center prediction models for clinical outcomes in radiation oncology. In fact, if every medical center transformed their data into FAIR through the ontology, data analytics can be performed on a broader dataset reducing possibilities of over-fitting.

Ontologies and Semantic Web technologies will provide the infrastructure to query in an easy way the data needed by the model. Data will not need to leave the hospital, since being now FAIR, will be queried using SPARQL during the model training / validation. This application, known as distributed learning has been recently presented in literature as a promising application in radiation oncology [16,19,20].

In addition, Semantic Web and ontologies allow connecting different databases. In fact, data are transformed into universal concepts connected between each other: linked data. New relationships between entities are created and discovered, representing a dynamic body of knowledge that is continuously accessible and increasing. In the examples we showed in the result section, we successfully integrated data coming from different sources: clinical databases, clinical trials bank, and scientific literature databases to answer questions of clinical interest.

Finally, semantic databases (e.g. a collection of RDF records) have all the advantages from relational databases. Furthermore, they include the additional possibility to develop machine-readable records (URIs). Recently, we faced a transition from relational databases to semantic databases. The reason is that semantic databases utilize an expanding semantic model that readily incorporates new varieties of data sources, and more easily adjusts to changed requirements as they arise. Subsequently, linking disparate data sets is far easier in a semantic graph setting. In addition, semantic graphs allow to discover hidden relationships between underlying data. In fact, the granular nature of semantics allows to determine relationships between different elements.

Dynamic body of knowledge

We decided to put the ontology publicly available on BioPortal for users to test and validate with the aim of 1) developing a dynamic and growing body of knowledge; 2) guaranteeing the broadest coverage for the radiation oncology data domain.

In addition, the latest version of the ROO is published on the GitHub: users are able to insert enhancements and open issues, making the ontology development a collaborative process.

Limitations

In this work we explored the ontology-based data integration with data from rectal cancer databases. We were able to map all the entities present in the databases with concept and properties from the ROO ontology. However, the ROO should be tested

also on larger databases, other diseases and routine clinical data to check if all the main information could be covered.

In addition, this work did not include a systematic evaluation. Further investigations on evaluating the system performance need to be considered such as comparing the query time between SPARQL and traditional databases.

Future developments

Future development is to extend the ROO to guarantee a broader coverage for an extensive use in the radiation oncology field. In particular, we would like to expand our ontology with:

- Detailed concepts for mapping radiation oncology annotations including organ at risks and nodal involvement
- Detailed concepts for mapping treatment-related concepts and properties such as Dose Volume histograms (DVH)

Furthermore, we want to expand the number of users. In this context, we will continue proposing the ontology as underlying data representation for advanced modelling applications such as distributed learning. In addition, we will try to use the ROO combined with other ontologies under development to combine and link: DICOM information, clinical data and quantitative features computed on patients' images and variables.

Conclusion

We successfully demonstrated that is possible to convert clinical data following FAIR principles using the combination of ontologies and Semantic Web technologies. We developed a broad Radiation Oncology Ontology that can be used in the domain of radiation oncology for data integration.

In addition, we showed how Semantic Web technologies based on developed ontologies allows to efficiently and easily query data from different (relational database) sources without knowing a priori their structures. This result opens the possibility to use ontologies and Semantic Web technologies to further produce and analyze linked data in radiation oncology.

References

1. Aspinall MG, Hamermesh RG. Realizing the promise of personalized medicine. *Harv Bus Rev* 2007; 85(10):108–117, 165
2. Duffy MJ, Crown J. A Personalized Approach to Cancer Treatment: How Biomarkers Can Help. *Clin Chem* 2008;54(11):1770–1779. doi:10.1373/clinchem.2008.110056
3. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. doi:10.1038/ sdata.2016.18
4. Alonso-Calvo R, Perez-Rey D, Paraiso-Medina S, Claerhout B, Hennebert P, Bucur A. Enabling semantic interoperability in multi-centric clinical trials on breast cancer. *Comput Methods Programs Biomed* 2015;118(3):322–329. doi:10.1016/j.cmpb.2015.01.003
5. Rubin DL, Rodriguez C, Shah P, Beaulieu C. iPad: Semantic Annotation and Markup of Radiological Images. *AMIA Annu Symp Proc* 2008;2008:626–630
6. Van Soest J, Lustberg T, Grittner D, Marshall MS, Persoon L, Nijsten B, *et al.* Towards a semantic PACS: Using Semantic Web technology to represent imaging data. *Stud Health Technol Inform* 2014;205:166–170
7. Min H, Manion FJ, Goralczyk E, Wong YN, Ross E, Beck JR. Integration of Prostate Cancer Clinical Data Using an Ontology. *J Biomed Inform* 2009;42(6):1035–1045. doi:10.1016/j.jbi.2009.05.007
8. Meldolesi E, van Soest J, Alitto AR, Autorino R, Dinapoli N, Dekker A, *et al.* VATE: VAlidation of high TEchnology based on large database analysis by learning machine. *Colorectal Cancer* 2014;3(5):435–450. doi:10.2217/crc.14.34
9. Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, *et al.* Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. *Radiother Oncol* 2016. doi:10.1016/j.radonc.2016.10.002
10. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems I: Terminology and typology. *Methods Inf Med* 2000;39(1):16–21
11. Gruber TR. A translation approach to portable ontology specifications. *Knowl Acquis* 1993;5(2):199–220. doi:10.1006/knac.1993.1008
12. Bechhofer S, van Harmelen F, Hendler J, Horrocks I, McGuinness DL, Patel-Schneider PF, *et al.* OWL Web Ontology Language:80
13. Berners-Lee T, Hendler J. Publishing on the semantic web 2001:2
14. van Stiphout RGPM, Valentini V, Buijsen J, Lammering G, Meldolesi E, van Soest J, *et al.* Nomogram predicting response after chemoradiotherapy in rectal cancer using sequential PETCT imaging: A multicentric prospective study with external validation. *Radiother Oncol* 2014;113(2):215–222. doi:10.1016/ j.radonc.2014.11.002
15. Noy NF, Crubézy M, Fergerson RW, Knublauch H, Tu SW, Vendetti J, *et al.* Protégé-2000: An Open-Source Ontology-Development and Knowledge-Acquisition Environment. *AMIA Annu Symp Proc* 2003;2003:953
16. Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, *et al.* Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clin Transl Radiat Oncol* 2017;4:24–31. doi:10.1016/j.ctro.2016.12.004
17. Hert M, Reif G, Gall HC. A comparison of RDB-to-RDF mapping languages, ACM Press; 2011. doi:10.1145/ 2063518.2063522
18. Yunianta A, Barukab OM, Yusof N, Dengen N, Haviluddin H, Othman MS. Semantic data mapping technology to solve semantic data problem on heterogeneity aspect. *Int J Adv Intell Inform* 2017;3(3):161–172. doi:10.26555/ijain.v3i3.131
19. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, *et al.* 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol* 2013;109(1):159–164. doi:10.1016/j.radonc.2013.07.007
20. Naqa IE, Li R, Murphy MJ, editors. *Machine Learning in Radiation Oncology: Theory and Applications*. Springer International Publishing; 2015

Chapter 7

Towards a semantic PACS: Using Semantic Web technologies to represent imaging data

Authors
Johan van Soest, Tim Lustberg, Detlef Grittner, M. Scott Marshall, Lucas Persoon, Bas Nijsten, Peter Feltens, Andre Dekker

Adapted from
Studies in Health Technology and Informatics, 2014, Volume 205, pages 166 – 170
DOI: 10.3233/978-1-61499-432-9-166

Abstract

The DICOM standard is ubiquitous within medicine. However, improved DICOM semantics would significantly enhance search operations. Furthermore, databases of current PACS systems are not flexible enough for the demands within image analysis research. In this paper, we investigated if we can use Semantic Web technology, to store and represent metadata of DICOM image files, as well as linking additional computational results to image metadata. Therefore, we developed a proof of concept containing two applications: one to store commonly used DICOM metadata in an RDF repository, and one to calculate imaging biomarkers based on DICOM images, and store the biomarker values in an RDF repository. This enabled us to search for all patients with a gross tumor volume calculated to be larger than 50 cc. We have shown that we can successfully store the DICOM metadata in an RDF repository and are refining our proof of concept with regards to volume naming, value representation, and the applications themselves.

Introduction

The DICOM standard [1] is ubiquitous within medicine, especially in diagnostic related departments (e.g. radiology, radiation therapy or nuclear medicine). The standard fulfils a twofold function, describing the image (meta)data using a key-value pair system (often called DICOM *tags*), and also as a communication protocol for transmitting this kind of data.

Over the years, the DICOM standard has added support for imaging modalities and steadily more metadata. One of these metadata features is the referencing structure based on unique identifiers (UIDs) to indicate relations between modalities, for instance, to state that a PET-CT image contains two separate image series that were scanned at the same time. Within radiotherapy, the DICOM standard is extended to record the tumor delineations of a physician (an RTSTRUCT), or to define the actual radiation treatment plan (RTPLAN) including dosimetry information (RTDOSE).

The relationships between RTSTRUCT, RTDOSE, and RTPLAN are unidirectional: RTSTRUCT refers to CT/PET/MRI, RTPLAN refers to RTSTRUCT, and RTDOSE refers to RTPLAN. This directionality leads to difficulties when traversing the DICOM modalities, especially in the opposite direction. For example, if a user has a CT scan, and wants to retrieve the radiation treatment plan (RTPLAN), he has to search for the RTSTRUCT object based on the specific CT scan, and from there search for the RTPLAN object based on the RTSTRUCT object. This is an inefficient operation because all RTSTRUCT files, and all RTPLAN files for the patient need to be processed to find the correct treatment plan. Another issue is related to imaging biomarkers. Imaging biomarkers are essential to developing prognostic and predictive models for radiotherapy [2,3]. As this field is evolving rapidly, we need a flexible solution that can be easily extended with new features in order to store biomarkers. We need to store imaging biomarkers while maintaining a link between the biomarker and the images/modalities from which the biomarker was calculated [4]. A direct link between image and biomarker can be established based on the UID of the DICOM file(s). However, additional metadata of the image should then be processed before storage, which is not common practice in a picture archiving and communication system (PACS).

To address the issues described above, we need a flexible storage solution for (additional) metadata, while still being able to define a formal structure to represent data already available in the DICOM standard. One of these flexible storage solutions, different from traditional databases is based on the Semantic Web technology [5]. Our definition of the Semantic Web is that it uses linked data and standardized terminologies (ontologies), therefore making it possible for everyone to contribute to a web of knowledge [6]. This linked data can be stored in a Resource Description Framework (RDF) repository, which can be viewed as a single table with three columns (subject, predicate and object) and where every row is called a triple. When we unify the object of a triple with its equal in another triple's subject, we build a graph of the data. This

RDF repository can be queried using a specific query language: SPARQL (SPARQL Protocol and RDF Query Language), which performs pattern matching on the available graph of data [7].

In this study, we hypothesize that we can use Semantic Web technology to store and represent metadata of DICOM image files and additional computational results linked to the image metadata.

Methods

To test our hypothesis, we separated the actions into three different tasks. The first task addressed the storage and representation of DICOM metadata in an RDF repository. The second task targeted the storage of additional computational results, related to the images. The third task was to implement a proof of concept.

Storage and representation of DICOM metadata

To store the data in an RDF repository (and make it accessible via a SPARQL endpoint), we first developed an ontology representing the most common DICOM metadata elements. This ontology serves as a definition of our data domain, and as a contract regarding the structure and data elements if anyone wants to contribute or query our data [8]. Our ontology can be found at the NCBO Bioportal [9] ontology browser as "Semantic DICOM Ontology" [10]. Although the ontology is still under development, the basic concepts/structure is stable, as it is based on the DICOM standard itself.

This ontology is used by our Semantic DICOM (SeDI) conversion service. This service is written in Java, using the dcm4che library [11]. For every DICOM file, the application will check whether specific tags defined in the ontology are present in the DICOM file, and creates RDF triples based on the rules defined in the ontology. As our conversion service keeps the basic ontological rules (e.g. regarding domain and range) in mind, the data representation will follow our ontological structure. After creating RDF triples, the triples are exposed via a SPARQL endpoint.

Storage of additional computation results

To calculate additional imaging biomarkers, we developed a pipeline application, which also accepts data using the DICOM communication protocol and uses a plug-in architecture for different computational modules. This pipeline application is developed in Java using the dcm4che [11] and Spring Dependency Injection [12] libraries.

Computational modules can be developed outside the direct scope of our pipeline application as with an API for calculation modules, every type of calculation can be plugged into the pipeline (when wrapped with Java code). This enhances the flexibility to add computations when needed.

Afterwards the results can be exported to different types of storage, including an RDF repository. When using the RDF export module (which triggers a SPARQL update/insert query) the application tries to execute a pre-defined query (containing placeholders for calculated values) on a given RDF repository, and can insert/update triples. Users can create their own insert/update query and are not bound to a specific ontology, which leaves room for end-users to commit the results to the ontology of their choice.

Proof of concept: query patients and their tumor volume

To test our hypothesis, we implemented a proof of concept setup (Figure 1) where the results of two systems were stored in separate RDF repositories (Figure 1). We used 10 randomly selected patients form our PACS (with different tumor locations) which were sent to both the SeDI conversion service, and the computation pipeline. Both of the applications and triple stores resided on computers in different countries.

To structure the output of our computational pipeline, we used the Radiation Oncology Ontology (under development), which is able to extend the Semantic DICOM ontology for the radiotherapy domain. For our hypothesis test, we linked our calculation output to the identification of the delineated volume of interest (representing the gross tumor volume), stored in the RDF repository of the SeDI conversion service.

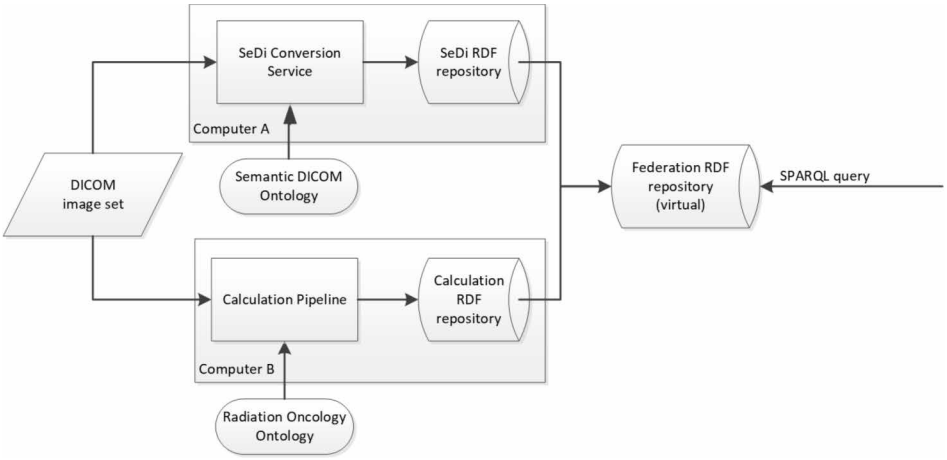


Figure 1 - Setup proof of concept

Results

We successfully parsed the DICOM tags specified in the ontology and stored the resulting triples in our RDF repository. The created concepts were structured according to the DICOM hierarchy of patients, studies, series and images, as specified in the SeDI ontology.

For 6 out of 10 patients, our computation pipeline could calculate the gross tumor volume. 4 out of 10 patients resulted in no calculation result, because the gross tumor volume was not named correctly using our local clinical standard for volume descriptions.

When performing a query on the federated endpoint for patients with a tumor volume larger than 50 cc, 5 patients were returned. When querying for all patients and tumor volumes, the query resulted in all 6 patients, where one patient had a volume of 2 cc, and was therefore not returned in our primary query results.

Discussion

We successfully implemented a proof of concept to store the semantics of imaging data important to biomarker research, and were able to use it to retrieve data which was more extensive and enriched in comparison to information available in a PACS database. Therefore, our results confirmed our hypothesis.

The storage of DICOM information in an RDF repository has an advantage over generic solutions where PACS databases store the entire (or very small parts of the) header. Such PACS relational databases are less flexible in changing data structures (and data), where the Semantic Web technology is more flexible in nature. For instance, within an RDF repository, adding extra information at a later point in time does not alter or change the existing data. Furthermore, it creates the possibility for everyone to link their own additional findings to the already available datasets. Therefore, we expect that the Semantic Web technology will provide a more suitable framework for research purposes, where questions regarding data (and/or calculated features) are still changing. Such a framework can increase the possibilities for data mining, especially when resources are linked to eCRF or EHR data stored in other RDF repositories which are accessible as linked data and through SPARQL endpoints.

We still plan to address several issues. At first, 4 out of 10 patients did not have a proper name describing the gross tumor volume (within the DICOM metadata). This description is a free-text field in many delineation applications; therefore we cannot expect that delineation descriptions are always correct. We need a proper (ontological) mapping to determine relevant delineation descriptions. The RTOG made a first attempt to standardize delineation naming [13], however it would take several years to reach compliance to this standard. Second, we have created two different applications, which operate without coordination. In future work, we plan to implement a messaging system where the SeDI conversion service will notify the computation pipeline that the

DICOM metadata has been stored in a RDF repository, and that computation on this image set can start. Third, although we are able to store the data in an RDF repository, we have not made use of the Value Representation (VR) types available in the DICOM metadata. Currently, all values are stored as text, instead of using the Value Representation (VR) type. This will be solved in our ontology and incorporated into our SeDI conversion service.

We have shown that a semantic representation of DICOM related information enables us to query for image features that are important to biomarkers in radiotherapy. Our work towards a semantic PACS for radiotherapy can eventually be generalized to store and retrieve images using the semantics of other domains.

References

1. The National Electronical Manufacturers Association. *Digital Imaging and Communications in Medicine (DICOM)*. NEMA Publications; 2011
2. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, *et al.* Rapid-learning system for cancer care. *J Clin Oncol* 2010;28(27):4268–4274
3. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, *et al.* 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol* 2013;109(1):159–164. doi:10.1016/j.radonc.2013.07.007
4. Levy MA, Freymann JB, Kirby JS, Fedorov A, Fennessy FM, Eschrich SA, *et al.* Informatics Methods to Enable Sharing of Quantitative Imaging Research Data. *Magn Reson Imaging* 2012;30(9):1249–1256. doi:10.1016/j.mri.2012.04.007
5. Berners-Lee T, Hendler J, Lassila O. The semantic web. *Sci Am* 2001;284(5):28–37
6. Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. *Int J Semantic Web Inf Syst* 2009;5(3):1–22
7. Allemang D, Hendler J. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. 1 edition. Amsterdam ; Boston: Morgan Kaufmann; 2008
8. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems I: Terminology and typology. *Methods Inf Med* 2000;39(1):16–21
9. Rubin DL, Moreira DA, Kanjamala PP, Musen MA. BioPortal: A Web Portal to Biomedical Ontologies:4
10. Grittner D, Van Soest J, Lustberg T, Marshall MS, Dekker A. Semantic DICOM Ontology 2015. http://bioportal.bioontology.org/ontologies/SEDI [accessed March 27, 2018]
11. dcm4che. https://www.dcm4che.org/ [accessed February 2, 2014]
12. Spring Framework. http://spring.io/ [accessed February 2, 2014]
13. Santanam L, Hurkmans C, Mutic S, van Vliet-Vroegindeweij C, Brame S, Straube W, *et al.* Standardizing Naming Conventions in Radiation Oncology. *Int J Radiat Oncol Biol Phys* 2012;83(4):1344–1349. doi:10.1016/j.ijrobp.2011.09.054

Chapter 8

Radiation oncology terminology linker: A step towards a linked data knowledge base

Authors
Tim Lustberg, Johan van Soest, Peter Fick, Rianne Fijten, Tim Hendriks, Sander Puts, Andre Dekker

Adapted from
Studies in Health Technology and Informatics, 2018, Volume 247, pages 855 – 859
DOI: 10.3233/978-1-61499-852-5-855

Abstract

Performing image feature extraction in radiation oncology is often dependent on the organ and tumor delineations provided by clinical staff. These delineation names are free text DICOM metadata fields resulting in undefined information, which requires effort to use in large-scale image feature extraction efforts. In this work we present a scale-able solution to overcome these naming convention challenges with a REST service using Semantic Web technology to convert this information to linked data. As a proof of concept an open source software is used to compute radiation oncology image features. The results of this work can be found in a public Bitbucket repository.

Introduction

Outcome registration is of vital importance to move towards a rapid learning healthcare system for cancer patients [1]. It takes months to see the effect of the radiation or chemotherapy treatment on cancer and years to estimate the severity of some side effects. Together with the outcome data, the radiation oncology treatment features provide the knowledge which is needed to innovate. However, in daily clinical practice, this information is not stored as structured data, they need to be computed using the DICOM objects created during radiation treatment planning.

The DICOM standard does not enforce a terminology for creating delineations of organs in the radiation treatment planning process because the labels are free text. There are efforts in the field of radiation oncology to standardize the delineation practice [2]. However, clinics often have their own implementation of such a standard. Earlier studies have shown that the routine clinical data is a valuable source of information if you can get through the issues of missing, ambiguous and contradicting data [3–5].

To overcome the challenges of messy clinical data and automatically generate radiation oncology treatment features, a software platform is needed. The local naming terminology for the organ and tumor delineation needs to be linked to a terminology understood by the computational platform. Linking this data to other data sources, such as Semantic DICOM (SeDI) [6] and the Radiation Oncology Ontology (ROO) [7] creates a knowledge base to facilitate future studies to predict treatment outcome for cancer patients treated with radiation.

The aim of this study is to provide such a platform as an expandable open source software library that can be used to generate and store radiation oncology treatment features as linked data with minimal user interaction.

Methods

In previous work, Semantic Web technology was used to enhance the query-ability of the information stored in the DICOM metadata, using the Semantic DICOM software and ontology (http://semantic-dicom.org). Linking the DICOM header data to radiation oncology treatment features enabled quick access to for instance CT-scans with a certain tumor size quickly. One of the limitations of this study was the manual work required to pre-process the tumor volumes. In this work, we added three features to eliminate these limitations: Terminology mapping service to link delineation naming terminologies, an ontology to link all the results together and a flexible computational platform for calculating imaging features.

Terminology mapping service

In the radiation oncology community, there are efforts to standardize the organ and tumor delineation naming conventions. However, when investigating routine clinical data, the delineation names will often deviate from the convention due to for instance: typographical errors, interpretation of a convention or the preference to use a native language in non-English speaking countries. To make this data machine-interpretable these local terminologies need to be mapped to a terminology that the computational platform understands.

The terminology Mapping Service (Figure 1) provides a REST API to solve this problem. The REST API uses an SQL database to store the terminology mappings and the governance of these mappings, recording the user and institute that created the mapping. Any application can use this service to store and retrieve synonyms to enable an image analysis application to choose the correct delineations when computing image features.

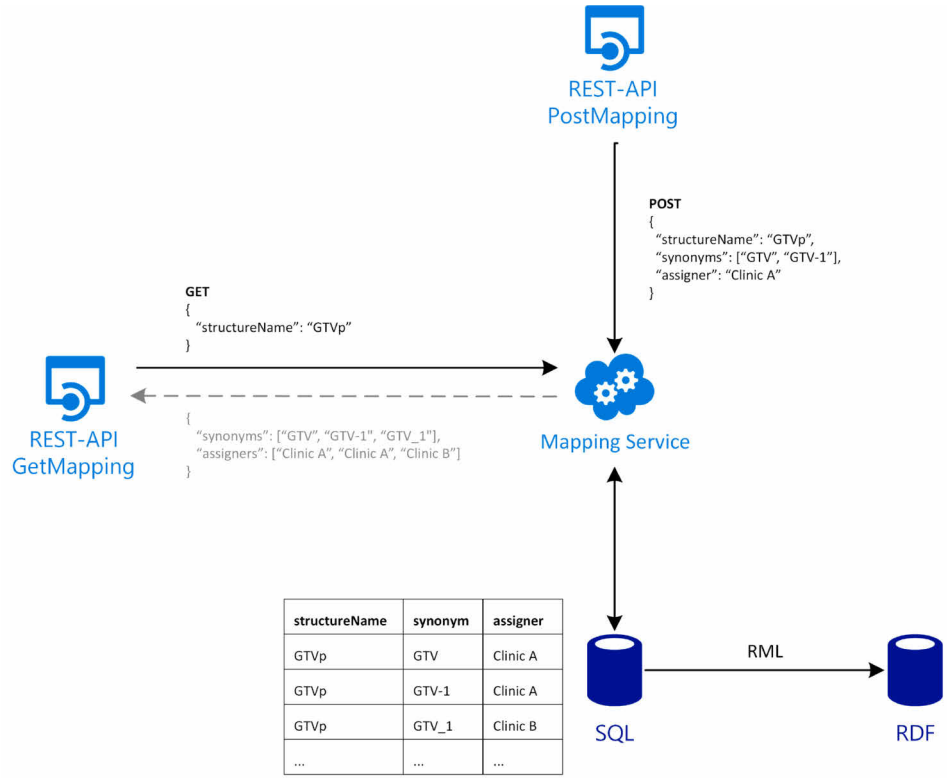


Figure 1 - Mapping Service overview. Communicating with the REST API using JSON data transfer objects. The mappings are converted into linked data using R2RML

Medical Image Analysis Platform

As a continuation of the feature calculation software used in the Semantic DICOM project, the Medical Image Analysis platform was created. The goal was to make the software more scalable and easier to use. To accomplish this, the software was reimplemented with a micro service design, based on the Netflix OSS framework (https://netflix.github.io/), where every functional part is its own REST service (Figure 2). The example calculations (worker) are implemented in Matlab 2015b with an object-oriented design to extract image features from radiation oncology DICOM objects. However, the framework was designed to compute features using any image feature extraction software in any programming language.

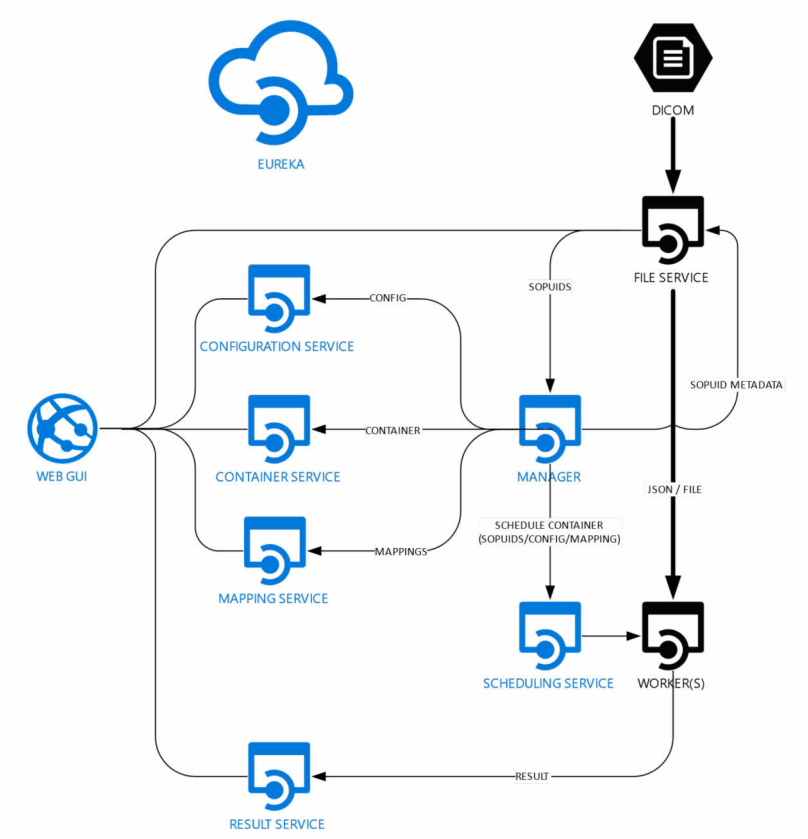


Figure 2 - MIA overview. A micro-service platform to manager the steps (services) needed for image feature extraction using a web-based user-interface.

Linked data representation of image features using Semantic Web technologies

To create the linked data knowledge base needed to answer complex questions the data generated by the software components described in this work, we added RDF Mapping Language (https://www.w3.org/TR/r2rml/) conversions scripts to create triples from the SQL data following the Radiation Oncology Ontology (Figure 3). This reproduces the result of the SeDI project where the image features are linked to the DICOM metadata to answer more complex questions a regular PACS system cannot [6].

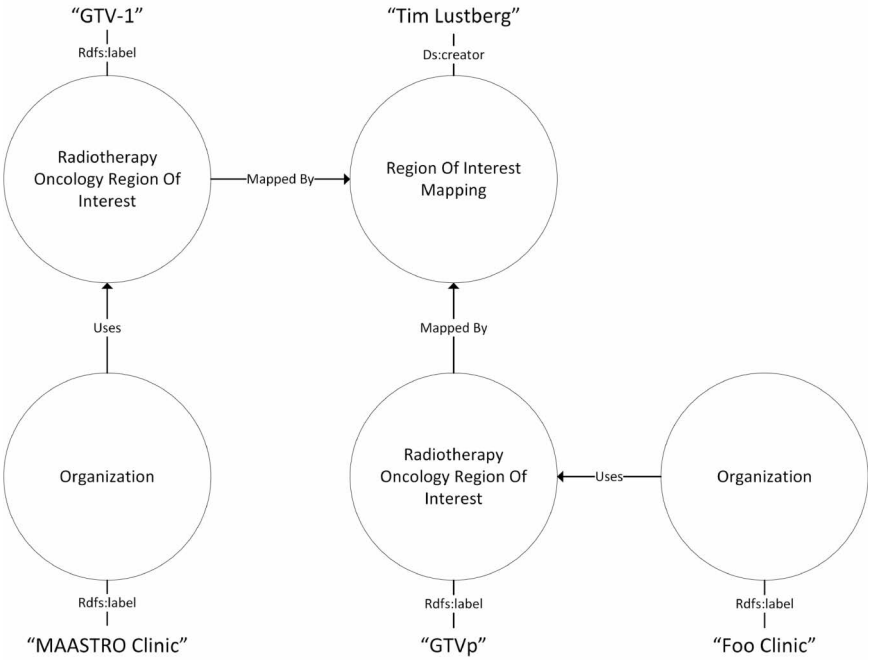


Figure 3 - Structure name mapping ontology based on the Radiation Oncology Ontology

Results

The software is available as on Open Source project on a public Bitbucket repository (https://bitbucket.org/account/user/maastrosdt/projects/MIA). The current release of the MIA and the Mapping Service works as described in the methods and is able to handle large volumes of DICOM data to quickly extract image features.

Discussion

The Mapping Service provides an easy API for software developers to work with when creating a terminology mapping application. Because the problem we wanted to solve in this work was to compute clinical image delineations labelled with free text, the Mapping Service was used to achieve this goal. However, in future implementation efforts this service could be reused to map any terminology. Using new R2RML scripts to link these mappings to an ontology could be of interest in other fields.

On top of the analysis uncertainties, there is a far bigger concern when computing image features in multiple centers based on human delineations. It has been reported that delineation practices vary between radiation oncologists in the same hospital using the same guidelines and even more so when comparing delineations from multiple centers.

In the future, we would like to create a system where the Mapping Service will become obsolete, as delineation labels will be based on a proper terminology and every clinician in the world used the same guidelines. However, we have to acknowledge that this might never be achieved because it requires a level of consensus which in rarely reached in medicine. The linked data systems presented in this work provide a scalable platform to deal with these imperfections to create a knowledge base for further research.

The Mapping Interface included with the MIA could be improved in numerous ways. Using the actual imaging data attached to the delineation names could provide a knowledge base for the software to learn what an organ looks like and how organs are positioned relative to each other. This information could be used to for instance identify the heart if the two delineations of the two lungs are known. This would further decrease the need for human intervention. Another improvement would be to show the images with the delineations on the Mapping Interface. This would make it easier for the user to identify the correct delineation when mapping it to a terminology, as opposed to guessing based on the names provided.

References

1. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, *et al.* 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol* 2013;109(1):159–164. doi:10.1016/j.radonc.2013.07.007
2. Santanam L, Hurkmans C, Mutic S, van Vliet-Vroegindeweij C, Brame S, Straube W, *et al.* Standardizing Naming Conventions in Radiation Oncology. *Int J Radiat Oncol Biol Phys* 2012;83(4):1344–1349. doi:10.1016/j.ijrobp.2011.09.054
3. Dekker A, Vinod S, Holloway L, Oberije C, George A, Goozee G, *et al.* Rapid learning in practice: A lung cancer survival decision support system in routine patient care data. *Radiother Oncol* 2014;113(1):47–53. doi:10.1016/j.radonc.2014.08.013
4. Lustberg T, Bailey M, Thwaites DI, Miller A, Carolan M, Holloway L, *et al.* Implementation of a rapid learning platform: Predicting 2-year survival in laryngeal carcinoma patients in a clinical setting. *Oncotarget* 2016;7(24):37288–37296. doi:10.18632/oncotarget.8755
5. Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, *et al.* Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clin Transl Radiat Oncol* 2017;4:24–31. doi:10.1016/j.ctro.2016.12.004
6. Van Soest J, Lustberg T, Grittner D, Marshall MS, Persoon L, Nijsten B, *et al.* Towards a semantic PACS: Using Semantic Web technology to represent imaging data. *Stud Health Technol Inform* 2014;205:166–170
7. Dekker A, Van Soest J, Traverso, Alberto. *Radiation Oncology Ontology*. 2017

Chapter 9

Validation of a rectal cancer outcome prediction model with a cohort of Chinese patients

Authors
Lijun Shen, Johan van Soest, Jiazhou Wang, Jialu Yu, Weigang Hu, Yutao U. T. Gong, Vincenzo Valentini, Ying Xiao, Andre Dekker, Zhen Zhang

Adapted from
Oncotarget, 2015, Volume 6, Issue 35, Pages 38327 – 38335
DOI: 10.18632/oncotarget.5195

Abstract

The risk of local recurrence (LR), distant metastases (DM) and overall survival (OS) of locally advanced rectal cancer after preoperative chemoradiation can be estimated by prediction models and visualized using nomograms, which have been trained and validated in European clinical trial populations. Data of 277 consecutive locally advanced rectal adenocarcinoma patients treated with preoperative chemoradiation and surgery from Shanghai Cancer Center, were retrospectively collected and used for external validation. Concordance index (C-index) and calibration curves were used to assess the performance of the previously developed prediction models in this routine clinical validation population. The C-index for the published prediction models was 0.72 ± 0.079, 0.75 ± 0.043 and 0.72 ± 0.089 in predicting 2-year LR, DM and OS in the Chinese population, respectively. Kaplan-Meier curves indicated good discriminating performance regarding LR, but could not convincingly discriminate a low-risk and medium-risk group for distant control and OS. Calibration curves showed a trend of underestimation of local and distant control, as well as OS in the observed data compared with the estimates predicted by the model.

In conclusion, we externally validated three models for predicting 2-year LR, DM and OS of locally advanced rectal cancer patients who underwent preoperative chemoradiation and curative surgery with good discrimination in a single Chinese cohort. However, the model overestimated the local control rate compared to observations in the clinical cohort. Validation in other clinical cohorts and optimization of the prediction model, perhaps by including additional prognostic factors, may enhance model validity and its applicability for personalized treatment of locally advanced rectal cancer

Introduction

Colorectal cancer (CRC) is one of the three leading causes of cancer mortality worldwide and approximately one-third of the cancers arise in the rectum. The annual incidence of rectal cancer in China exceeds 200,000 cases and about 40% patients have stage III disease at the time of diagnosis. For locally advanced rectal cancer, preoperative radiotherapy combined with a fluoropyrimidine followed by total mesorectal excision (TME) is the current treatment standard [1]. However, the improvement in locoregional control observed in several randomized trials has not translated into improved survival [2–4]; the development of distant metastases is now the predominant cause of failure in locally advanced rectal cancer [5].

As many as 20% of patients may have a complete pathologic response after preoperative chemoradiation and have a very good prognosis, but the remainder of the patients have a higher risk for local recurrence and/or distant metastasis after treatment [6,7]. We believe that rectal cancer treatment strategies should be personalized according to the expected outcome. Rectal cancer patients, for whom a poor outcome is expected, may benefit from intensified local or systemic treatment. In contrast, patients with an expected pathologic complete response (pCR) after preoperative chemoradiation may be considered for organ-preserving nonsurgical treatment strategies such as wait-and-see [8–11].

To personalize treatment, validation of prediction models is needed to create an evidence base for treatment decisions [12,13]. Several nomograms for predicting follow-up outcome for colorectal cancer have been proposed, but models for locally advanced rectal cancer are scarce [14–16]. Valentini et al. [17] developed prediction models (visualized using nomograms) for locally advanced rectal cancer patients treated with long-course chemoradiation (CRT) followed by surgery, based on data from large randomized trials. The discriminative capability of these nomograms was assessed during external validation, and was determined by measuring the concordance index (C-index) of local recurrence, distant metastases and overall survival (0.68, 0.73 and 0.70 respectively) after 5 years of follow-up. As this study performed an external validation on clinical trial data, it is unknown whether the findings are generalizable to an Asian routine patient population.

The aim of this study is to test the hypothesis that the published model to classify the probability of survival for locally advanced rectal cancer, developed by Valentini et al. [17], is generalizable to a routine clinical dataset from an Asian cancer center.

Materials and methods

Patients

Between March 2006 and December 2012, a consecutive series of 338 patients with MRI/CT staged locally advanced (cT3–4 and/or cN1–2) rectal adenocarcinoma underwent long-course conventional chemoradiation (total dose was between 45–55Gy and mean dose was 50Gy) in daily fraction from Monday to Friday with a concomitant 5-fluorouracil-based chemotherapy and surgery in Shanghai Cancer Center. Patient records were retrospectively extracted from the clinical databases including electronic medical record and treatment planning system. All patients underwent a physical examination before neoadjuvant therapy, including digital rectal examination, and flexible endoscopy; computed tomographic (CT) scans of the chest, abdomen; and CT and/or magnetic resonance imaging (MRI) of the pelvis. Sixty-one patients were excluded for the following various reasons. Twelve patients were diagnosed with a non–skin cancer within 5 years of the diagnosis of rectal cancer. Twenty-seven patients did not undergo radical rectal resection and 2 patients did not complete radiation. Four patients had a metastatic disease before or at the time of surgery. For 16 patients the adjuvant chemotherapy information or pathological results were not available. Finally, a total of 277 patients with complete clinicopathological information were included in this study.

Patients were followed-up every 3–6 months during the first 2 years, every 6 months in the later 3 years, and after 5 years only once every year. Follow-up evaluation consisted of physical examination, imaging examinations, endoscopic study and laboratory examination. Local relapse was defined as tumor recurrence within the pelvis while distant metastasis was defined as out of the pelvis. All relapse cases were diagnosed either by histology or imaging exams. Overall survival was defined as the time difference between date of diagnosis and death from any cause in this study. Recurrence-free survival and metastases-free survival were defined as the time from the start of RT to local recurrence or distant metastasis. The protocol was approved by the hospital's Medical Ethics Committee.

Statistical analysis

The variables required for the nomograms were gender, age, clinical tumor stage, radiotherapy dose, concomitant chemotherapy, surgery procedure, pathological tumor stage, pathological nodal stage and adjuvant chemotherapy. For each patient, the 2-year predicted probability of local control, distant control and overall survival was calculated using the previously published prediction model [17]. The model we used in this manuscript was come from the training cohort of the original paper. We used the same cohort to calculate the prediction performances for training dataset and used the full

routine clinical dataset as validation dataset. In this previous publication, the authors trained a Cox proportional hazards model, and validated the binary outcomes for local recurrence, distant metastasis and overall survival at 5-years of follow-up. This means that the baseline hazard for a follow-up of 5 years was used. In this study, we evaluated the outcomes at 2 years of follow-up. Therefore, we used the baseline hazard for a follow-up of 2 years.

For the nomograms, this means that the sum of scores (the sum of all scores per prediction parameter) stays the same, but the probability related to the sum of scores value changed. To make a fair comparison, we recalculated the concordance index (C-index) on the external validation dataset used by Valentini et al. The C-index is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant. The C-index measures predictive information derived from a set of predictor variables in a model, and is identical to the area under a "receiver operating characteristic" (ROC) curve for binary outcomes [18]. This AUC value represents the ability of a model to assign a higher probability to positive outcomes (e.g. survival). Calibration refers to the agreement between observed outcomes and predictions. Perfect predictions should be on the 45-degree line in calibration curve (the ideal calibration) [18]. Using the given cut-off values (that is, for local recurrence, the two probability thresholds were 8% and 20% while for distant metastasis and overall survival, the thresholds were 15% and 25%), patients were grouped in good, medium and poor prognosis groups [17].

The clinicopathologic characteristics of modelling and validation cohorts were analyzed by a chi-square test or Fisher's exact test. Clinical survival outcomes were assessed using the Kaplan-Meier survival curves, and prognostic groups were compared with the log-rank test implemented in SPSS version 18.0 (SPSS, Chicago IL). A two-sided *p*-value of less than 0.05 was considered statistically significant.

To reduce selection bias and improve validation reliability, we bootstrapped (R = 1000) the predicted and observed outcomes and calculated the C-index for each bootstrap sample. Afterwards, we used the mean C-index from all bootstrap samples as the final C-index for our validation. Calibration curves were applied to assess the agreement between observed outcomes and predictions [19]. The C-index and calibration curves were implemented in Matlab version 7.1 (MathWorks, Natick, MA).

Results

Patient characteristics

The distribution of clinicopathologic characteristics between the European training cohort and the Chinese clinical validation cohort were shown in Table 1. The median age of the validation cohort was 56 years old. The median follow-up was 26 months

(ranging from 3 to 87 months). Of the 277 patients, 20 (7.2%) patients developed local recurrence, 57 (20.6%) patients developed distant metastasis and 42 (15.2%) patients died during follow-up. There were significant distribution differences between two cohorts except the sex distribution.

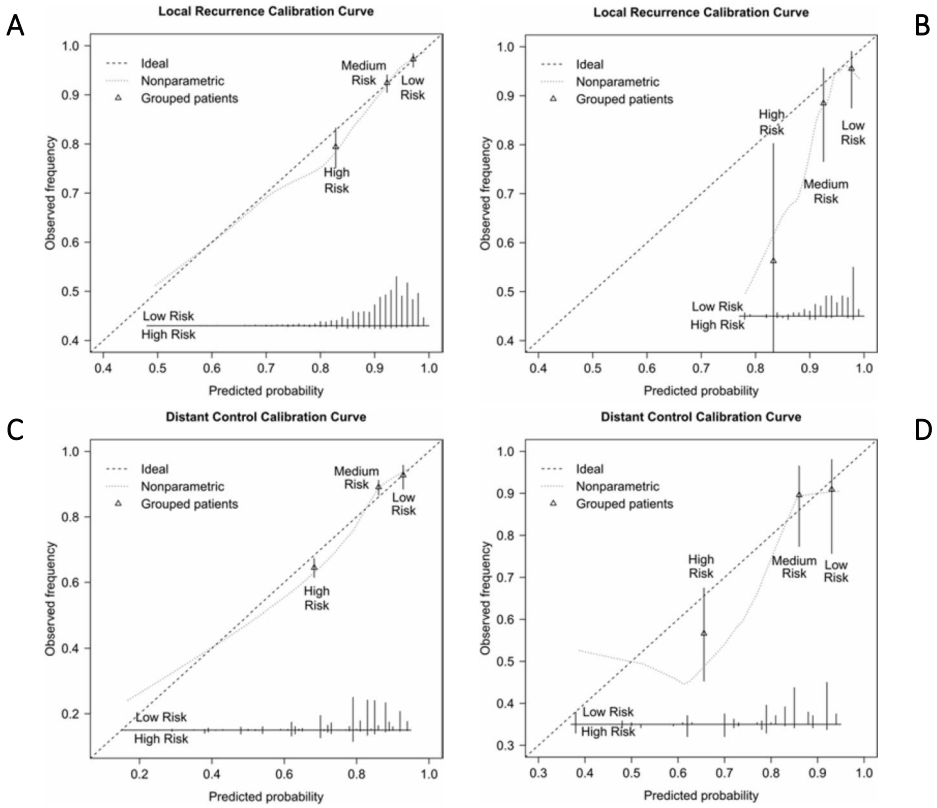
Table 1 - Patient characteristics of the European training (N=2795) and current clinical routine validation (N=277) cohorts

Variable	Training Cohort (N = 2235)	Current validation cohort (N = 277)	p-value
Sex			0.780
Male	1575 (70.5)	198 (71.5)	
Female	660 (29.5)	79 (28.5)	
Age,years			<0.001
≤49	290 (13.0)	107 (38.6)	
50–59	607 (27.2)	78 (28.2)	
60–69	917 (41.0)	69 (24.9)	
≥70	421 (18.8)	23 (8.3)	
Tumor location			<0.001
Low	786 (35.2)	116 (41.9)	
Mid	1146 (51.3)	158 (57.0)	
High	303 (13.6)	3 (1.1)	
cT stage			<0.001*
2	18 (6.1)	9 (3.2)	
3	1887 (84.4)	231 (83.4)	
4	193 (8.6)	37 (13.4)	
Treatments			
Radiotherapy dose, Gy			<0.001
<45[†]	95 (4.3)	20 (7.2)	
45	1558 (69.7)	39 (14.1)	
>45	582 (26.0)	218 (78.7)	
Concomitant chemotherapy			<0.001
No	862 (38.6)	5 (1.8)	
Yes	1373 (61.4)	272 (98.2)	
Surgery procedure			<0.001
LAR	1373 (61.4)	100 (36.1)	
APR	862 (38.6)	177 (63.9)	
Adjuvant chemotherapy			<0.001
No	836 (37.4)	15 (5.4)	
Yes	1399 (62.6)	262 (94.6)	
ypT stage			<0.001
0	205 (9.2)	64 (23.1)	
1–2	810 (36.2)	88 (31.8)	
3	1163 (52.0)	119 (43.0)	
4	57 (2.6)	6 (2.2)	

Nomogram validation

When re-validating the previous prediction models from the training cohort of the original paper for at 2 years of follow-up, we found C-indexes of 0.72 ± 0.021, 0.73 ± 0.013 and 0.68 ± 0.018 for the prediction of local recurrence, distant metastasis and overall survival, respectively. The evaluation in the Chinese clinical routine population achieved C-indexes of 0.72 ± 0.079, 0.75 ± 0.043 and 0.72 ± 0.089, respectively.

Figure 1 showed the calibration curves comparing the observed outcomes with the predicted LR, DM and OS probabilities in the original training cohort and the current validation cohort. Calibration curves in original training cohort suggested that the prediction models were well calibrated (Fig. 1A, 1C and 1E). Although the trend of observed incidence was lower than the predicted incidence in Chinese cohort, calibration curves were not different from ideal since the error bars touched the ideal line. The bars represent 95% confidence interval suggesting that we are for 95% sure that calibration group is similar to observation group. This means the nomograms have a trend to overestimate the probabilities of local and distant control, as well as overall survival in the validation population, especially in the high risk group.



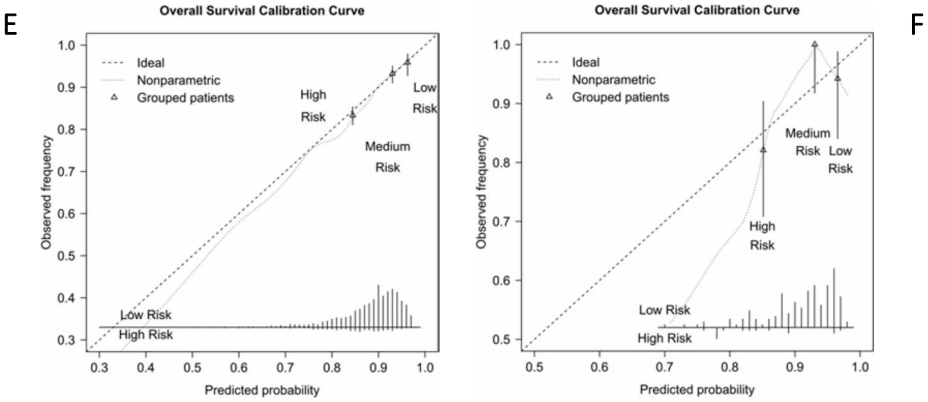


Figure 1 - Calibration curves for prediction models in European trianing cohort and Chinese validation cohort. A. Calibration curve for local control in training cohort. B. Calibration curve for local control in current validation cohort. C. Calibration curve for distant control in training cohort. D. Calibration curve for distant control in current validation cohort. E. Calibration curve for overall survival in training cohort. F. Calibration curve for overall survival in current validation cohort.

Figure 2 showed the Kaplan-Meier curves stratifying patients into low, medium or high risk groups based on predicted outcome. This figure showed that the local control prediction model had discriminative power to stratify three risk groups (high, medium, low; $p = 0.002$), but it should be noticed that only 32 patients were assigned to the high risk group. The distant control and overall survival prediction models were able to significantly discriminate between the high-risk group and the medium- or low-risk groups ($p < 0.001$ and $p < 0.001$, respectively) but couldn't discriminate between medium- and low-risk groups.

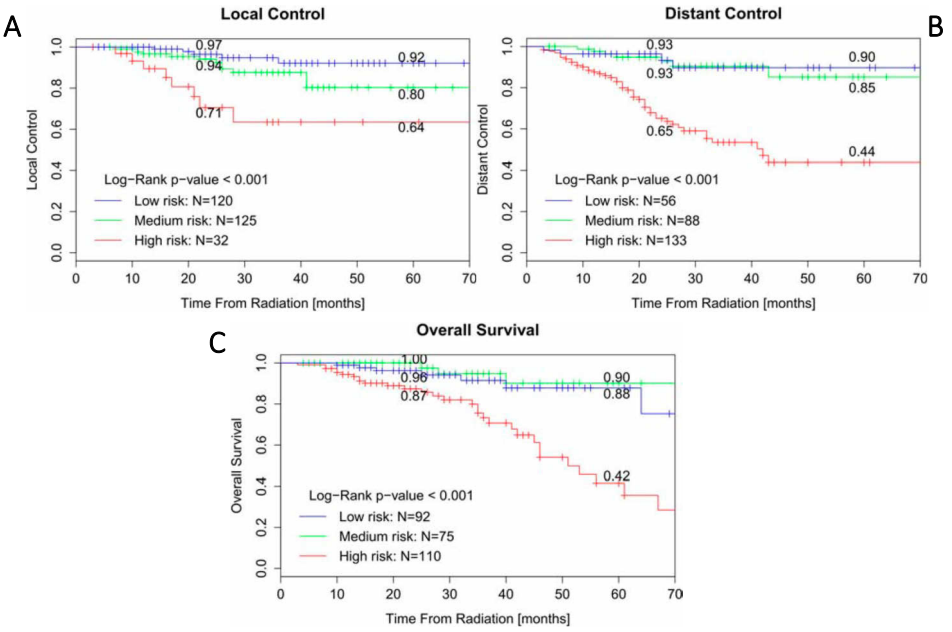


Figure 2 - Kaplan-Meier curves for patients stratified by nomogram-predicted survival. A. Kaplan-Meier curve for patients stratified by local recurrence risk of nomogram-predicted survival. B. Kaplan-Meier curve for patients stratified by distant metastasis risk of nomogram-predicted survival. C. Kaplan-Meier curve for patients stratified by death risk of nomogram-predicted survival.

Discussion

Nomograms as a rapid learning method and decision support system have been developed for several types of cancer [20–22] and radiation oncology [13]. One significant advantage of nomograms is their ability to easily deduce a (survival or curative effect) probability for individual patients. In locally advanced rectal cancer, the prediction models, and subsequent nomograms, developed by Valentini et al. appeared to be an easy-to-use tool to stratify patients into three risk groups for local recurrence, distant metastases, and overall survival. It should be noticed that these models were trained and validated on European clinical trial populations; their generalization to the routine hospital population is unknown. It is necessary to validate models in routine clinical setting to make sure models are actually performing well. We can try to optimize models with more features further before, however the question will be how generalizable they are. As to our knowledge, this is the first study validating these prediction models in a routine clinical, non-European patient population. Therefore, further external validation of these prediction models in other independent (routine clinical) datasets should be performed, since differences in population distribution and treatment may influence the accuracy and calibration of the model.

In our study, there are some differences in the distribution of clinicopathologic features between European clinical trial patients (training set) and this Chinese routine clinical cohort (Table 1). Firstly, patients in our cohort are younger than in the training set. Secondly, most cancers are located in the low to middle distance from the anal verge in our cohort and there is a higher abdominal-perineal resection (APR) surgery rate. This might be due to European has a taller body shape and higher peritoneal reflection than Asian population, and a decision bias from surgeons. Thirdly, radiation with more than 45Gy is more common in Chinese clinical cohort and there is a relatively higher ypT0 rate. The impact of an uneven distribution may lead to the underestimation of the probabilities of relapse and survival in the validation population.

We observed a good discriminative capability of the prediction models with a C-index of 0.72 ± 0.079, 0.75 ± 0.043 and 0.72 ± 0.089 for 2-year local recurrence, distant metastases and overall survival, respectively. Kaplan-Meier curves with stratification based on the predicted score indicated a good stratified performance of local control in the clinical cohort, however, couldn't discern low-risk and medium-risk groups well in distant control and overall survival. One important reason is that the number of patients in these risk groups is small in comparison to the original validation. The cut-off ranges of different risk groups in the nomograms are not optimized for this dataset and therefore introduce variation in risk group sizes in the clinical cohort. Since cut-off value is relatively subjective in different studies, we choose a 'tailored' cut-off to see if the model works well. After we redistributed the patients into different risk groups (the proportion approximately equals to European training data) by new cut-off value, the discrimination of low-risk and medium-risk groups seems better in Kaplan-Meier curve of OS (Figure 3). Therefore, not only the proper size of the population but also the 'tailored' cut-off values are important for the application of nomograms.

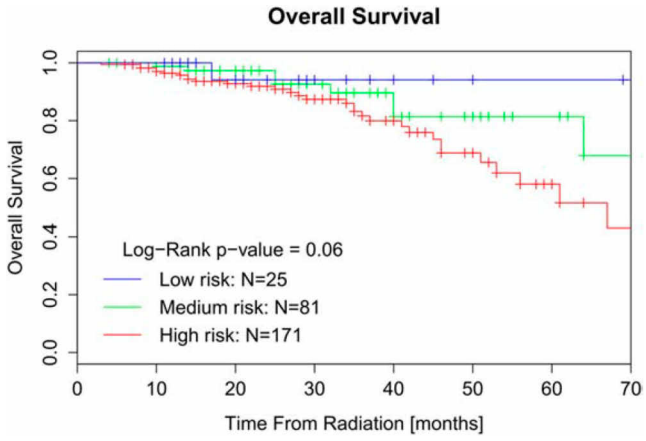


Figure 3 - Kaplan-Meier curve of OS for patients stratified by 'tailored' cut-off values

Several other limitations have to be taken into account in this study. Since preoperative chemoradiation for locally advanced rectal cancer in China started after the publication of several important randomized clinical trials, the follow-up time in the validation cohort was not long enough (median follow-up was 26 months and 23 of 277 patients lost to follow-up). Moreover, our data were from a single institution with small sample size which may not be representative of rest of the Chinese population. More patients from multiple centers are needed to enlarge the validation group. Furthermore, the current validation was based on 2-year follow-up, where the original prediction models were validated on 5-year follow-up. Although we used the adjusted baseline hazard, the original development (and variable selection) was based on 5-year follow up, and may have resulted in some differences. Nonetheless, we have shown that the prediction models still hold discriminative power when applied to a 2-year follow-up dataset.

In this retrospective study, 16 patients whose adjuvant chemotherapy information or pathological results were not available were excluded for validation, introducing a possible selection bias. Moreover, in the original trial datasets, local recurrence was diagnosed by histology, and distant metastasis by at least two imaging exams. These rules were not as strictly applied in our Chinese clinical routine validation cohort. This could cause an overestimation of relapse compared to the original trial population.

Given these limitations, the nomograms derived from European clinical trial population might not be ideally applicable to a routine Chinese population, however, they would work better with tailored cut-off values. Furthermore, as this is a validation based on routine clinical data, this validation better resembles the noise in data (incomplete data, other guidelines/policies) to be expected when applying a prediction model in the clinical setting [23]. Future work including Chinese and other clinical care populations in the training set are expected to improve its calibration as well as generalizability to the clinical care situation. As our understanding of the progress of rectal cancer, more specific clinical, pathologic, and biologic factors can be incorporated to refine this predictive model.

References

1. National Comprehensive Cancer Network Guidelines for Rectal Cancer. http.//www.nccn.org/ professionals/physician_gls/f_guidelines.asp#rectal
2. Bosset JF, Collette L, Calais G, Mineur L, Maingon P, Radosevic-Jelic L, *et al.* Chemotherapy with preoperative radiotherapy in rectal cancer. *N Engl J Med* 2006;355(11):1114–1123. doi:10.1056/NEJMoa060829
3. Kapiteijn E, Marijnen CA, Nagtegaal ID, Putter H, Steup WH, Wiggers T, *et al.* Preoperative radiotherapy combined with total mesorectal excision for resectable rectal cancer. *N Engl J Med* 2001;345(9):638–646
4. Sauer R, Liersch T, Merkel S, Fietkau R, Hohenberger W, Hess C, *et al.* Preoperative Versus Postoperative Chemoradiotherapy for Locally Advanced Rectal Cancer: Results of the German CAO/ARO/AIO-94 Randomized Phase III Trial After a Median Follow-Up of 11 Years. *J Clin Oncol* 2012;30(16):1926–1933. doi:10.1200/JCO.2011.40.1836
5. Engelen SME, Maas M, Lahaye MJ, Leijtens JWA, van Berlo CLH, Jansen RLH, *et al.* Modern multidisciplinary treatment of rectal cancer based on staging with magnetic resonance imaging leads to excellent local control, but distant control remains a challenge. *Eur J Cancer* 2013;49(10):2311–2320. doi:10.1016/j.ejca.2013.03.006
6. Zorcolo L, Rosman AS, Restivo A, Pisano M, Nigri GR, Fancellu A, *et al.* Complete Pathologic Response after Combined Modality Treatment for Rectal Cancer and Long-Term Survival: A Meta-Analysis. *Ann Surg Oncol* 2012;19(9):2822–2832. doi:10.1245/s10434-011-2209-y
7. Capirci C, Valentini V, Cionini L, De Paoli A, Rodel C, Glynne-Jones R, *et al.* Prognostic Value of Pathologic Complete Response After Neoadjuvant Therapy in Locally Advanced Rectal Cancer: Long-Term Analysis of 566 ypCR Patients. *Int J Radiat Oncol* 2008;72(1):99–107. doi:10.1016/j.ijrobp.2007.12.019
8. Bökkerink GM, de Graaf EJ, Punt CJ, Nagtegaal ID, Rütten H, Nuyttens JJ, *et al.* The CARTS study: Chemoradiation therapy for rectal cancer in the distal rectum followed by organ-sparing transanal endoscopic microsurgery. *BMC Surg* 2011;11:34. doi:10.1186/1471-2482-11-34
9. Habr-Gama A, Gama-Rodrigues J, São Julião GP, Proscurshim I, Sabbagh C, Lynn PB, *et al.* Local Recurrence After Complete Clinical Response and Watch and Wait in Rectal Cancer After Neoadjuvant Chemoradiation: Impact of Salvage Therapy on Local Disease Control. *Int J Radiat Oncol* 2014;88(4):822–828. doi:10.1016/j.ijrobp.2013.12.012
10. Pucciarelli S, De Paoli A, Guerrieri M, La Torre G, Maretto I, De Marchi F, *et al.* Local Excision After Preoperative Chemoradiotherapy for Rectal Cancer: Results of a Multicenter Phase II Clinical Trial. *Dis Colon Rectum* 2013;56(12):1349–1356. doi:10.1097/DCR.0b013e3182a2303e
11. Weiser MR, Beets-Tan R, Beets G. Management of complete response after chemoradiation in rectal cancer. *Surg Oncol Clin N Am* 2014;23(1):113–125. doi:10.1016/j.soc.2013.09.012
12. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, *et al.* 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol* 2013;109(1):159–164. doi:10.1016/j.radonc.2013.07.007
13. 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* 2013;10(1):27–40. doi:10.1038/nrclinonc.2012.196
14. Weiser MR, Landmann RG, Kattan MW, Gonen M, Shia J, Chou J, *et al.* Individualized Prediction of Colon Cancer Recurrence Using a Nomogram. *J Clin Oncol* 2008;26(3):380–385. doi:10.1200/JCO.2007.14.1291
15. Massacesi C, Norman A, Price T, Hill M, Ross P, Cunningham D. A clinical nomogram for predicting long-term survival in advanced colorectal cancer. *Eur J Cancer* 2000;36(16):2044–2052. doi:10.1016/S0959-8049(00)00286-0
16. Kattan MW, Gönen M, Jarnagin WR, DeMatteo R, D'Angelica M, Weiser M, *et al.* A Nomogram for Predicting Disease-specific Survival After Hepatic Resection for Metastatic Colorectal Cancer. *Ann Surg* 2008;247(2):282. doi:10.1097/SLA.0b013e31815ed67b

17. Valentini V, Van Stiphout RG, 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* 2011;29(23):3163–3172
18. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, *et al.* Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology* 2010;21(1):128–138. doi:10.1097/EDE.0b013e3181c30fb2
19. Hilden J, Habbema JD, Bjerregaard B. The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods Inf Med* 1978;17(4):227–237
20. Kattan MW, Karpeh MS, Mazumdar M, Brennan MF. Postoperative Nomogram for Disease-Specific Survival After an R0 Resection for Gastric Carcinoma. *J Clin Oncol* 2003;21(19):3647–3650. doi:10.1200/JCO.2003.01.240
21. Gronchi A, Miceli R, Shurell E, Eilber FC, Eilber FR, Anaya DA, *et al.* Outcome Prediction in Primary Resected Retroperitoneal Soft Tissue Sarcoma: Histology-Specific Overall Survival and Disease-Free Survival Nomograms Built on Major Sarcoma Center Data Sets. *J Clin Oncol* 2013;31(13):1649–1655. doi:10.1200/JCO.2012.44.3747
22. Hansen J, Rink M, Bianchi M, Kluth LA, Tian Z, Ahyai SA, *et al.* External validation of the updated briganti nomogram to predict lymph node invasion in prostate cancer patients undergoing extended lymph node dissection. *The Prostate* 73(2):211–218. doi:10.1002/pros.22559
23. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. *Br J Cancer* 2014;110(3):551–555. doi:10.1038/bjc.2013.725

Chapter 10

Prospective validation of pathologic complete response models in rectal cancer: Transferability and reproducibility

Authors
Johan van Soest, Elisa Meldolesi, Ruud van Stiphout, Roberto Gatta, Andrea Damiani, Vincenzo Valentini, Philippe Lambin, Andre Dekker

Adapted from
Medical Physics, September 2017, Volume 44, Issue 9, pages 4961 – 4967
DOI: 10.1002/mp.12423

Abstract

Purpose: Multiple models have been developed to predict pathologic complete response (pCR) in locally advanced rectal cancer patients. Unfortunately, validation of these models normally omit the implications of cohort differences on prediction model performance. In this work, we will perform a prospective validation of three pCR models, including information whether this validation will target transferability or reproducibility (cohort differences) of the given models.

Methods: We applied a novel methodology, the cohort differences model, to predict whether a patient belongs to the training or to the validation cohort. If the cohort differences model performs well, it would suggest a large difference in cohort characteristics meaning we would validate the transferability of the model rather than reproducibility. We tested our method in a prospective validation of three existing models for pCR prediction in 154 patients.

Results: Our results showed a large difference between training and validation cohort for one of the three tested models [Area under the Receiver Operating Curve (AUC) cohort differences model: 0.85], signaling the validation leans towards transferability. Two out of three models had a lower AUC for validation (0.66 and 0.58), one model showed a higher AUC in the validation cohort (0.70).

Discussion: We have successfully applied a new methodology in the validation of three prediction models, which allows us to indicate if a validation targeted transferability (large differences between training/validation cohort) or reproducibility (small cohort differences).

Introduction

As the field of radiation oncology is moving towards individualized medicine, the need to identify (sub-)groups of patients on the basis of patient and/or tumor features is emerging [1]. Machine learning techniques using (routine) clinical patient information are needed to identify these features. Furthermore, machine learning can be used to develop a prognostic model for disease development, or to develop a predictive model where the outcome may vary, based on the applied intervention(s). These prognostic and predictive models are the building blocks for clinical decision support systems (CDSS) [2]. The promise of these CDSSs is to handle and adapt to insights found in research, relieving the clinical staff from the burden of keeping up with the high volume of publications and the rapidly increasing amount of knowledge [3,4].

Before implementing clinical prediction models into a CDSS, these models need validation on different levels [5]. These levels can be classified using the TRIPOD statement [6]. Although in many studies internal/external validations are included, they normally do not describe validation results to their full extent. According to Justice et al. [7], validation of prediction models should describe two aspects: Accuracy validation (performance of the model) and generalizability (how similar/dissimilar are training and validation cohorts and why and how do these differences influence the performance of the model).

Accuracy, or model performance validation, describes the statistical validity of a prognostic or predictive model [8]. In general, model performance (or fitness) is determined by the discriminative ability and calibration of a prognostic/predictive model [9]. The discriminative ability describes how well a model correctly classifies a subject into the correct group. Calibration describes the agreement of the frequency between observed and predicted events.

The second aspect, generalizability, can be divided into two components: reproducibility and transferability. Reproducibility describes the accuracy of a prediction model on similar cohorts, where transferability tests the accuracy of a prediction model on cohorts with different characteristics. Similarities or differences between two cohorts are affected by temporal, methodological or geographic aspects [7]. An example of a temporal difference is the emerging influence of HPV on head & neck cancer patients [10]. Methodological differences could originate from different treatments being applied in the same patients or different levels of quality (e.g., clinical routine versus clinical trials). Geographical different origins of the training and validation cohort could make these different in, for example, race and socioeconomic factors. Often these are interrelated with geographical differently located cancer centers treating different patients differently at different times [11,12]. Often, (external) validation of prediction models only describe the accuracy. The method described by Debray et al. [13] can be used to estimate the difference between the training and validation cohort, measuring the level of generalizability (same characteristics) versus transferability (different characteristics) between training and validation cohorts. By adding this measurement, next

to the model performance on the validation set, it gives more insight (without hard boundaries) in which situations a prediction model does (not) work (Figure 1). Therefore, it is imperative to add this measure in the general model validation process, as it better describes for which cohorts a prediction model was tested.

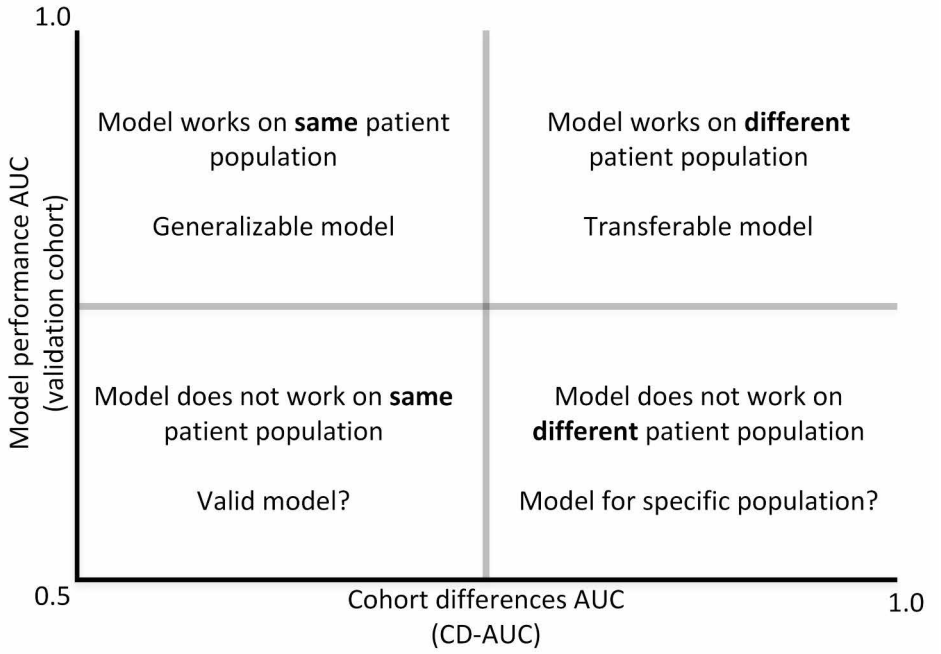


Figure 1 - Model performance in perspective of validation performance. The x-axis shows the cohort differences AUC (as described by Debray et al [13]), where the y-axis describes the model performance. Boundaries between quadrants are only indicative.

In this work, we aim to investigate this reproducibility and transferability metric in a prospective validation of three prediction models for pathologic complete response (pCR) in rectal cancer patients. These models have been developed and retrospectively validated by van Stiphout et al. [14] based on prior work identifying prognostic factors for pathologic response [15–17]. We hypothesize that this prospective validation tests for reproducibility, with comparable (or slightly reduced) model performances.

Materials and methods

The three models we validated were learned on three different training cohorts as published previously [14]. These models predict pathological complete response (pCR) based on different groups of available data: (a) only clinically available parameters (clinical model), (b) Clinically available parameters + pre-treatment PET parameters (pre-

treatment PET model), (c) Clinically available parameter + pre-treatment PET + post-treatment PET parameters (post-treatment PET model). For the PET parameters, tumors were semi-automatically contoured on PET-CT scans using commercial software (TrueD, Siemens Medical, Erlangen, Germany). Standardized Uptake Value (SUV) thresholding was performed using the gluteus muscle to set the threshold for the automatic contouring, within pre-defined boundaries [18]. The response index (RI) describes the ratio between the pre-treatment and posttreatment SUV value of the primary tumor [14]. Pathological complete response was determined as having a T0N0M0 based on the surgical specimen, extracted from the pathology report. Based on the three different datasets, an exhaustive feature selection was performed to train a proximal Support Vector Machine (SVM). Internal validation was performed using a leave-one-out cross-validation [14]. Original cohort datasets for training and validation were at our disposal. The cohort used for our prospective validation was the THUNDER trial cohort (NCT00969657). This cohort consists of 154 patients, from two participating centers (MAASTRO Clinic, Maastricht University Medical Centre+, Netherlands and Sacred Heart University Hospital, Rome, Italy). All patients included in this THUNDER trial gave written informed consent before data was collected.

Univariate cohort differences were tested for statistical significance using Wilcoxon rank sum test [19] (for continuous variables) or Fisher's exact test [20] (for categorical variables). To correct for multiple (univariate) testing, we calculated an adjusted P-value using the Bonferroni correction which multiplies the P-values by the number of comparisons. In our case, multiplying the P-values by the number of model input and output parameters. Cohort characteristics for the prediction model variables are shown in Tables 1, 2 and 3.

Table 1- Cohort characteristics clinical prediction model (pre-treatment, without PET features). The adjusted p-value determines a statistically significant difference if < 0.05; based on the original training cohort, and current (prospective) validation cohort. Wilcoxon rank sum test was used for continuous variables, and Fisher's exact test for categorical variables.

Variable	Training	Validation (pros)	p-value	p-value adjusted
# Patients	677	112		
Tumor length [cm] (SD)	4.97 (1.73)	5.03 (1.81)	$9.63 \cdot 10^{-1}$	$9.63 \cdot 10^{-1}$
cT			$5.70 \cdot 10^{-9}$	$2.28 \cdot 10^{-8}$
1	4 (0%)	0 (0%)		
2	18 (3%)	16 (14%)		
3	583 (86%)	70 (63%)		
4	72 (11%)	26 (23%)		
cN			$1.45 \cdot 10^{-9}$	$4.34 \cdot 10^{-8}$
0	154 (23%)	9 (8%)		
1	307 (45%)	35 (31%)		
2	216 (32%)	68 (61%)		
pCR	134 (20%)	29 (26%)	$1.60 \cdot 10^{-1}$	$3.30 \cdot 10^{-1}$

Table 2- Cohort characteristics pre-treatment prediction model including PET features. Tumor location measures the distance from the anal verge.

Variable	Training	Validation (pros)	p-value	p-value adjusted
# Patients	114	98		
Max tumor diameter [cm] (SD)	7.01 (2.15)	5.70 (1.75)	$3.14 \cdot 10^{-7}$	$1.26 \cdot 10^{-6}$
cN			$1.60 \cdot 10^{-7}$	$7.99 \cdot 10^{-7}$
0	28 (24%)	8 (8%)		
1	56 (49%)	28 (29%)		
2	30 (26%)	62 (63%)		
Tumor location			$2.42 \cdot 10^{-6}$	$7.27 \cdot 10^{-6}$
0-5 cm	56 (49%)	29 (30%)		
5-10 cm	38 (33%)	15 (15%)		
10-15 cm	20 (17%)	54 (55%)		
SUV max (SD)	13.65 (6.23)	17.20 (8.29)	$9.61 \cdot 10^{-4}$	$1.92 \cdot 10^{-3}$
pCR	17 (15%)	25 (25%)	$5.91 \cdot 10^{-2}$	$5.91 \cdot 10^{-2}$

Table 3 - Cohort characteristics post-treatment prediction model (with PET features)

Variable	Training	Validation (pros)	p-value	p-value adjusted
# Patients	107	53		
Response Index SUV max pre/ post (SD)	56.79 (27.24)	63.32 (20.19)	$2.9 \cdot 10^{-1}$	1
Tumor length [cm] (SD)	5.54 (2.33)	5.12 (1.75)	$3.5 \cdot 10^{-1}$	1
SUV max (post-treatment) (SD)	5.94 (3.13)	5.25 (2.20)	$3.6 \cdot 10^{-1}$	1
pCR	26 (24%)	13 (24%)	1	1

Next, we calculated the multivariate cohort differences (MCD) using the method proposed by Debray et al [13]. This method assesses the ability to predict whether a specific patient in our cohort belongs to cohort A (training) or B (validation). When we are able to predict to which cohort patients belong, it would mean that (several of) the underlying prediction model variables have very different distributions (pointing to a validation which would test *transferability*). In contrast, when we cannot predict to which cohort patients belong, it would mean that the model variables are more homogeneous among the training and validation cohort (which would be a validation set which is suitable to test *reproducibility*). As this method predicts the originating cohort for a specific patient, we can apply generic accuracy validation measures; in this case we will use the Area under the Receiver Operating Curve (AUC) [21]. In this situation, an AUC close to 0.5 indicates no predictive performance and hence little differences in cohort characteristics. An AUC deviating from 0.5 will indicate differences in cohort characteristics. We considered cohorts equal for an AUC =< 0.6, moderately different between 0.6 and 0.8, and highly different for an AUC > 0.8. To avoid confusion with the actual evaluation of the prediction model, we will use the term Cohort Differences AUC (CD-AUC) to denote the result of the method explained above.

After performing tests to describe univariate and multivariate cohort differences, we compared the distributions of predicted probabilities in the training and validation cohorts by calculating mean probabilities and corresponding standard deviations. Furthermore, we evaluated the prediction model performance on both cohorts using the Area under the Receiver Operating Curve (AUC) [21], Hosmer-Lemeshow C-statistic [22] and Brier score [23] to determine the discriminative ability, calibration and accuracy, respectively [24,25]. These performance measures have different characteristics: The AUC specifies the ability to make a threshold, separating the probabilities for a given outcome into a binary yes/no result (discriminative performance). Unfortunately, this AUC doesn't take the distance between a probability and the actually measured outcome into account, hence only determines the best operating (threshold) point. In contrast, calibration measures how well the predicted probability is comparable to the actual incidence of the outcome. For example, the Hosmer–Lemeshow C-statistic splits patients into n groups, based on ordered prediction probabilities, and uses the Chi-square test to assess statistical significant differences between observed and predicted outcomes [22]. Finally, aspects from both discriminative ability and calibration are available in accuracy measurements. One of these measurements is the Brier score, which is the mean squared error between probabilities and the observed outcome [23]. This score is not suitable as a single measure; however is useful when comparing different models with equal outcomes and/or cohorts [9]. For more information regarding these model performance metrics, we would like to refer to Steyerberg et al [24].

For robustness purposes, we used bootstrapping as a resampling technique (R = 1000), and applied this method to the discrimination (AUC) and accuracy (Brier score) measurements. All calculations and statistical analysis were performed using R (version 3.3.2) [26]. A generalized workflow of the applied methods is shown in Figure 2.

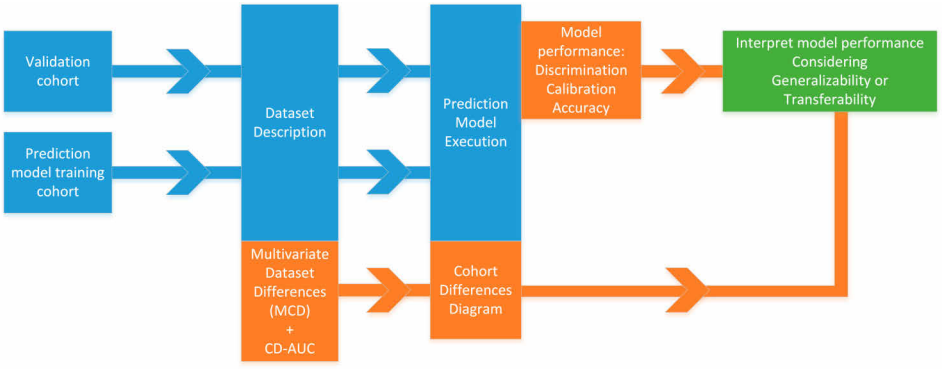


Figure 2 - Generalized workflow for validation of existing prediction models, where model performance is put into respect of generalizability/transferability of the evaluated prediction model.

Results

The multivariate cohort differences are shown in Table 4. For every prediction model, we made a separate multivariate model to predict whether a case belongs to the training or validation cohort, and determined the AUC [further referred as the Cohort Differences AUC (CD-AUC)]. The CD-AUCs were 0.69, 0.85 and 0.62 for the clinical, pre-treatment PET and post-treatment PET prediction model variables, respectively.

Table 4- Multivariate differences model for clinical, pre-treatment PET and post-treatment PET prediction model, based on original and current prospective validation cohort.

Variable	Coefficient	p-value	CD-AUC (95% CI)
Clinical prediction model			
Intercept	2.72	$5.05*10^{-4}$	0.69 (0.61 – 0.71)
Tumor length	0.04	$5.09*10^{-1}$	
cT	0.11	$6.69*10^{-1}$	
cN	-1.01	$5.65*10^{-9}$	
pCR	0.57	$2.36*10^{-2}$	
Pre-treatment PET prediction model			
Intercept	1.32	$1.23*10^{-1}$	0.85 (0.78 – 0.89)
Max tumor diameter	0.51	$5.18*10^{-6}$	
cN	-1.36	$1.06*10^{-6}$	
Tumor location	-0.66	$1.59*10^{-3}$	
SUV max	-0.07	$6.21*10^{-3}$	
pCR	-0.81	$8.31*10^{-2}$	
Post-treatment PET prediction model			
Intercept	0.35	$7.71*10^{-1}$	0.62 (0.51 – 0.67)
Response Index SUV max pre/post	-0.01	$4.91*10^{-1}$	
Tumor length	0.09	$3.35*10^{-1}$	
SUV max (post-treatment)	0.04	$6.30*10^{-1}$	
pCR	0.31	$4.69*10^{-1}$	

For the clinical prediction model, the multivariate differences model showed statistically significant differences in clinical nodal stage and pCR. For the clinical + pre-treatment PET prediction model, the multivariate differences model showed a high discriminative ability (CD-AUC: 0.85). In this model, almost all variables showed a statistically significant difference; except for pCR. Finally, for the clinical + pre- and post-treatment PET prediction model, the multivariate differences model had a low discriminative ability (CD-AUC: 0.62). None of the variables in this last model showed a statistically significant difference.

Based on the CD-AUC values in Table 4, we can probably state that the clinical and pre–post PET prediction models are being validated for reproducibility, where the pre-

treatment PET prediction model is being validated for transferability. The CD-AUC of the pre-treatment PET prediction model indicated a high predictive ability whether a patient belongs to the training/validation cohort. This is also expressed in the (multivariate) cohort differences model coefficients (Table 4) deviating from 0.

The comparison of distributions of predicted probabilities in the training and validation cohorts for all three prediction models are shown in Figure 3. For the clinical + pre-treatment PET model, the mean and standard deviations for the predicted probabilities are almost equal in both training and validation cohorts. The post-treatment PET model shows a higher average probability in the validation dataset, with a smaller standard deviation. The latter could be due to the small number of patients available for this prediction model.

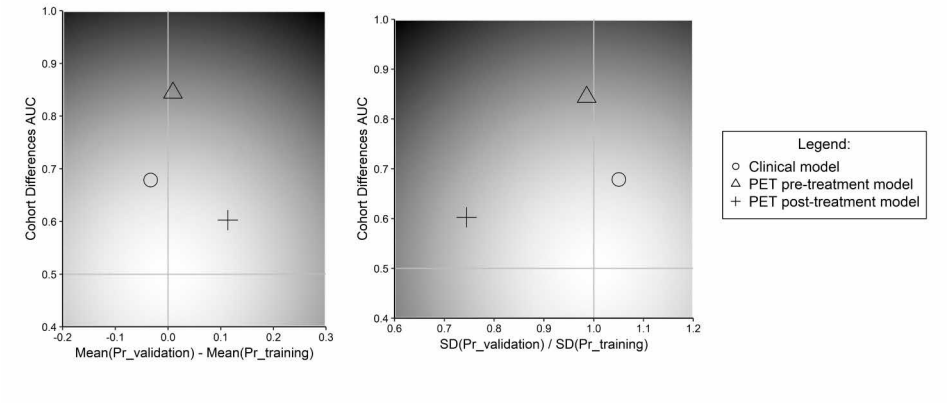


Figure 3 – Diagrams displaying the differences between training and validation cohorts on different aspects. For both graphs, the y-axis represents differences in cohort characteristics (CD-AUC). The x-axis shows the difference in mean probability of pCR (left figure, < 0 indicates lower mean probability in validation), or ratio of standard deviations (SD, right figure, > 1.0 indicates larger SD in validation) in the training and validation cohort.

After describing the (dis)similarity of training and validation cohorts, we will present the result of the model performance on both cohorts. The prediction model performance results for both cohorts are shown in Table 5 For both AUC and Brier score, standard deviations (SD) are given, based on bootstrapping the validation cohort. In addition, Figure 4 show the calibration plots of predicted and observed outcomes for both training and validation cohorts. For the clinical prediction model, the AUC increased in the validation cohort, the Hosmer-Lemeshow p-value showed a larger deviation from perfect calibration (P-value < 0.05), and the Brier score increased in the validation cohort (indicating a decrease in overall model accuracy). For both pre- and post-treatment PET model, validation metrics showed a similar trend with an exception for the AUC (decrease instead of increase). For the post-treatment PET model, the decrease in AUC was 2.8 times the standard deviation in the validation. As this standard deviation (0.10) was considered large, we would address this to the distribution in probabilities described

before (higher mean probability; smaller standard deviation) and subsequently the population size applicable for this prediction model.

Table 5 - Prediction performance results on both training and validation cohort for all three prediction models. Performance is measured in terms discrimination (AUC; the higher the better), calibration (Hosmer-Lemeshow C-statistic); p-value > 0.05 the "better") and accuracy (Brier score; the lower the better). For both AUC and Brier score, standard deviations (SD) are given, based on bootstrapping the validation cohort.

Model	AUC		H-L p-value		Brier	
	Training (SD)	Validation (SD)	training	validation	Training (SD)	Validation (SD)
Clinical	0.62 (0.03)	0.70 (0.06)	$2.6*10^{-2}$	$4.2*10^{-3}$	0.126 (0.008)	0.153 (0.021)
PET pre	0.74 (0.06)	0.66 (0.07)	$1.2*10^{-5}$	$3.14*10^{-2}$	0.118 (0.009)	0.149 (0.013)
PET post	0.86 (0.04)	0.58 (0.10)	$8.4*10^{-7}$	$8*10^{-3}$	0.135 (0.007)	0.164 (0.012)

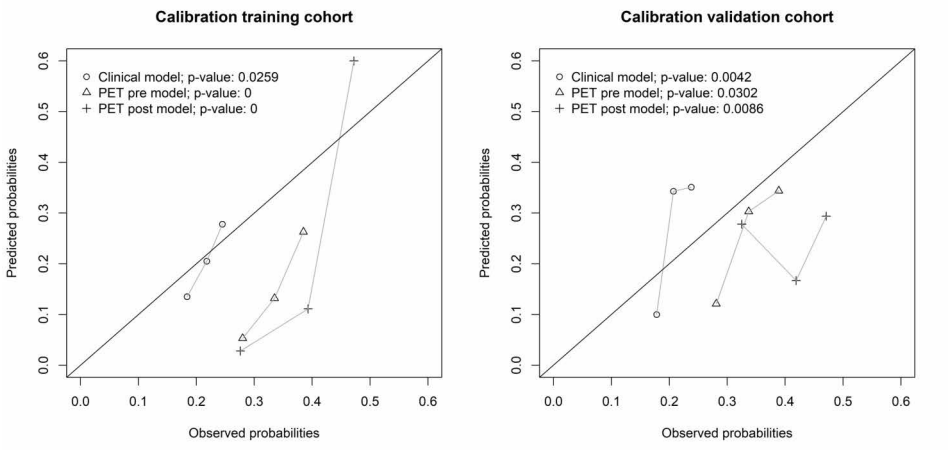


Figure 4 – Calibration plots for both training (left) and validation (right) cohorts, for all prediction models. All groups have equal number of patients; incidence within groups may vary.

Discussion

The In this work, we have successfully performed a prospective validation (TRIPOD statement [6] type 4) of three previously developed prediction models, and applied an additional method to assess the differences between training and validation cohorts. In addition to the traditional accuracy validation, our analysis gives additional information to clinicians whether the validation was performed on a similar or different cohort (in terms of population characteristics), and therefore whether the validation assessed the reproducibility (possible same clinical setting), or transferability (possible different clinical setting) of a prediction model. As these measures are relatively easy to interpret, they could be used when commissioning prognostic models for use in clinical practice,

by assessing whether the population in a certain clinic is different from the population where the model was trained on.

We would advise to validate prediction models on trial and routine clinical cohorts as also suggested by Booth and Tannock [27] and proposed in the VATE project [28,29]. The quality of cohorts from clinical trials are needed to identify which variables need to be reported in clinical practice. Afterwards, training/validating models (using the methods explained here) on routine clinical data would increase the cohorts available to learn/validate upon as was done by Shen et al [30]. Furthermore, validation in a clinical setting could also reduce the turnaround time between developing/validating and using predictive models in clinical practice; enabling rapid learning healthcare and subsequently decision support [2,3].

When evaluating the results, the significance of the univariate differences (P-values between training and validation cohort; Tables 1, 2 and 3) generally overlapped with the multivariate cohort differences, described by the covariate weight P-values (Table 4). But several variables which were significant in the univariate variable assessment lost their significance in the multivariate assessment (e.g., clinical T-stage); or became significant (e.g., pCR). In our opinion, this could be affected by differences in sample sizes between training and validation, or the effect of testing one variable versus testing the complete characteristic of a patient. Although this correlation reduces the added value of the cohort differences metric, we still think this metric is an added value as a single measure to assess cohort differences: to determine whether external validation results measure reproducibility or transferability. Secondly, statistical tests only measure significant differences; the cohort differences model can reveal subtle differences which only become apparent in a multivariate analysis.

For the post-treatment PET model, our main hypothesis is that the prediction model was overfitted on the training cohort (pCR positive outcomes = 26). When calculating a sample size for model training, we would use 10 events per variable as used in this rule of thumb [31]. As an example, the training cohort would need 30, 40 and 30 events (pCR) for the clinical, pre-treatment PET and post-treatment PET model, respectively. When considering a pCR percentage of 20%, this would result in a population size of 150, 200 and 150 patients, respectively. As a result, only the clinical prediction model training cohort would be considered large enough. Regarding the validation cohort, the only studies investigating model validation cohort sizes up to our knowledge are by Collins et al. [32] and Vergouwe et al [33]. They do state that 100 events would be a minimum, meaning that the required sample size would be 500 patients, considering a pCR event rate of 20%. Therefore we have to state that our validation might be underpowered, however, could only be accomplished by large multicenter trials. This also means that the cohort difference model and AUC values cannot reliably detect a difference in cohorts in underpowered datasets.

Future work would include the validation of the clinical pCR prediction model in a routine clinical cohort, and investigate applicability of prediction models in clinical practice.

Conclusion

In general, we would advise to apply the explained methods when validating (existing) prediction models, as it puts prediction model performance in perspective of the heterogeneity between training and validation cohorts. Our workflow (Figure 2) could therefore be used as a guideline.

Based on these results, we can also state that the clinical prediction model performed well when reproducing results in the current prospective validation. The pre- and post-treatment PET prediction models were unfortunately underpowered in both training and validation cohorts.

References

1. 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* 2013;10(1):27–40. doi:10.1038/nrclinonc.2012.196
2. Lambin P, Zindler J, Vanneste BGL, De Voorde LV, Eekers D, Compter I, *et al.* Decision support systems for personalized and participative radiation oncology. *Adv Drug Deliv Rev* 2016. doi:10.1016/j.addr.2016.01.006
3. Lambin P, Roelofs E, Reymen B, Velazquez ER, Buijsen J, Zegers CML, *et al.* 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol* 2013;109(1):159–164. doi:10.1016/j.radonc.2013.07.007
4. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, *et al.* Rapid-learning system for cancer care. *J Clin Oncol* 2010;28(27):4268–4274
5. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, *et al.* Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med* 2013;10(2):e1001381. doi:10.1371/journal.pmed.1001381
6. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015;13(1). doi:10.1186/s12916-014-0241-z
7. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130(6):515–524
8. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19(4):453–473
9. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media; 2008
10. Lajer CB, Buchwald CV. The role of human papillomavirus in head and neck cancer. *APMIS* 2010;118(6–7):510–519. doi:10.1111/j.1600-0463.2010.02624.x
11. Jochems A, Troost EGC, Dekker A, Lambin P, Oberije C. Improving prediction models in the era of rapid learning health care: weighting data to reflect relative importance, *ESTRO FORUM*. Barcelona: 2015. doi:10.3252/pso.eu.estro2015.2015
12. Dekker A, Vinod S, Holloway L, Oberije C, George A, Goozee G, *et al.* Rapid learning in practice: A lung cancer survival decision support system in routine patient care data. *Radiother Oncol* 2014;113(1):47–53. doi:10.1016/j.radonc.2014.08.013
13. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;68(3):279–289. doi:10.1016/j.jclinepi.2014.06.018
14. 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* 2011;98(1):126–133. doi:10.1016/j.radonc.2010.12.002
15. Janssen MHM, Öllers MC, Riedl RG, van den Bogaard J, Buijsen J, van Stiphout RGPM, *et al.* Accurate Prediction of Pathological Rectal Tumor Response after Two Weeks of Preoperative Radiochemotherapy Using 18F-Fluorodeoxyglucose-Positron Emission Tomography-Computed Tomography Imaging. *Int J Radiat Oncol* 2010;77(2):392–399. doi:10.1016/j.ijrobp.2009.04.030
16. Janssen MHM, Öllers MC, van Stiphout RGPM, Riedl RG, van den Bogaard J, Buijsen J, *et al.* PET-Based Treatment Response Evaluation in Rectal Cancer: Prediction and Validation. *Int J Radiat Oncol* 2012;82(2):871–876. doi:10.1016/j.ijrobp.2010.11.038
17. van den Bogaard J, Janssen MHM, Janssens G, Buijsen J, Reniers B, Lambin P, *et al.* Residual metabolic tumor activity after chemo-radiotherapy is mainly located in initially high FDG uptake areas in rectal cancer. *Radiother Oncol* 2011;99(2):137–141. doi:10.1016/j.radonc.2011.04.004

18. Öllers M, Bosmans G, van Baardwijk A, Dekker A, Lambin P, Teule J, *et al.* The integration of PET-CT scans from different hospitals into radiotherapy treatment planning. *Radiother Oncol* 2008;87(1):142–146. doi:10.1016/j.radonc.2007.12.025
19. Wilcoxon F. Individual Comparisons by Ranking Methods. *Biom Bull* 1945;1(6):80. doi:10.2307/3001968
20. Fisher RA. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *J R Stat Soc* 1922;85(1):87. doi:10.2307/2340521
21. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29–36
22. Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol* 1982;115(1):92–106
23. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 1950;78(1):1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
24. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, *et al.* Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology* 2010;21(1):128–138. doi:10.1097/EDE.0b013e3181c30fb2
25. Peek N, Abu-Hanna A. Clinical prognostic methods: Trends and developments. *J Biomed Inform* 2014;48:1–4. doi:10.1016/j.jbi.2014.02.016
26. R Core Team. *R A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2016
27. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. *Br J Cancer* 2014;110(3):551–555. doi:10.1038/bjc.2013.725
28. Meldolesi E, van Soest J, Alitto AR, Autorino R, Dinapoli N, Dekker A, *et al.* VATE: VAlidation of high TEchnology based on large database analysis by learning machine. *Colorectal Cancer* 2014;3(5):435–450. doi:10.2217/crc.14.34
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* 2014;112(1):59–62
30. Shen L, van Soest J, Wang J, Yu J, Hu W, Gong YUT, *et al.* Validation of a rectal cancer outcome prediction model with a cohort of Chinese patients. *Oncotarget* 2015
31. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984;3(2):143–152
32. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study: Sample size considerations for validating a prognostic model. *Stat Med* 2015. doi:10.1002/sim.6787
33. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;58(5):475–483. doi:10.1016/j.jclinepi.2004.06.017

Chapter 11

Updated prognostic models for local recurrence, distant metastases and overall survival in a pooled dataset of rectal cancer patients

Authors
Vincenzo Valentini*, Johan van Soest*, Carlota Massiocchi, Elisa Meldolesi, Giuditta Chiloiro Nicola Dinapoli, Andrea Damiani, Andre Dekker
* Authors contributed equally

Work in Progress

Introduction

In the last decades, new strategies as the standardization of surgical procedures, the intensification of radiotherapy (RT) dose and multimodalities approaches have significantly decreased the rates of local recurrence (LR), increasing the probability of overall survival (OS) and improving the quality of life in locally advanced rectal cancer (LARC) patients [1,2].

Next to this debate on outcomes, different forms of treatments have historically been investigated. Many different clinical trials, comparing different treatment combinations using components like neo-adjuvant chemotherapy (CT), concurrent CT, RT, surgery and adjuvant CT, have been conducted over time [3–14]. These treatment components also have varying characteristics such as different chemotherapy compounds, different RT methods (e.g. varying in dose and/or radiation methods), and different surgery procedures. This large heterogeneity in both components and their characteristics introduces challenges when comparing multiple clinical trials. To overcome these challenges, several meta analyses have been performed to compare different trials based on (meta)data characteristics in their respective trial arms [15–17]. Next to these meta-analyses, other groups have pooled datasets from different clinical trials (with comparable treatment protocols) to group comparable trial arms and perform a pooled analysis [18,19].

Smith and Brown have attempted to stratify rectal cancer patients into three risk groups, depending on the risk of local recurrence or distant metastases by defining “good”, “bad”, and “ugly” cases based on available prognostic factors [20]. Unfortunately, taking into account all of these prognostic factors is difficult, as the number of factors quickly outgrows the human cognitive capacity of information to process [21].

Prediction models might help physicians to define these groups which might lead to a more personalized treatment in oncology [22]. Using a model-based approach might support physicians to choose the best treatment: LARC patients with a poor prognosis might benefit from an intensified treatment (when physical conditions allow), while those with a prediction of pathologic complete response (pCR; the absence of tumor tissue in the pathologic specimen) might benefit from de-escalating therapy as an organ-preserving strategy [23–25].

Several prediction models have been developed for LARC to predict pCR as an intermediate endpoint [26–28], or follow-up endpoints like LR, DM OS, or disease-free survival (DFS) [29–31].

In this work, our aim is to update the previously developed prediction models for LR, DM and OS developed by Valentini et al. [29]. These models were based on a pooled analysis of five LARC clinical trial datasets, and have been validated on Chinese and European routine clinical datasets [32,33]. In the cohort of patients used for the Chinese validation, the local control (LC) rate resulted overestimated compared to the observation in the clinical one. For this reason, authors concluded that the model applicability on an Asian

patient population needed to include additional prognostic factors, suggesting that an internal review of the models is necessary before applying them on your own data. In the European cohort the addition of neoadjuvant rectal score combined with the models was used to better stratify the patient risk at 5-years survival.

The pooled dataset used for the current analysis was extended with more (non-) European trials, including a higher number of patients, longer follow-up time and more in-depth analysis. Our main hypothesis entails that the new prediction models have equal or better performance than the previous models and are more robust by using more patients and longer FU time. Moreover, for each outcome multiple follow-up months (12,24,60 and 120 months) are presented (including the respective time-related performance) to support decisions at different follow-up times.

Methods

Study Population

The dataset used for our analysis contained 14 international trial cohorts (N=9667) investigating locally advanced rectal cancer (Table 1). All trials varied in treatment protocols, patient characteristics, and accrual years. After merging all trial cohorts, mean imputation was performed on variables having missing data for less than 10% of all cases, stratified per trial. Exclusion criteria for patients or trial arms were: (a) short-course radiotherapy (e.g. 5x5Gy) or no radiotherapy, (b) M+ patients, (c) (neo-)adjuvant chemotherapy regimens containing oxaliplatin, (d) Adjuvant chemo-radiotherapy (CRT), (e) Surgery procedures different than anterior-resection or abdominoperineal resection, (f) T1 patients, (g) patients with a treatment duration < 0 or > 80 days (time between first and last RT fraction).

The variables under investigation in this analysis are: sex, age at diagnosis or randomization, clinical tumor and nodal stage, tumor distance from anal verge, tumor length, radiotherapy dose, neo-adjuvant chemotherapy administered, concurrent chemotherapy administered, surgery procedure, pathological tumor and nodal stage. The dataset was split into a training (80%) and testing (20%) dataset using a uniform random assignment, stratified per trial. Outcome variables were the right-censored variables describing LR, DM and OS. Variable characteristics and their distributions between training and validation dataset are shown in Table 2.

Table 1 - Trials included in the pooled analysis dataset

Trial name	Treatment arm	RT Dose	Inclusion Criteria	Accrual	# Patients	
					Total	Included
EORTC 22921 [3]	Preoperative RT	45	T3 or resected T4M0 No history of cancer Age < 80 years WHO 0-1 Tumor within 15cm of the anorectal verge	1993 - 2003	1011	908
	Preoperative CRT	45				
	Preoperative RT + postoperative CT	45				
	Preoperative CRT + postoperative CT	45				
FFCD 2903 [4]	Preoperative RT	45	T3 or resectable T4, M0 Untreated rectal adecarcinoma Age < 75 years WHO 0-1	1993 - 2003	742	679
	Preoperative CRT	45				
CAO/ARO/AIO 94 [5]	Preoperative CRT	50	Stage II-III No history of cancer Age < 75 years Tumor within 16cm of the anorectal verge	1995-2002	799	372
	Postoperative CRT	55				
Polish I [34]	Preoperative RT	25	cT3-4, M0 unresectable tumor Age < 75 years WHO 0-2 Lower border of tumor 615 cm from anal verge No previous Radiotherapy to the pelvis	1999-2002	312	131
	Preoperative CRT	50				
ACCORD [7]	CRT (Capecitabine), 45Gy	45	T2 tumors located in the anterior and lower rectum CT3 or resectable cT4-M0 Age < 80 years WHO performance status of 0 or 1	2005-2008	598	201
	CRT (Oxaliplatin), 50Gy	45				
Dutch TME [8]	Preoperative RT + surgery	25	-	-	1731	0
	Surgery alone	-				
Swedish [9]	Preoperative RT + surgery	25	M0	-	908	0
	Surgery alone	-				

Trial name	Treatment arm	RT Dose	Inclusion Criteria	Accrual	# Patients	
					Total	Included
I-CNR-RT [10]	CRT	45	Adenocarcinoma of the extraperitoneal portion of the rectum (within 15 cm from the anal margin) cT3-4 Age < 75 years WHO of 0 or 1 Leucosyte count > 3000/L platelet count >130000/L, serum creatinine below 1,2 mmol/L	-	634	533
	CRT + adjuvant chemo	45				
Glynne-Jones [35]	CRT	45	Age < 18 years, rectal adenocarcinoma, located < 15 cm from the anal verge, or below the peritoneal reflection WHO 0-1 Adequate haematologic renal, and hepatic function M0 ypT0-T4 N0-N2	2005-2008	113	106
	CRT + Adjuvant chemo	45				
INTERACT [36]	CRT (Capacitabine + Oxaliplatin)	50	cT2-3-4 Age < 78 years	2006-2013	538	168
	CRT (Capacitabine)	55				
CAO/ARO/AIO 04 [11,37]	CRT (5FU), 50Gy	50	cT2-3-4 or N+-M0 Age > 18 years Age < 78 years Adenocarcinoma with an inferior margin no more than 12 cm above the anal verge ECOG 0-2 No prior radiotherapy or chemotherapy	2008-2010	1236	608
	CRT (5FU + Oxaliplatin),	50				

Trial name	Treatment arm	RT Dose	Inclusion Criteria	Accrual	# Patients	
					Total	Included
TROG 01-04 [12,13]	Preoperative RT	25	CT3 M0 Adenocarcinoma with lower borders within 12 cm of the anal verge ECOG 0-2 Neutrofil count 1.5 109/l; platelet count 100 109/l; bilirubin and ALT 1.5 times the upper limit of normal; and serum creatinine 1.5 times the upper limit of normal	2001-2006	323	154
	Preoperative RT	50				
Polish II [14]	Preoperative RT followed by 4 cycles folfox	25	Primary or locally recurrent rectal cancer involving or abutting adjacent organs or structure (cT4) or a palpably fixed cT3 lesion, Age < 76 WHO <3 M0	-	515	0
	Preoperative CRT (5FU+Oxaliplatin)	50				
Total					9667	3680

Table 2 - Included patient characteristics for both training and validation dataset. The covariate distributions were evaluated by calculating the significance (p<0.05) using Pearson's chi-square (°) or Mann-Whitney test (*)

Variable	Training (N=3087)		Validation (N=773)		
	Value / Count (Percentage)	Missing (Percentage)	Value / Count (Percentage)	Missing (Percentage)	p-value
Age		35 (1.1%)		6 (0.7)	0.71*
Median	62.0		62.0		
Range	22.0 – 82.0		27.0 – 82.0		
Gender		2 (0.06%)		2 (0.26%)	0.20*
Male	2168 (70%)		523 (68%)		
Female	917 (30%)		248 (32.1%)		
RtDose		0 (0%)		0 (0%)	0.67*
Median	45.0		45		
Range	40.9 – 61.2		41.4 – 59.0		
tDistance (cm)		77 (2.5%)		16 (2.1%)	0.36*
Median	6.0		5.5		
Range	0.0 – 20.0		0.0 – 15.0		
Tumor length (cm)		1508 (49%)		376 (49%)	0.63*
Median	4.0		4.0		
Range	0.0 – 37.0		0.0 – 90.0		
Clinical tumor stage		300 (10%)		79 (10%)	0.89°
2	65 (2.1%)		18 (2.3%)		
3	2477 (80%)		619 (80%)		
4	245 (8%)		57 (7.4%)		
Clinical nodal stage		1275 (41.3%)		338 (44%)	0.85°
0	851 (27.6%)		203 (26.3%)		
1	881 (28.6%)		210 (27.1%)		
2	80 (2.6%)		22 (3%)		
Pathological tumor stage		35 (1.1%)		7 (1%)	0.59°
0/1	592 (19.2%)		16 (21.1%)		
2	946 (31%)		223 (29%)		
3	1433 (46.4%)		358 (46.3%)		
4	81 (3%)		22 (3%)		
Pathological nodal stage		29 (1%)		13 (1.7%)	0.54°
0	2158 (70%)		540 (70%)		
1	698 (23%)		178 (23%)		
2	202 (6%)		42 (5.4%)		
Surgery procedure		0 (0%)		0 (0%)	0.65°
APR	1065 (34.5%)		274 (35.5%)		
AR-based	2022 (65.5%)		499 (64.5%)		
Adjuvant chemo		0 (0%)		0 (0%)	1°
5FU-based	1962 (63.5%)		491 (63.5%)		
No chemo	1125 (36.4%)		282 (36.5%)		

Variable	Training (N=3087)		Validation (N=773)		
	Value / Count (Percentage)	Missing (Percentage)	Value / Count (Percentage)	Missing (Percentage)	p-value
Neo-adjuvant chemo		0 (0%)		0 (0%)	0.85°
5FU-based	2444 (79.1%)		615 (80%)		
No chemo	643 (21%)		158 (20%)		
Local recurrence		69 (2%)		20 (26%)	0.50°
Yes	322 (10.4%)		88 (11%)		
No	2696 (87.3%)		665 (86%)		
Distant metastases		64 (2%)		20 (26%)	0.52°
Yes	861 (28%)		205 (26%)		
No	2162 (70%)		548 (71%)		
Survival		4 (0.1%)		4 (0.5%)	0.50°
No	901 (29%)		214 (28%)		
Yes	2182 (71%)		555 (72%)		
FU time local recurrence		69 (2.4%)		20 (2.6%)	0.61*
Median	52.75		52.35		
Range	0 – 212.59		0 – 207.0		
FU time distant metastases		65 (2.3%)		20 (2.6%)	0.94*
Median	48.71		47.18		
Range	0 – 212.59		0 – 207.0		
FU time overall survival		9 (0.3%)		6 (0.8%)	0.65*
Median	55.53		54.77		
Range	0 – 212.59		2.77 – 207		

Model development and validation

The significance in variable distributions differences between training and validation sets were evaluated by using Pearson's chi-square test for categorical variables and Mann-Whitney test for numerical variables. For exploratory purposes, univariate analysis of variables for all outcomes (LR, DM and OS) was performed using a log-rank test on the Kaplan-Meier estimates of 5-year event rates.

To answer our hypothesis, we trained three accelerated failure time (AFT) models, for prediction of local control (LC), distant control (DC) and overall survival (OS). We chose the AFT model over a Cox Proportional Hazards model, as most of the local/distant recurrence events occurred relatively early during follow-up (see Figure 1). This early occurrence of events resulted in a violation of the proportional hazards assumption in the Cox model (all variables have the same influence over time), where the AFT model is more suitable in these kinds of situations.

Variable selection was performed using a stepwise approach, which is a wrapper method exploring different models by stepwise removing/adding variables, re-learning and evaluating the model. Evaluation was performed using the Akaike's Information Criterion (AIC). The AIC measures the error of a model, while penalizing for more input

variables. Hence, it will reduce the chance of overfitting a model by penalizing more complex models.

Final models were tested on the training and validation datasets. Measures for performance were discrimination and calibration; using the Area under the Receiver Operating Curve (AUC) and Brier score statistic, respectively. Discrimination measures whether a model can make a perfect split between positive and negative outcomes. Calibration measures the error between the probabilities, and the actual outcome. Calibration therefore combines both discriminative (best split) and accuracy (how well does the probability compare to the actual outcome). As a resampling method, we bootstrapped (R=1000) our validation to test the standard deviation in performance measures.

Afterwards, we identified three risk groups by extracting the 33th and 66th percentile of the predictions calculated by the model for each outcome at each time-point (24,36,60 and 120 months) to visualize risk group outcomes in a Kaplan-Meijer curve.

For comparison purposes, we also applied the previously developed models on both training and validation datasets and calculated the same performance measures (AUC; Brier score) as for the newly developed models.

R software version 3.3.1 was used for the entire analysis.

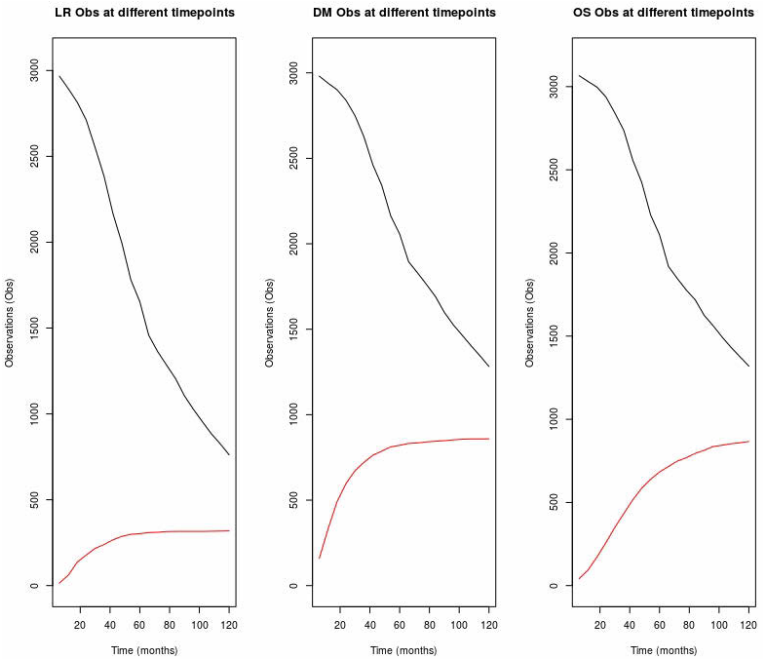


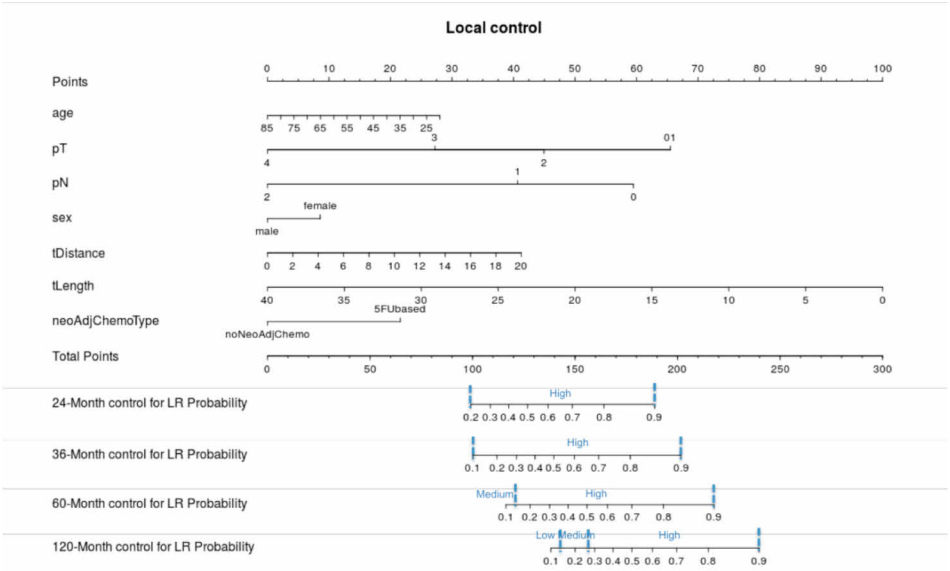
Figure 1 – All three figures show the total number of observations (black line) and the corresponding number of events (red line) at different timepoints. The three figures represent local recurrence (LR), distant metastases (DM) and overall survival (OS). For each outcome, the majority of events occurred within 40 months for LR and DM, and within 60 months for overall survival.

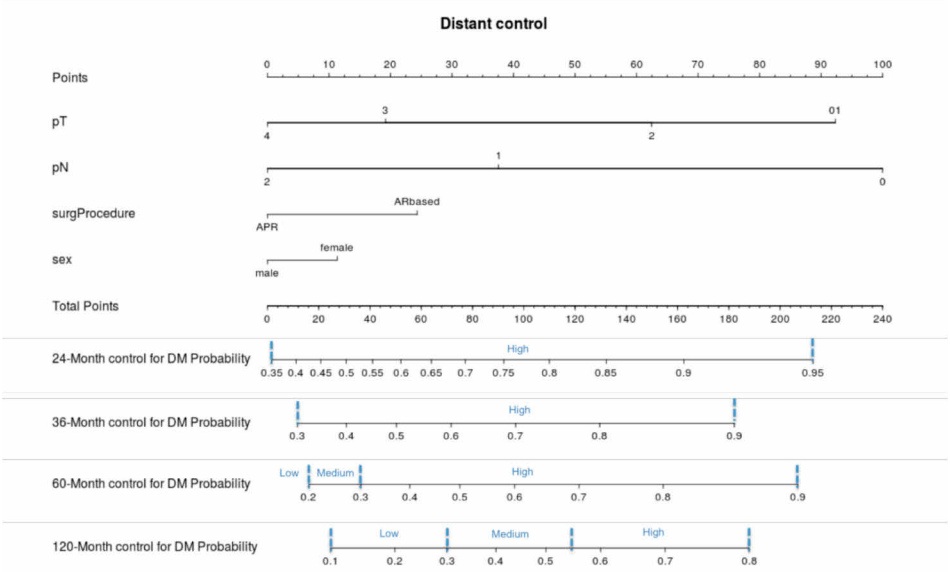
Results

3680 out of 9667 locally advanced rectal cancer patients who underwent pre-operative complete long-course RT, neoadjuvant and/or adjuvant 5-FU based chemotherapy, followed by surgery were analyzed in this study. 3087 patients were used as training set and 773 as validation. The dataset characteristics are shown in Table 2: no statistically significant differences in the distributions of clinical, pathological, treatment and outcome variables were observed. The outcome event rates at 5 years of follow-up were 26% for mortality, 29% for DM and 12% for LR. Median follow-up time increased in comparison to the previous study to 72.5 months for OS, 61 months for LR and 66 months for DM.

Kaplan-Meier curves for statistically significant univariate variables are shown in supplementary materials. Clinical nodal stage and radiotherapy dose were not included in the analysis due to a large number of missing values (41%) and the heterogeneity in the data compilation among trials (see supplemental material for further information).

The nomogram representation of the developed prediction models for LR, DM and OS are shown in Figure 2. The respective performance of these models at different follow-up intervals are shown in Table 3. In comparison to the previous pooled analysis, the predictive role of some variables is different. The clinical tumor stage is not showing its predictive power anymore for any of the three considered outcomes: a possible explanation is that none of the used trials was an RMN based study. Adjuvant chemotherapy appears to be still consistent in the prediction for OS while it is not a strong variable anymore in predicting either distant metastasis or local recurrence. Furthermore, sex has been selected as an important feature in the prediction of all the considered outcomes: LR, DM and OS. Finally, both tumor's length and distance from the anal verge have been identified as important new variables in the prediction of the local recurrence. The rest of the variables included in the nomograms are similar to the previous analysis, with mostly the same predictive power.


Local control
Points
0 10 20 30 40 50 60 70 80 90 100
age
85 75 65 55 45 35 25
pT
3 01
4 2
pN
1
2 0
sex
female
male
tDistance
0 2 4 6 8 10 12 14 16 18 20
tLength
40 35 30 25 20 15 10 5 0
neoAdjChemoType
5FUbased
noNeoAdjChemo
Total Points
0 50 100 150 200 250 300
24-Month control for LR Probability
High
0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
36-Month control for LR Probability
High
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
60-Month control for LR Probability
Medium High
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
120-Month control for LR Probability
Low Medium High
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9


Distant control
Points
0 10 20 30 40 50 60 70 80 90 100
pT
3 01
4 2
pN
1
2 0
surgProcedure
ARbased
APR
sex
female
male
Total Points
0 20 40 60 80 100 120 140 160 180 200 220 240
24-Month control for DM Probability
High
0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8 0.85 0.9 0.95
36-Month control for DM Probability
High
0.3 0.4 0.5 0.6 0.7 0.8 0.9
60-Month control for DM Probability
Low Medium High
0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
120-Month control for DM Probability
Low Medium High
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8

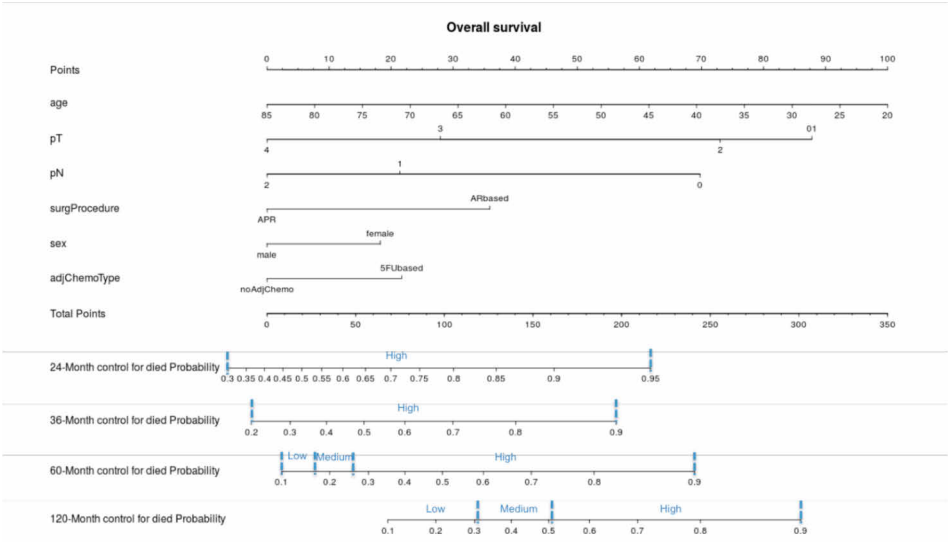


Figure 2 - Nomograms developed for 24, 36, 60 and 120 months of follow-up prediction of local control, distant control and overall survival. The Low, Medium and High categories are describing chance of control/survival, based on a 33% and 66% threshold of the available events at these timepoints.

Table 3 - Performance measurements of the three developed prediction models, specified for 2, 3, 5 and 10 years of follow-up.

	AUC		Brier score (calibration)		Data size	
LR	Original	Current	Original	Current	Original	Current
2-Year	T: 0.74 V: 0.68	0.74 0.66	0.13 0.14	0.06 0.07	T: 2408 V: 594	1304 323
3-Year	T: 0.73 V: 0.72	0.74 0.69	0.19 0.19	0.09 0.10	T: 2114 V: 527	1094 273
5- Year	T: 0.71 V: 0.71	0.75 0.73	0.24 0.22	0.14 0.16	T: 1517 V: 376	675 168
10-Year	T: 0.70 V: 0.68	0.75 0.70	0.25 0.25	0.26 0.29	T: 688 V: 174	235 67
DM						
2-Year	T: 0.75 V: 0.67	0.75 0.68	0.21 0.23	0.14 0.15	T: 2798 V: 694	2801 692
3-Year	T: 0.75 V: 0.67	0.75 0.68	0.25 0.27	0.17 0.19	T: 2585 V: 664	2588 633
5- Year	T: 0.75 V: 0.69	0.74 0.70	0.25 0.27	0.21 0.22	T: 2021 V: 492	2025 491
10-Year	T: 0.75 V: 0.66	0.74 0.67	0.19 0.21	0.24 0.27	T: 1258 V: 300	1262 299
OS						
2-Year	T: 0.71 V: 0.62	0.72 0.65	0.08 0.17	0.15 0.15	T: 2600 V: 649	2876 719
3-Year	T: 0.72 V: 0.67	0.73 0.68	0.12 0.28	0.17 0.19	T: 2426 V: 603	2683 665
5- Year	T: 0.70 V: 0.67	0.72 0.70	0.20 0.36	0.21 0.22	T: 1917 V: 465	2067 505
10-Year	T: 0.70 V: 0.63	0.73 0.64	0.22 0.27	0.24 0.26	T: 1185 V: 284	1289 309

Discussion and conclusion

During the last decade, one of the main oncological challenges has been the investigation of the complex interplay between treatment protocols to explain the different outcomes for specific patients. Most of these investigations have resulted statistical analyses, sometimes accompanied by the development of specific predictive models [29,30,38,39]. The possibility to have predictive models implemented in decision support systems (DSS) and assisting medical doctors in predicting outcomes for specific patients would mean a major step for treatment personalization. Especially in situations where controversy regarding treatment options is apparent, such as the benefit of an adjuvant chemotherapy [40], and whether to individualize treatment modalities and follow-up times.

The goal of this paper was to update the previously developed prediction models for LR, DM and OS in LARC, published by Valentini et al. [29] in a bigger population and including longer follow-up periods. Six more trials (five European and one non-European) have been added to the five LARC clinical trials previously used, where for the original trials the follow-up time was updated. Furthermore, a different analysis has been introduced, to better reflect the relatively early incidence of outcomes.

The primary hypothesis was that the new prediction models have equal or better performance than previously developed and are more robust by using more patients, longer follow-up time and more in-depth statistical analysis. For every outcome (LC, DC, OS) a separate prediction model has been trained. These models were graphically represented as nomograms, showing four different probability scales to predict outcomes at 24, 36, 60 and 120 months of follow-up. On each of these scales, three different risk groups have been identified, as in the previous paper, to be able to better personalize treatments (wait and see, FU-based regimens, combination of more drugs) depending on these three categories (see Figure 2). As a secondary aim, this paper introduces this combined dataset for future, more elaborate analysis.

Updated models

Based on the results shown above, the models have in general a similar performance in comparison to the previously developed models [29]. Although these models were trained on more patients, and for some trials a longer follow-up, it shows that adding more data in this case does not increase model performance. Hence, we would hypothesize that model improvement for future models can only be achieved when including more/different variables or using more advanced machine learning models which can better (or more intuitively) handle existing variables (and for example their interactions). To investigate whether more advanced models would be beneficial in the current pooled dataset, we could use the method described by Deist et al. [41].

In the new models, the total radiotherapy dose (RTdose) variable was not included in the analysis as it was often dependent on different factors (e.g. inclusion criteria of the considered trials and trial arm protocols) causing interactions with other (unobserved however relevant) variables. Furthermore, the definition of rtDose was not always consistent, as it can describe the prescribed, planned or delivered total radiotherapy dose. Hence, we decided to exclude this variable in our current analysis to mitigate inexplicable results. The subsections below will discuss the specific models more in-depth.

Local control prediction

Our revised model for LR includes the following variables: age, pathological T and N, both tumor's length and location (measured as the distance from the anal verge), sex and neoadjuvant chemotherapy. The AFT model shows an increased performance when

predicting 5 and 10 years of follow-up, in comparison to the previous model. Predicting 2 and 3 years of follow-up resulted in having a similar performance in both discrimination and accuracy (see Table 3).

The main difference with the previous nomogram is the loss of influence of adjuvant chemotherapy that, as similar for DM, is no longer included in the new LR prediction's model. Furthermore, the clinical tumor stage is not showing a good prognostic power in the multivariate model anymore. On the contrary, administration of 5-FU based neo-adjuvant chemotherapy (NADCT) were selected as a variable predicting local control (where administering 5-FU based NADCT resulted in better local control). Oxaliplatin-based combinations were excluded from the analysis due to the high variability of the included regimens. Regarding patient characteristics, young patients showed a better prognosis than elderly patients. In regards of the tumor and disease characteristics, both tumor size and its location measured as distance from the anal verge showed prognostic performance in determining local control. This was also found by Yun et al. [42]; the smaller the tumor and the further the tumor is located from the anal verge, the better the prognosis in terms of local control.

Finally, the stratification of high, medium and low-risk patients is even more important in prediction of local control. Patients having a good prognosis (high risk of control, low risk of recurrence) could possibly safely avoid TME in a "wait and watch" approach [2,43].

Distant recurrence/control prediction

In terms of performance, the new distant control model appears consistent with the previously developed model (Table 3). The new model includes different variables, which may be due to the inclusion of more trials (and a larger number of patients) or due to the different method of variable selection and prediction algorithm used. The model resulted in selecting four variables: pathologic T and N stage, surgery procedure and sex.

The main difference that emerges is the exclusion of the adjuvant chemotherapy that is no longer included in the new DM prediction's model. The role of adjuvant CT in patients with LARC treated with CRT and TME is still controversial [44–48]. Despite the current guidelines recommend the usage of adjuvant CT in LARC patients [48], there is limited rectal cancer data available to demonstrate the efficacy of this treatment. Furthermore, recommendations are still mainly based on colon cancer's data [45,47]. Moreover, the compliance of the patients in receiving this regimen is often quite low mainly due to surgery complications (e.g. slow recovery, delayed closure of the temporary ileostomy) causing a reduction in overall survival as well.

Furthermore, consistent with the literature [49,50], gender is becoming a relevant variable in the metastases prediction's process with a worse prognosis for men.

In the last decade, performing neo-adjuvant chemotherapy and radiotherapy has been correlated to pathologic complete response (ypT0N0); where subsequently ypT0N0 has shown a correlation to long-terms outcomes such as local/distant control

[51,52]. This explains why the current models do emphasize the role of the pathological T and N stage in predicting the risk of developing metastases, as measuring pathologic complete response is measured after treatment, and therefore has a better long-term outcome prognosis.

Surgical procedure maintains its predicting role in our latest nomogram for DM. To date total mesorectal excision (TME) is considered the standard surgical technique in rectal cancer. Abdomino-perineal resection (APR) results in higher LR rates and lower 5-year DFS compared with sphincter-saving procedures (SSP). This can be explained by the higher risk of residual cancer after APR and consequently higher risk of DM [53]. To increase a negative margin in the surgical specimen, a new surgical procedure (cylindrical or extralevator APR - ELAPE) has been tested in the last years, especially in distal rectal cancers [2]. Despite the promising results in terms of lower rates of positive CRM and LR, ELAPE technique did not demonstrate the expected clinical benefit. More studies must be conducted to evaluate this new technique.

Overall survival

In comparison to the previous model for overall survival, there appears to be a slightly better performance in the new prediction model for all follow-up time points (Table 3).

The best new model for OS included the following variables: age, pathologic T and N stage, surgery procedure, sex and adjuvant Chemotherapy.

In comparison to the previous model [29], the clinical T stage was excluded during the variable selection. For other variables, the model did not show substantial change. Elderly patients showed a worst survival than younger ones, females appeared have a better prognosis than males, LAR surgery procedures showed a better overall survival than APR and complete response (ypT0N0) after preoperative CRT showed an improved long-term survival. Although still controversial [40,46], 5-FU based adjuvant chemotherapy maintained its relevance in predicting OS. Finally, Oxaliplatin-based combinations were excluded from the analysis due to the high variability of administration variability among the included regimens.

In the last decade, the development of the multimodality treatment has drastically improved the outcomes in patients with LARC. Pathologic complete response (pCR; ypT0N0) remains the most relevant predictor of a low risk of local recurrence. Several studies have been developed in the last years exploring the impact of a radiomic analyses on LR in rectal cancer. Radiomics characterizes tumor phenotypes by extracting multiple quantitative features from radiologic images and provides a comprehensive view of the entire intratumor heterogeneity. Studies have shown that radiomics has potential for predicting pCR after neoadjuvant chemo-radiotherapy [54,55]. The challenge now is to identify different treatment approaches that could maintain or even improve the oncologic outcomes while improving patients' quality of life as well. In this direction, predictive models are useful tools that can help medical doctors in the diffi-

cult process of personalize treatments identifying the best treatment regimens (wait-and-see, follow-up based regimens, combination of more drugs) for each of the three categories.

Data analysis

Due to the nature of clinical trial data, the developed prediction models need extensive validation in routine clinical data to validate both statistical and operational value [56]. Furthermore, several variables resulted in a moderate number of (not randomly) missing values (see Table 2). Although we attempted imputation, we had to exclude specific trials or subgroups of patients; reducing the number of eligible patients for analysis. This resulted in a trade-off between more specific variables in a model however including less patients, or less specific variables and more patients. In this study, we have chosen for the latter case. Our aim was to build prediction/prognostic models which were robust and included longer term follow-up information. Including many specific variables would have negatively impacted this aim.

Next to these variabilities in data recording, different trials applied different treatment protocols using optional (neo-)adjuvant chemotherapy and/or surgery. In combination with different inclusion criteria, the treatment-specific variables (e.g. surgery type, radiotherapy dose, chemotherapy regimen) could be influenced. Based on these insights, we would advise hospitals implementing these models in clinical practice to perform extensive statistical validation [57]. Commissioning of prediction models (similar to treatment modalities) is needed to be aware of hospital population characteristics in comparison to the patients included in this study.

Conclusion

Our study suggests the consistency of the previous identified models in supporting clinical decision making for different outcomes in LARC patients. The analysis of a larger study population, along with longer follow-up information, pointed out a similar performance in terms of discrimination and accuracy for these new prediction models. However, due to longer follow-up information available for some trials, the model stabilized for predictions at these timepoints. By introducing accelerated failure time models which are more geared towards early incidence of outcomes, the models more reliably predict the effect of variables over time. Finally, for each of the nomograms predicting DM, LR or OS, we have developed predictions for multiple follow-up timepoints (24, 36, 60 and 120 months). By reporting multiple predictions and the performance at different timepoints, the models give better guidance to personalized treatments. When implemented and visually represented correctly in decision support systems, shared decision making among physicians and patients can guide the decision of possible treatments, while considering short- and long-term treatment effects.

References

1. Artac M, Korkmaz L, El-Rayes B, Philip PA. An update on the multimodality of localized rectal cancer. *Critical Reviews in Oncology/Hematology* 2016;108:23–32. doi:10.1016/j.critrevonc.2016.10.004
2. Smith JJ, Garcia-Aguilar J. Advances and Challenges in Treatment of Locally Advanced Rectal Cancer. *Journal of Clinical Oncology* 2015;33(16):1797–1808. doi:10.1200/JCO.2014.60.1054
3. Bosset JF, Collette L, Calais G, Mineur L, Maingon P, Radosevic-Jelic L, *et al.* Chemotherapy with preoperative radiotherapy in rectal cancer. *N Engl J Med* 2006;355(11):1114–1123. doi:10.1056/NEJMoa060829
4. Gerard JP. Preoperative Radiotherapy With or Without Concurrent Fluorouracil and Leucovorin in T3-4 Rectal Cancers: Results of FFCD 9203. *Journal of Clinical Oncology* 2006;24(28):4620–4625. doi:10.1200/JCO.2006.06.7629
5. Sauer R, Becker H, Hohenberger W, Rödel C, Wittekind C, Fietkau R, *et al.* Preoperative versus postoperative chemoradiotherapy for rectal cancer. *New England Journal of Medicine* 2004;351(17):1731–1740
6. Bujko K, Nasierowska-Guttmejer A, Wyrwicz L, Malinowska M, Krynski J, Kosakowska E, *et al.* Neoadjuvant treatment for unresectable rectal cancer: An interim analysis of a multicentre randomized study. *Radiotherapy and Oncology* 2013;107(2):171–177. doi:10.1016/j.radonc.2013.03.001
7. Gerard JP, Azria D, Gourgou-Bourgade S, Martel-Lafay I, Hennequin C, Etienne PL, *et al.* Clinical Outcome of the ACCORD 12/0405 PRODIGE 2 Randomized Trial in Rectal Cancer. *Journal of Clinical Oncology* 2012;30(36):4558–4565. doi:10.1200/JCO.2012.42.8771
8. Kapiteijn, E. Klein Kranenbarg, W. E. Total Mesorectal Excision (TME) with or without Preoperative Radiotherapy in the Treatment of Primary Rectal Cancer: Prospective Randomised Trial with Standard Operative and Histopathological Techniques. *The European Journal of Surgery* 1999;165(5):410–420. doi:10.1080/110241599750006613
9. Folkesson J. Swedish Rectal Cancer Trial: Long Lasting Benefits From Radiotherapy on Survival and Local Recurrence Rate. *Journal of Clinical Oncology* 2005;23(24):5644–5650. doi:10.1200/JCO.2005.08.144
10. Sainato A, Cernusco Luna Nunzia V, Valentini V, De Paoli A, Maurizi ER, Lupattelli M, *et al.* No benefit of adjuvant Fluorouracil Leucovorin chemotherapy after neoadjuvant chemoradiotherapy in locally advanced cancer of the rectum (LARC): Long term results of a randomized trial (I-CNR-RT). *Radiotherapy and Oncology* 2014;113(2):223–229. doi:10.1016/j.radonc.2014.10.006
11. Sauer R, Liersch T, Merkel S, Fietkau R, Hohenberger W, Hess C, *et al.* Preoperative Versus Postoperative Chemoradiotherapy for Locally Advanced Rectal Cancer: Results of the German CAO/ARO/AIO-94 Randomized Phase III Trial After a Median Follow-Up of 11 Years. *Journal of Clinical Oncology* 2012;30(16):1926–1933. doi:10.1200/JCO.2011.40.1836
12. Ngan SY, Burmeister B, Fisher RJ, Solomon M, Goldstein D, Joseph D, *et al.* Randomized Trial of Short-Course Radiotherapy Versus Long-Course Chemoradiation Comparing Rates of Local Recurrence in Patients With T3 Rectal Cancer: Trans-Tasman Radiation Oncology Group Trial 01.04. *Journal of Clinical Oncology* 2012;30(31):3827–3833. doi:10.1200/JCO.2012.42.9597
13. Ansari N, Solomon MJ, Fisher RJ, Mackay J, Burmeister B, Ackland S, *et al.* Acute Adverse Events and Postoperative Complications in a Randomized Trial of Preoperative Short-course Radiotherapy Versus Long-course Chemoradiotherapy for T3 Adenocarcinoma of the Rectum: Trans-Tasman Radiation Oncology Group Trial (TROG 01.04). *Annals of Surgery* 2017;265(5):882–888. doi:10.1097/SLA.0000000000001987
14. Bujko K, Wyrwicz L, Rutkowski A, Malinowska M, Pietrzak L, Kryński J, *et al.* Long-course oxaliplatin-based preoperative chemoradiation versus 5 × 5 Gy and consolidation chemotherapy for cT4 or fixed cT3 rectal cancer: results of a randomized phase III study. *Annals of Oncology* 2016;27(5):834–842. doi:10.1093/annonc/mdw062
15. Bujko K, Glimelius B, Valentini V, Michalski W, Spalek M. Postoperative chemotherapy in patients with rectal cancer receiving preoperative radio(chemo)therapy: A meta-analysis of randomized trials comparing surgery ± a fluoropyrimidine and surgery + a fluoropyrimidine ± oxaliplatin. *European Journal of Surgical Oncology (EJSO)* 2015;41(6):713–723. doi:10.1016/j.ejso.2015.03.233

16. Wei Y, Royston P, Tierney JF, Parmar MKB. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data. *Statistics in Medicine* 2015:n/a-n/a. doi:10.1002/sim.6556
17. Burbach JPM, den Harder AM, Intven M, van Vulpen M, Verkooijen HM, Reerink O. Impact of radiotherapy boost on pathological complete response in patients with locally advanced rectal cancer: A systematic review and meta-analysis. *Radiotherapy and Oncology* 2014;113(1):1–9. doi:10.1016/j.radonc.2014. 08.035
18. Cheung WY, Shi Q, O'Connell M, Cassidy J, Blanke CD, Kerr DJ, *et al.* The Predictive and Prognostic Value of Sex in Early-Stage Colon Cancer: A Pooled Analysis of 33,345 Patients from the ACCENT Database. *Clinical Colorectal Cancer* 2013;12(3):179–187. doi:10.1016/j.clcc.2013.04.004
19. Maas M, Nelemans PJ, Valentini V, Das P, Rödel C, Kuo LJ, *et al.* Long-term outcome in patients with a pathological complete response after chemoradiation for rectal cancer: a pooled analysis of individual patient data. *The Lancet Oncology* 2010;11(9):835–844. doi:10.1016/S1470-2045(10)70172-8
20. Smith N, Brown G. Preoperative staging of rectal cancer. *Acta Oncologica* 2008;47(1):20–31. doi:10.1080/02841860701697720
21. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, *et al.* Rapid-learning system for cancer care. *Journal of Clinical Oncology* 2010;28(27):4268–4274
22. 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* 2013;10(1):27–40. doi:10.1038/nrclinonc.2012.196
23. Verseveld M, de Graaf EJR, Verhoef C, van Meerten E, Punt CJA, de Hingh IHJT, *et al.* Chemoradiation therapy for rectal cancer in the distal rectum followed by organ-sparing transanal endoscopic microsurgery (CARTS study). *British Journal of Surgery* 2015;102(7):853–860. doi:10.1002/bjs.9809
24. Habr-Gama A, Gama-Rodrigues J, São Julião GP, Proscurshim I, Sabbagh C, Lynn PB, *et al.* Local Recurrence After Complete Clinical Response and Watch and Wait in Rectal Cancer After Neoadjuvant Chemoradiation: Impact of Salvage Therapy on Local Disease Control. *International Journal of Radiation Oncology*Biology*Physics* 2014;88(4):822–828. doi:10.1016/j.ijrobp.2013.12.012
25. Pucciarelli S, De Paoli A, Guerrieri M, La Torre G, Maretto I, De Marchi F, *et al.* Local Excision After Preoperative Chemoradiotherapy for Rectal Cancer: Results of a Multicenter Phase II Clinical Trial. *Diseases of the Colon & Rectum* 2013;56(12):1349–1356. doi:10.1097/DCR.0b013e3182a2303e
26. van Stiphout RGPM, Valentini V, Buijsen J, Lammering G, Meldolesi E, van Soest J, *et al.* Nomogram predicting response after chemoradiotherapy in rectal cancer using sequential PETCT imaging: A multicentric prospective study with external validation. *Radiotherapy and Oncology* 2014;113(2):215–222. doi:10.1016/j.radonc.2014.11.002
27. Sun Y, Chi P, Lin H, Lu X, Huang Y, Xu Z, *et al.* A nomogram predicting pathological complete response to neoadjuvant chemoradiotherapy for locally advanced rectal cancer: implications for organ preservation strategies. *Oncotarget* 2017;8(40). doi:10.18632/oncotarget.18821
28. Liu Z, Zhang XY, Shi YJ, Wang L, Zhu HT, Tang Z, *et al.* Radiomics Analysis for Evaluation of Pathological Complete Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer. *Clinical Cancer Research* 2017;23(23):7253–7262. doi:10.1158/1078-0432.CCR-17-1038
29. Valentini V, Van Stiphout RG, 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. *Journal of Clinical Oncology* 2011;29(23):3163–3172
30. van Gijn W, van Stiphout RGPM, van de Velde CJH, Valentini V, Lammering G, Gambacorta MA, *et al.* Nomograms to predict survival and the risk for developing local or distant recurrence in patients with rectal cancer treated with optional short-term radiotherapy. *Ann Oncol* 2015. doi:10.1093/ annonc/mdv023
31. Honda M, Oba K, Akiyoshi T, Maeda H, Kashiwabara K, Kanda M, *et al.* Development and validation of a prognostic nomogram for colorectal cancer after radical resection based on individual patient data from three large-scale phase III trials. *Oncotarget* 2017;8(58). doi:10.18632/oncotarget.21845

32. Shen L, van Soest J, Wang J, Yu J, Hu W, Gong YUT, *et al.* Validation of a rectal cancer outcome prediction model with a cohort of Chinese patients. *Oncotarget* 2015
33. Keränen SR, Frasson M, García-Granero E, Navarro S, Campos S, Jordá E, *et al.* Stratification of patients with locally advanced rectal cancer (LARC) treated with preoperative chemoradiation (ChR), according to Valentini's nomograms (VN) and the Neoadjuvant Rectal Score (NAR). External validation in a single Institution. *Annals of Oncology* 2016;27(suppl_6). doi:10.1093/annonc/mdw370.132
34. Bujko K, Nowacki MP, Nasierowska-Guttmejer A, Michalski W, Bebenek M, Kryj M, *et al.* Long-term results of a randomized trial comparing preoperative short-course radiotherapy with preoperative conventionally fractionated chemoradiation for rectal cancer. *British Journal of Surgery* 2006;93(10):1215–1223. doi:10.1002/bjs.5506
35. Glynne-Jones R, Counsell N, Quirke P, Mortensen N, Maraveyas A, Meadows HM, *et al.* Chronicle: results of a randomised phase III trial in locally advanced rectal cancer after neoadjuvant chemoradiation randomising postoperative adjuvant capecitabine plus oxaliplatin (XELOX) versus control. *Annals of Oncology* 2014;25(7):1356–1362. doi:10.1093/annonc/mdu147
36. Valentini V, De Paoli A, Barba MC, Friso ML, Lupattelli M, Rossi R, *et al.* OC-0494: Capecitabine based preoperative chemo-RT in rectal cancer intensified by RT or oxaliplatin: the INTERACT trial. *Radiotherapy and Oncology* 2014;111:S193. doi:10.1016/S0167-8140(15)30599-5
37. Rödel C, Liersch T, Becker H, Fietkau R, Hohenberger W, Hothorn T, *et al.* Preoperative chemoradiotherapy and postoperative chemotherapy with fluorouracil and oxaliplatin versus fluorouracil alone in locally advanced rectal cancer: initial results of the German CAO/ARO/AIO-04 randomised phase 3 trial. *The Lancet Oncology* 2012;13(7):679–687. doi:10.1016/S1470-2045(12)70187-0
38. Zhang J, Cai Y, Hu H, Lan P, Wang L, Huang M, *et al.* Nomogram basing pre-treatment parameters predicting early response for locally advanced rectal cancer with neoadjuvant chemotherapy alone: a subgroup efficacy analysis of FOWARC study. *Oncotarget* 2016;7(4). doi:10.18632/oncotarget.6469
39. Peng J, Ding Y, Tu S, Shi D, Sun L, Li X, *et al.* Prognostic Nomograms for Predicting Survival and Distant Metastases in Locally Advanced Rectal Cancers. *PLoS ONE* 2014;9(8):e106344. doi:10.1371/ journal.pone.0106344
40. Maas M, Nelemans PJ, Valentini V, Crane CH, Capirci C, Rödel C, *et al.* Adjuvant chemotherapy in rectal cancer: Defining subgroups who may benefit after neoadjuvant chemoradiation and resection: A pooled analysis of 3,313 patients: Adjuvant chemotherapy in rectal cancer. *International Journal of Cancer* 2015;137(1):212–220. doi:10.1002/ijc.29355
41. Deist TM, Dankers FJWM, Valdes G, Wijsman R, Hsu IC, Oberije C, *et al.* Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. *Medical Physics* 2018;45(7):3449–3459. doi:10.1002/mp.12967
42. Yun JA, Huh JW, Kim HC, Park YA, Cho YB, Yun SH, *et al.* Local recurrence after curative resection for rectal carcinoma: The role of surgical resection. *Medicine* 2016;95(27):e3942. doi:10.1097/ MD.0000000000003942
43. Dossa F, Chesney TR, Acuna SA, Baxter NN. A watch-and-wait approach for locally advanced rectal cancer after a clinical complete response following neoadjuvant chemoradiation: a systematic review and meta-analysis. *The Lancet Gastroenterology & Hepatology* 2017;2(7):501–513. doi:10.1016/S2468-1253(17)30074-2
44. Milinis K. Adjuvant chemotherapy for rectal cancer: Is it needed? *World Journal of Clinical Oncology* 2015;6(6):225. doi:10.5306/wjco.v6.i6.225
45. Petersen SH, Harling H, Kirkeby LT, Wille-Jørgensen P, Mocellin S. Postoperative adjuvant chemotherapy in rectal cancer operated for cure. *Cochrane Database of Systematic Reviews* 2012. doi:10.1002/ 14651858.CD004078.pub2
46. Breugom AJ, van Gijn W, Muller EW, Berglund A, van den Broek CBM, Fokstuen T, *et al.* Adjuvant chemotherapy for rectal cancer patients treated with preoperative (chemo)radiotherapy and total mesorectal excision: a Dutch Colorectal Cancer Group (DCCG) randomized phase III trial. *Annals of Oncology* 2014. doi:10.1093/annonc/mdu560

47. André T, Boni C, Navarro M, Tabernero J, Hickish T, Topham C, *et al.* Improved Overall Survival With Oxaliplatin, Fluorouracil, and Leucovorin As Adjuvant Treatment in Stage II or III Colon Cancer in the MOSAIC Trial. *Journal of Clinical Oncology* 2009;27(19):3109–3116. doi:10.1200/JCO.2008.20.6771
48. Benson AB, Bekaii-Saab T, Chan E, Chen YJ, Choti MA, Cooper HS, *et al.* Rectal Cancer. *J Natl Compr Canc Netw* 2012;10(12):1528–1564. doi:10.6004/jnccn.2012.0158
49. Yang Y, Wang G, He J, Ren S, Wu F, Zhang J, *et al.* Gender differences in colorectal cancer survival: A meta-analysis: Gender differences and colorectal cancer survival. *International Journal of Cancer* 2017;141(10):1942–1949. doi:10.1002/ijc.30827
50. Qiu M, Hu J, Yang D, Cosgrove DP, Xu R, Qiu M, *et al.* Pattern of distant metastases in colorectal cancer: a SEER based study. *Oncotarget* 2015;6(36):38658–38666. doi:10.18632/oncotarget.6130
51. Capirci C, Valentini V, Cionini L, De Paoli A, Rodel C, Glynne-Jones R, *et al.* Prognostic Value of Pathologic Complete Response After Neoadjuvant Therapy in Locally Advanced Rectal Cancer: Long-Term Analysis of 566 ypCR Patients. *International Journal of Radiation Oncology*Biology*Physics* 2008;72(1):99–107. doi:10.1016/j.ijrobp.2007.12.019
52. Cui J, Fang H, Zhang L, Wu YL, Zhang HZ. Advances for achieving a pathological complete response for rectal cancer after neoadjuvant therapy. *Chronic Diseases and Translational Medicine* 2016;2(1):10–16. doi:10.1016/j.cdtm.2016.06.001
53. Marr R, Birbeck K, Garvican J, Macklin CP, Tiffin NJ, Parsons WJ, *et al.* The Modern Abdominoperineal Excision: The Next Challenge After Total Mesorectal Excision. *Annals of Surgery* 2005;242(1):74–82. doi:10.1097/01.sla.0000167926.60908.15
54. Cusumano D, Dinapoli N, Boldrini L, Chiloiro G, Gatta R, Masciocchi C, *et al.* Fractal-based radiomic approach to predict complete pathological response after chemo-radiotherapy in rectal cancer. *La Radiologia Medica* 2017. doi:10.1007/s11547-017-0838-3
55. Dinapoli N, Barbaro B, Gatta R, Chiloiro G, Casà C, Masciocchi C, *et al.* Magnetic Resonance, Vendor-independent, Intensity Histogram Analysis Predicting Pathologic Complete Response After Radiochemotherapy of Rectal Cancer. *International Journal of Radiation Oncology*Biology*Physics* 2018. doi: 10.1016/j.ijrobp.2018.04.065
56. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. *British Journal of Cancer* 2014;110(3):551–555. doi:10.1038/bjc.2013.725
57. van Soest J, Meldolesi E, van Stiphout R, Gatta R, Damiani A, Valentini V, *et al.* Prospective validation of pathologic complete response models in rectal cancer: transferability and reproducibility. *Medical Physics* 2017

Chapter 12

Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer

Authors
Tim Lustberg, Johan van Soest, Mark Gooding, Devis Peressutti, Paul Aljabar, Judith van der Stoep, Wouter van Elmpt, Andre Dekker

Adapted from
Radiotherapy and Oncology, 2018, volume 126, issue 2, pages 312-317
DOI: 10.1016/j.radonc.2017.11.012

Abstract

Background and purpose: Contouring of organs at risk (OARs) is an important but time consuming part of radiotherapy treatment planning. The aim of this study was to investigate whether using institutional created software-generated contouring will save time if used as a starting point for manual OAR contouring for lung cancer patients.

Material and methods: Twenty CT scans of stage I–III NSCLC patients were used to compare user adjusted contours after an atlas-based and deep learning contour, against manual delineation. The lungs, esophagus, spinal cord, heart and mediastinum were contoured for this study. The time to perform the manual tasks was recorded.

Results: With a median time of 20 min for manual contouring, the total median time saved was 7.8 min when using atlas-based contouring and 10 min for deep learning contouring. Both atlas based and deep learning adjustment times were significantly lower than manual contouring time for all OARs except for the left lung and esophagus of the atlas based contouring.

Conclusions: User adjustment of software generated contours is a viable strategy to reduce contouring time of OARs for lung radiotherapy while conforming to local clinical standards. In addition, deep learning contouring shows promising results compared to existing solutions.

Introduction

The contouring of organs at risk (OAR) and target volumes is an important aspect of treatment planning in radiation oncology. This process is time consuming and the quality of contours depends on the skill level of the observer [1]. Automatic contouring software could potentially speed up the process and improve consistency between observers. There are a number of commercially available products but these are not frequently used in clinical practice [2].

In the last two decades a lot of effort was put into learning new ways to recognize structures in a range of different imaging modalities (CT, PET, and MRI). Approaches range from knowledge-based algorithms such as atlas-based contouring, machine learning and statistical shape and appearance models; region-based methods such as adaptive thresholding, graph cuts and watershed contouring; or a combination of the knowledge- and region-based methods [2]. In products commercially available in 2014, all vendors use a form of atlas-based contouring and approximately half complement this with a model-based method but these are generally limited to certain OARs [2]. Recently, machine learning techniques, and deep learning methods in particular, have become popular for a wider range of tasks. These approaches, based on artificial neural networks, have shown outstanding capabilities, outperforming most classification and regression methods to date. The main advantage of deep learning methods is the ability to automatically learn the most suitable data representation for the task at hand.

The clinical applicability of automatic contouring software is well-reported for regions such as head and neck, breast, and abdomen [2–6]. For the lung, there are a number of studies reporting automatic contouring [7–14]. Some focus solely on a single method and OAR or the Gross Tumor Volume (GTV) and only evaluate the accuracy of the method without any clinical implications such as time gain. Two studies address the usability of atlas-based contouring for the thorax OARs [9–14]. Dolz et al. provides a framework for a region-based contouring technique in clinical practice [13], they advise further investigation into more accurate atlas selection methods to improve the clinical usability. All these methods should be followed by manual correction of the imperfections of the software contouring with the techniques currently available to assess the clinical usability of these methods. In this study, we hypothesize that using a software-generated contour created with an institution specific model as a starting point for OAR contouring will reduce contouring time for lung cancer patients in a clinical setting. For this we evaluate two contouring methods, atlas-based contouring and one novel method based on deep learning in a clinical representative scenario.

Materials and methods

Atlas-based contours

A commercial atlas-based contouring software (Mirada RTx 1.6 and Workflow Box 1.4, Mirada Medical Ltd., Oxford, United Kingdom) was used to automatically generate contours of the OARs. The atlas-based contouring employed 20 stage I NSCLC patients collected from clinical practice at our institute with minimal geometric distortions and small lesion volumes. These atlas patients were contoured by a senior radiotherapy technician specialized in the thorax region using institutional guidelines and carefully inspected by radiation oncologist for correctness.

Deep learning contours

A prototype of deep learning contouring software (''Mirada DLC Expert" prototype, Mirada Medical Ltd., Oxford, United Kingdom) was used to create contours of the OARs. Both automated contouring methods were performed on a standard desktop computer, additionally the DLC used a graphics card to perform the calculations. The prototype uses a deep learning model based on convolutional neural networks [15], a sub-class of deep learning techniques tailored to process imaging data. Convolutional neural networks employ models with a large number of degrees of freedom (in the order of millions), and are therefore able to learn complex non-linear relationships within the imaging data. Training such models requires a large amount of imaging data and uses a backpropagation algorithm based on a stochastic gradient descent to optimize the free parameters [15]. Contours of 450 lung patients were collected from clinical practice and used to train the model.

Patient selection

Twenty consecutive stage I-III NSCLC patients treated in the period January to February 2016 were selected from routine clinical practice and the mid-ventilation phase of a 4D CT scan (Siemens Biograph PET/CT or Sensation Open CT scanner) was used to contour the OARs. The OARs were defined by institutional guidelines and comprised the left lung, right lung, heart, spinal cord, esophagus, and mediastinum.

Contour methods

For this study we created 5 contour sets: A manual contour (MC); an atlas-based contour (AC); a user adjustment of the atlas-based contour (UAC), meaning the atlas-based contour was used as a starting point and adjustments by a radiotherapy technician were allowed to meet institutional guidelines; a deep learning contour (DLC); and a user adjustment of the deep learning contour (UDLC). A single radiotherapy technician performed contouring tasks to prevent any inter-observer variability. Manual contouring

tasks were performed using the software used in clinical practice (Eclipse, version 11.0, Varian, Palo Alto, United States of America). For each of the contouring tasks, the time required for completion was recorded per patient and OAR.

Subjective scoring of contours

The software generated contours were subjectively scored by the technician from one to four. (1) None of the results would form a useful basis for further editing, no time is expected to be saved contouring is expected compared to manual contouring; (2) Some of the results form a useful basis for further editing, little time would be saved contouring is expected compared to manual contouring; (3) Many of the results form a useful basis for further editing, a moderate time saving is expected compared to manual contouring; (4) Most of the results form a useful basis for further editing, a significant time saving is expected compared to manual contouring.

Contour consistency measurement

All contours were exported and analyzed in Matlab 8.6 (The MathWorks Inc., Natick, MA, USA). To assess similarity, the manual contour was compared to the software generated and user adjusted contours using several metrics. To quantify the similarity between the different contour sets, the contours were projected onto a three-dimensional grid matching the dimensions of the corresponding CT scan. The Dice index, which is the volume of the union normalized by the mean of the two volumes, of each OAR in comparison with the manual contour was calculated. The distance between the surfaces of each contour was measured using a nearest neighbor Euclidean distance calculation, the maximum nearest neighbor Euclidean distance, i.e. the Hausdorff distance, were compared. To evaluate differences in the results of the Dice index, Hausdorff distances and contouring time, a ranked Wilcoxon test was performed; p-values smaller than 0.05 were assumed to be statistically significant.

Results

Contouring time and consistency

The total median time saved was 7.8 min [range 2.2 to 13 min] and 10 min [range 5.2–15 min] for the UAC and UDLC respectively with respect to the MC. This is a large reduction compared to the median time required to contour all OARs for a lung case, which was 20 min. An overview of the recorded times per OAR is given in Figure 1. Both lungs and the spinal cord show significant time reductions for the UAC and UDLC ($p < 0.05$), except for the left lung UAC. This was the result of one large outlier in the dataset, where the time needed to contour was 0.9 min and the time to adjust the AC was 3.6

min. The AC was incorrect due to the fact the lung had collapsed (Figure 2A). Repeating the ranked Wilcoxon test without this sample resulted in a significant time gain for the left lung ($p = 0.014$). The DLC remained robust in the case where the lung had collapsed (Figure 2A). The UDLC performs best with a median time to evaluate and correct under one minute for the lungs and spinal cord. Comparing the Dice and Hausdorff distance for the DLC and UDLC to MC results in no significant differences ($p > 0.05$), this supports the assumption that the time needed to adjust these OARs represents the time the technician needs to decide the contour meets clinical guidelines and can be used without major adjustments. An overview of the Dice scores and Hausdorff distances can be found in Figures 3 and 4.

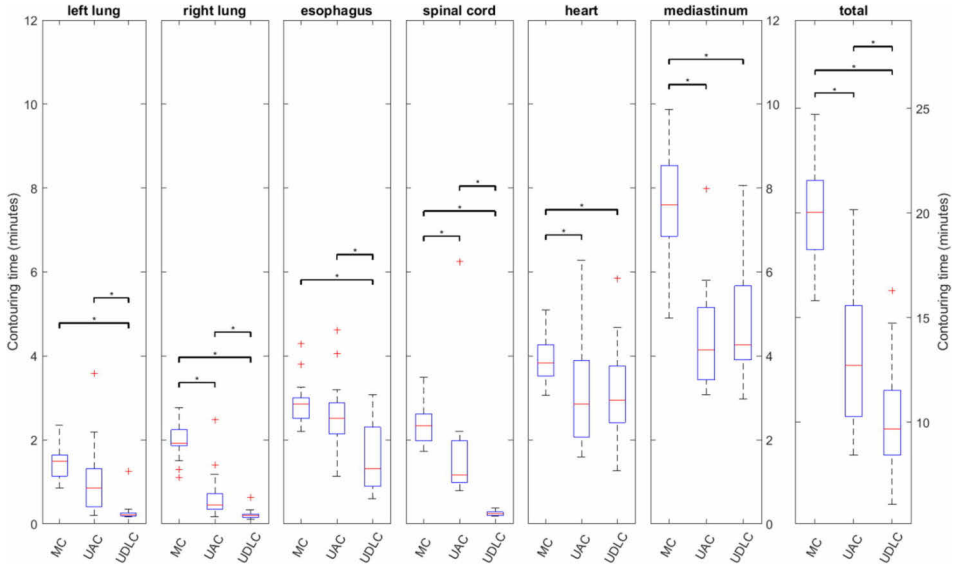


Figure 1 - Contouring time of the manual contour (MC) user adjusted atlas-based contour (UAC) and user adjusted deep learning contour (UDLC) displayed for each OAR and the total time of all OARs. * indicates significant difference between manual and adjustment timing ($p < 0.05$, ranked Wilcoxon test).

The UAC of the esophagus did not result in a significant time reduction with a median time saving of 0.3 min [range 1.9 to 1.9]. By contrast, the UDLC resulted in a significant time reduction of 1.5 min [range 0–3.2]. Figure 2B shows an example slice of the esophagus contour where the DLC matches the MC but the AC is different. The heart and mediastinum showed a significant time reductions for both UAC and UDLC when compared to the MC, however, no significant difference was observed using UDLC over UAC ($p = 0.65$ and $p = 0.32$ for the OARs respectively). Figure 2C shows an example of the heart, where the AC matches the MC but the DLC is different. The mediastinum had a median time reduction 3.4 min [range 0.6–5.7], which is the largest contribution to the overall time reduction. Comparing the AC and DLC to the UAC and UDLC for the esoph-

agus, heart and mediastinum resulted in a significant different time required when compared to the MC ($p < 0.05$).

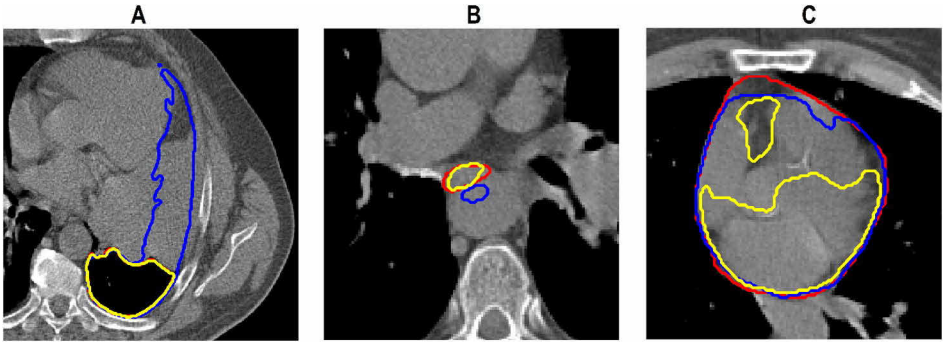


Figure 2 - Example cases showing manual contour (MC, red), atlas-based contour (AC, blue) and the deep learning contour (DLC, yellow) for left lung (A) the esophagus (B) and heart (C).

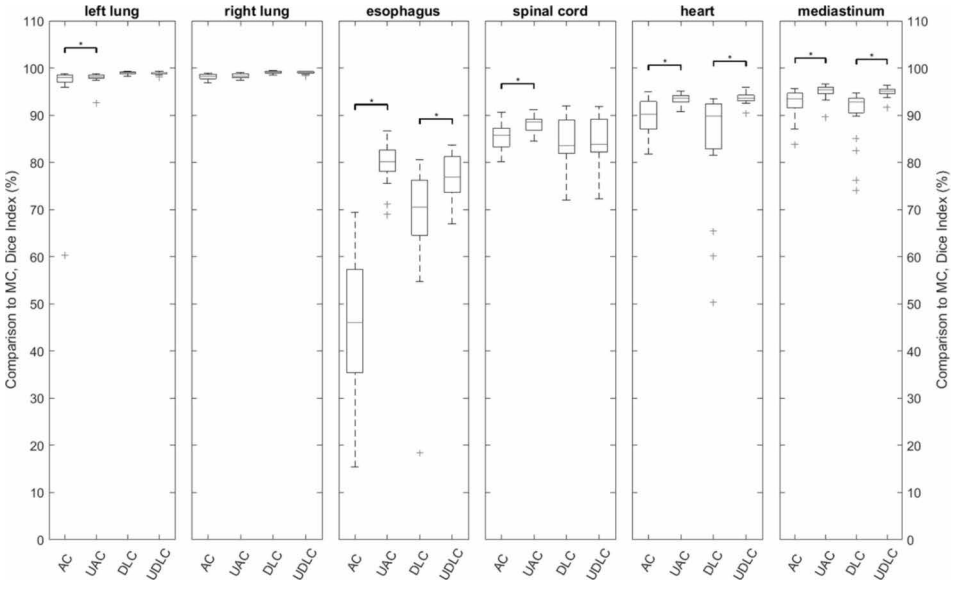


Figure 3 - Dice scores comparing the atlas contour (AC), user adjusted atlas contour (UAC), deep learning contour (DLC) and user adjusted deep learning contour (UDLC) to the manual delineation MC displayed for each OAR. *indicates significant difference between a software generated contour and user adjusted contour ($p < 0.05$, ranked Wilcoxon test)

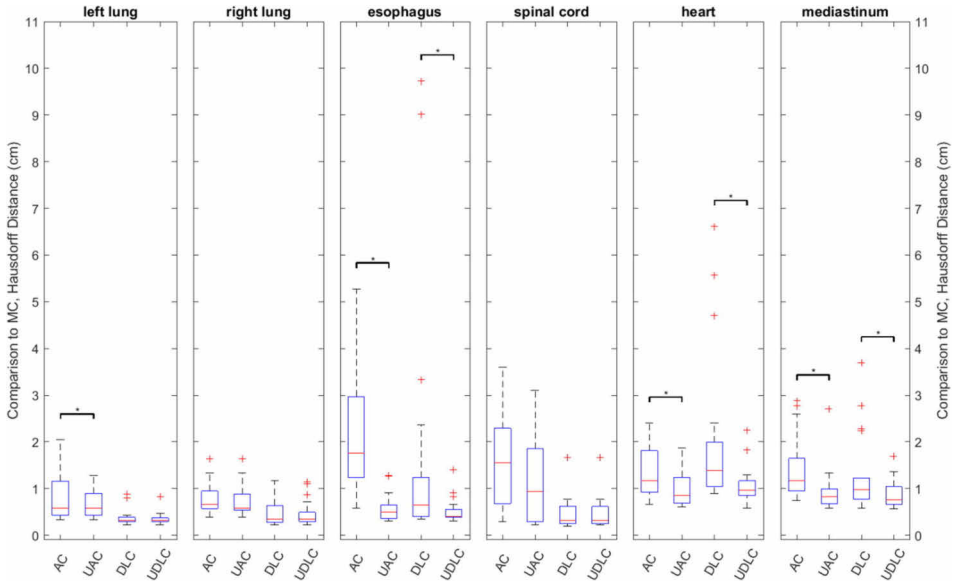


Figure 4 - Hausdorff distance comparing the atlas contour (AC), user adjusted atlas contour (UAC), deep learning contour (DLC) and user adjusted deep learning contour (UDLC) to the manual contour MC displayed for each OAR. *indicates significant difference between a software generated and user adjusted contour ($p < 0.05$, ranked Wilcoxon test).

Subjective scoring of contours

For the lungs and spinal cord, AC performed well according to the subjective score of the technician, with a median score of 4 [range 2–4]. DLC performed even better for these OARs, where every contour was scored as 4. The esophagus AC performed poorly according to the technician with all contours having a score of 1. The DLC was perceived to be an improvement in most cases with a median score of 3 [range 1–4]. The AC of the heart and mediastinum were perceived to be slightly better than the DLC with respective median scores of 3 [range 2–3] and 3 [range 1 to 3]. The OARs with a median subjective score of 4 were consistent (median Dice score >90%, median Hausdorff distance <1.5 cm) when compared against the MC, however the spinal cord performed slightly worse (median Dice score 83%, median Hausdorff distance 1.6 cm).

Grouping the contours per subjective score instead of the corresponding OAR or contouring method resulted in the Dice scores, Hausdorff distances and time saved as shown in Figure 5. Because the median time to contour an OAR varies greatly, e.g. 1.5 min for the left lung compared with 7.6 min for the mediastinum, the time saved was expressed in percentage of the time to manually contour the OAR. Comparing the manual contour to the software generated contours for OARs with a subjective score of 1 resulted in a median Dice score of 57% [range 16%–99%]; a median Hausdorff distance of 2.6 cm [range 0.4 cm–4.5 cm]; and a median time saved of 27% [range 70% to 92%]. Performing the same analysis for the OARs with a subjective score of 4 resulted in a

median Dice score of 98% [range 70%–99%]; a median Hausdorff distance of 0.4 cm [range 0.2 cm–1.3 cm]; and a median time saved of 79% [range 3% to 94%].

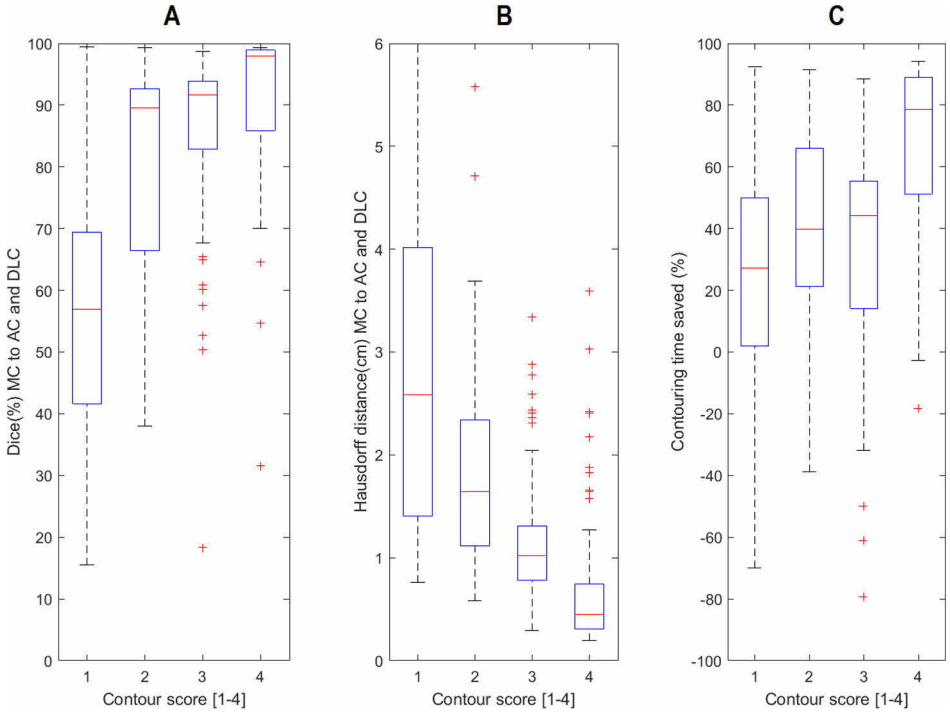


Figure 5 - The comparison results of the manual contour (MC) and the atlas based contour (AC) and deep learning contour (DLC) showing the Dice (A) and the Hausdorff distance (B) each subjective score group. Part C shows the time saved adjusting the software generated contours as a percentage of the manual contour time, displayed for each score group.

Discussion

The methods presented in this study can be used to evaluate any commercially available auto-contouring software. Sharp et al. provide a comprehensive overview of the software that was available a few years ago [2]. Evaluating contour consistency and recording time saved can be used to setup a commissioning protocol to use auto-contouring software in clinical practice and can be used to evaluate the cost-effectiveness of such products. Systematically creating a model for a treatment site, creating a manual, software generated and user adjusted contour and evaluating them will provide the necessary information to safely introduce contouring software into clinical practice. Having a low-tech alternative, i.e. user adjustment strategies as opposed to a software strategy [14], could support the clinical use of auto-contouring software in the near future.

In similar studies, it is shown that auto-contouring techniques can save time [16,17], while other studies show promising new techniques that could potentially save time if investigated [10,18]. Young et al. showed a potential time save for contouring lymph node for pelvic cancer patients [19]. Grouping the Dice scores, Hausdorff distances and time saved for each subjective score group comparing the manual to the user adjusted contours shows a relation between the consistency measurements and the subjective score. This demonstrates the ability of a radiation technician, while accounting for clinical guidelines, to judge whether contour quality is sufficient to lead to time savings compared with routine clinical practice. The technician can delete poor software generated contours and create a manual contour instead. This could from a basis of a viable and principled strategy for introducing imperfect automated contouring methods into clinical practice in order to save time while maintaining local clinical guidelines.

The results observed in the esophagus, heart and mediastinum highlight several issues with knowledge-based contouring software. First, the manual and user adjusted contours were significantly different while both were accepted as contours in compliance with local clinical guidelines by the same observer (intra-observer variability). For instance, the Dice score of the esophagus shows that even after adjusting the software generated contour to meet the clinical guidelines, the median Dice score is 78%, when comparing to the manual contour. Inter-observer variability in contours is reported to be an uncertainty that could greatly improve the quality of radiotherapy plans if reduced [1,3,20]. Van Baardwijk et al. investigated a model-based contouring method for the Gross Tumor Volume (GTV) in CT-PET scans and showed a decrease in Inter-observer variability [7]. Further investigation is needed to determine whether the automated contouring methods as described in this work could potentially reduce the intra- and inter-observer variability for OARs.

Using knowledge-based auto-contouring software, such as atlas-based and deep learning contouring, might improve the consistency of contours created by different observers because the software creates a contour which is the consensus of multiple observers when learning the model. In a future study, multiple observers should be included to determine the areas where they agree and where they do not. Comparing if the differences of the software generated contours happen in these specific parts of the contour will be a better quantification if the atlas is ready for clinical practice. This information, combined with the earlier suggested user adjustment method, could help with an important aspect of bringing automated contour techniques in the clinic, acceptance of these methods by the clinicians [2].

Deep learning is a state of the art machine learning technique that is utilized for many applications [15]. In health care specifically, there are several studies which investigate the utilization of deep learning for image contouring [21–23]. In our study, there are some promising results utilizing deep learning for the automatic contours of OARs for lung cancer patients. The deep learning contouring outperformed the atlas-based contouring for lungs and spinal cord. The deep learning performed better for the

esophagus but further improvements remain necessary. In some cases, the user adjustment of the DLC did not save much time because slices near the stomach were missing and the technician judged these are clinically relevant to include. A similar case can be made for the heart example given in Figure 2C. The clinical training data included substantial differences of opinion in the training data where some observers include or exclude the vessels at the top of the heart differently, therefore the deep learning contouring method is trying to combine these differences of opinion, leading to inaccuracy. For both cases, reaching consensus between observers prior to training DLC could provide improved performance.

As pointed out by Nelms et al. variation in contouring can have a huge impact on the dosimetric properties of a radiotherapy treatment plan for head and neck cancers [24]. In a contouring peer-review of lung cancer, the effect or contouring deviation on the resulting radiation treatment was shown [25]. The latter study discusses the dosimetric differences compared to the clinical guidelines and concludes that further investigation on the actual impact on tumor control and normal tissue toxicity is needed. In our study, we used the assessment of the technician to determine if contours meet clinical guidelines. Further investigation is needed to show the impact of these possible differences in contouring methods on dosimetry and finally treatment outcome.

Comparing the results of the automated contouring methods should be done with caution. The atlases are created from a highly curated set including an expert technician performing the contouring tasks that is validated by the radiation oncologist [26] while the DLC is learned on clinical data, which includes multiple observer preferences and possible imperfections. Comparing these two methods might be considered an unfair comparison, however, in this study we compared the current state of the art of both auto-contouring methods [22,26].

Time saving depends roughly on two main factors, the visualization of the boundary of the organ (e.g. lungs vs. esophagus) and the volume of the OAR to contour (e.g. mediastinal structures). High contrast edges, for instance the lungs, are easier to detect for both software and a human observer while low contrast edges, for instance the esophagus, are much harder. Automatic delineation methods typically are less accurate for small visible soft-tissue boundaries, which increases the time needed for adjustments, on top of this it is also more difficult for the human to distinguish where the contour should be, again increasing the time needed to adjust. Even if we assume that the auto-contouring techniques will reach human level contouring performance in the future, a human observer will probably still need time to evaluate difficult to contour OARs such as the esophagus.

Conclusion

Automatic contouring software as a starting point for clinical contours of OARs in lung radiation therapy allows for a significant time gain when contouring lungs, spinal cord, heart and mediastinum. DLC shows promising results with regard to the creation of institution-based models and to automatically generate high quality contours, providing a greater time saving compared to existing solutions. In addition, clinicians are able to assess if a software generated contour will potentially save time or not.

References

1. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. *Radiother Oncol* 2016;121(2):169–179. doi:10.1016/j.radonc.2016.09.009
2. Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, *et al.* Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Med Phys* 2014;41(5). doi:10.1118/1.4871620
3. Reed VK, Woodward WA, Zhang L, Strom EA, Perkins GH, Tereffe W, *et al.* Automatic Segmentation of Whole Breast Using Atlas Approach and Deformable Image Registration. *Int J Radiat Oncol* 2009;73(5):1493–1500. doi:10.1016/j.ijrobp.2008.07.001
4. Heimann T, Ginneken B van, Styner MA, Arzhaeva Y, Aurich V, Bauer C, *et al.* Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets. *IEEE Trans Med Imaging* 2009;28(8):1251–1265. doi:10.1109/TMI.2009.2013851
5. Stapleford LJ, Lawson JD, Perkins C, Edelman S, Davis L, McDonald MW, *et al.* Evaluation of Automatic Atlas-Based Lymph Node Segmentation for Head-and-Neck Cancer. *Int J Radiat Oncol* 2010;77(3):959–966. doi:10.1016/j.ijrobp.2009.09.023
6. Duc AKH, Eminowicz G, Mendes R, Wong SL, McClelland J, Modat M, *et al.* Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. *Med Phys* 42(9):5027–5034. doi:10.1118/1.4927567
7. van Baardwijk A, Bosmans G, Boersma L, Buijsen J, Wanders S, Hochstenbag M, *et al.* PET-CT–Based Auto-Contouring in Non–Small-Cell Lung Cancer Correlates With Pathology and Reduces Interobserver Variability in the Delineation of the Primary Tumor and Involved Nodal Volumes. *Int J Radiat Oncol* 2007;68(3):771–778. doi:10.1016/j.ijrobp.2006.12.067
8. Luo Y, Liao Z, Jiang W, Gomez D, Williamson R, Court L, *et al.* TU-H-CAMPUS-JeP2-05: Can Automatic Delineation of Cardiac Substructures On Noncontrast CT Be Used for Cardiac Toxicity Analysis? *Med Phys* 43(6Part37):3783–3783. doi:10.1118/1.4957688
9. Kim J, Han J, Ailawadi S, Baker J, Hsia A, Xu Z, *et al.* SU-F-J-113: Multi-Atlas Based Automatic Organ Segmentation for Lung Radiotherapy Planning. *Med Phys* 43(6Part10):3433–3433. doi:10.1118/1.4956021
10. Meng Q, Kitasaka T, Nimura Y, Oda M, Ueno J, Mori K. Automatic segmentation of airway tree based on local intensity filter and machine learning technique in 3D chest CT volume. *Int J Comput Assist Radiol Surg* 2017;12(2):245–261. doi:10.1007/s11548-016-1492-2
11. Kopriva I, Ju W, Zhang B, Shi F, Xiang D, Yu K, *et al.* Single-Channel Sparse Non-Negative Blind Source Separation Method for Automatic 3-D Delineation of Lung Tumor in PET Images. *IEEE J Biomed Health Inform* 2017;21(6):1656–1666. doi:10.1109/JBHI.2016.2624798
12. Rebouças Filho PP, Cortez PC, da Silva Barros AC, C. Albuquerque VH, R. S. Tavares JM. Novel and powerful 3D adaptive crisp active contour method applied in the segmentation of CT lung images. *Med Image Anal* 2017;35:503–516. doi:10.1016/j.media.2016.09.002
13. Dolz J, Kirişli HA, Fechter T, Karnitzki S, Oehlke O, Nestle U, *et al.* Interactive contour delineation of organs at risk in radiotherapy: Clinical evaluation on NSCLC patients. *Med Phys* 43(5):2569–2580. doi:10.1118/1.4947484
14. Schreibmann E, Marcus DM, Fox T. Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search. *J Appl Clin Med Phys* 15(4):22–38. doi:10.1120/jacmp.v15i4.4468
15. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436–444. doi:10.1038/nature14539
16. Haas B, Coradi T, Scholz M, Kunz P, Huber M, Oppitz U, *et al.* Automatic segmentation of thoracic and pelvic CT images for radiotherapy planning using implicit anatomic knowledge and organ-specific segmentation strategies. *Phys Med Biol* 2008;53(6):1751. doi:10.1088/0031-9155/53/6/017
17. Lim JY, Leech M. Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. *Acta Oncol* 2016;55(7):799–806. doi:10.3109/0284186X.2016.1173723

18. Wang J, Guo H. Automatic Approach for Lung Segmentation with Juxta-Pleural Nodules from Thoracic CT Based on Contour Tracing and Correction. *Comput Math Methods Med* 2016. doi:10.1155/2016/2962047
19. Young AV, Wortham A, Wernick I, Evans A, Ennis RD. Atlas-Based Segmentation Improves Consistency and Decreases Time Required for Contouring Postoperative Endometrial Cancer Nodal Volumes. *Int J Radiat Oncol* 2011;79(3):943–947. doi:10.1016/j.ijrobp.2010.04.063
20. Louie AV, Rodrigues G, Olsthoorn J, Palma D, Yu E, Yaremko B, *et al.* Inter-observer and intra-observer reliability for lung cancer target volume delineation in the 4D-CT era. *Radiother Oncol* 2010;95(2):166–171. doi:10.1016/j.radonc.2009.12.028
21. Hu P, Wu F, Peng J, Liang P, Kong D. Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution. *Phys Med Biol* 2016;61(24):8676. doi:10.1088/1361-6560/61/24/8676
22. Mansoor A, Cerrolaza JJ, Perez G, Biggs E, Nino G, Linguraru MG. Marginal Shape Deep Learning: Applications to Pediatric Lung Field Segmentation. *Proc SPIE-- Int Soc Opt Eng* 2017;10133. doi:10.1117/12.2254412
23. Suzuki K. Overview of deep learning in medical imaging. *Radiol Phys Technol* 2017;10(3):257–273. doi:10.1007/s12194-017-0406-5
24. Nelms BE, Tomé WA, Robinson G, Wheeler J. Variations in the Contouring of Organs at Risk: Test Case From a Patient With Oropharyngeal Cancer. *Int J Radiat Oncol • Biol • Phys* 2012;82(1):368–378. doi:10.1016/j.ijrobp.2010.10.019
25. Lo AC, Liu M, Chan E, Lund C, Truong PT, Loewen S, *et al.* The Impact of Peer Review of Volume Delineation in Stereotactic Body Radiation Therapy Planning for Primary Lung Cancer: A Multicenter Quality Assurance Study. *J Thorac Oncol* 2014;9(4):527–533. doi:10.1097/JTO.0000000000000119
26. Peressutti D, Schipaanboord B, Soest J van, Lustberg T, Elmpt W van, Kadir T, *et al.* TU-AB-202-10: How Effective Are Current Atlas Selection Methods for Atlas-Based Auto-Contouring in Radiotherapy Planning? *Med Phys* 43(6Part33):3738–3739. doi:10.1118/1.4957432

Chapter 13

Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data

Authors
Johan van Soest, Chang Sun, Ole Mussmann, Marco Puts, Bob van den Berg, Alexander Malic, Claudia van Oppen, David Townend, Andre Dekker, Michel Dumontier

Adapted from
Studies in Health Technology and Informatics, 2018, Volume 247, pages 581 – 585
DOI: 10.3233/978-1-61499-852-5-581

Abstract

Conventional data mining algorithms are unable to satisfy the current requirements on analyzing big data in some fields such as medicine, policy making, judicial, and tax records. However, applying diverse datasets from different institutes (both healthcare and non-healthcare related) can enrich information and insights. So far, analyzing this data in an automated, privacy-preserving manner does not exist to our knowledge. In this work, we propose an infrastructure, and proof-of-concept for privacy-preserving analytics on vertically partitioned data.

Introduction

Information exchange in the healthcare domain is becoming increasingly important. In first place for clinical purposes such as transfer of care documents among healthcare providers, however also increasingly for secondary use such as development of value-based healthcare and healthcare learning systems. For clinical purposes, information exchange systems are mostly targeted on the exchange of data, e.g. using syntactical standards (e.g. HL7) which facilitate semantic standards (e.g. terminological systems) [1]. For secondary use, purposes translate into e.g. business analytics, obligations to (governmental) registries, and scientific research. Although the use is different, the same clinical standards can be used for secondary purposes.

Although these standards provide transfer of information, they also raise questions about maintainability and ownership, and subsequently security and privacy. By transferring information between multiple health care providers, provenance and authorization become more complex. Furthermore, propagating provenance and authorization updates in all data duplications (e.g. changes in patient consent) becomes a complex task, as all health care providers who received the information need to re-validate their provenance and authorization, or even remove the data from their systems. Furthermore, public confidence regarding data security by large companies has been impaired by recent high-profile breaches (e.g. Equifax breach). Subsequently, policy makers and EU General Data Protection Regulations attempt to increase the requirements for data collection and use, however revise less what alternative options are [2].

One of the alternatives to data transfer is to investigate sending applications containing questions and algorithms to the data source. The goal of this paper is twofold: a) to develop an infrastructure to facilitate transfer and execution of algorithms, b) to apply this infrastructure in a proof-of-concept setup. This proof-of-concept (PoC) will focus on analyzing vertically partitioned data from two institutes. Beyond the scope of this paper, the PoC will be used as a baseline to investigate the causes of onset and progression of Diabetes Mellitus in a population cohort study; including socioeconomic and environmental factors. This paper is further organized into methods, results and conclusion/discussion. The methods section describes the development process of an infrastructure for communicating algorithms and results, called the Personal Health Train (PHT) infrastructure. Furthermore, this paragraph will explain the PoC setup. The results section briefly explains the developed infrastructure (open-source available), and a reference PoC implementation. Finally, the conclusion & discussion will explain our main findings, strengths and weaknesses, and future work.

Methods

The methods below are guided by the scientific question to perform analyses on a population cohort study, enriched with complementary information. Hence, we will first discuss the identification and development of required concepts, and afterwards define the PoC methods used for privacy-preserving processing using complementary data analytics.

Identification of complementary data analytics methods

We started this project by identifying options for complementary data analytics: performing analyses on datasets which have common patients, however have different data elements per patient. The main questions in this identification process were whether data should be transferred, and if yes, with or without patient identifiers. Afterwards, we identified current (commonly used) approaches to complementary data analytics. Finally, we chose the most appropriate starting point for implementation.

Development of the PHT infrastructure

The main goal of the PHT infrastructure is to provide a general-purpose infrastructure, where many different questions can be asked at multiple data owners (e.g. hospitals or even patients themselves). Using such an infrastructure, data owners should have more control over which questions and/or analytics are performed on their data. Furthermore, it should reduce data duplication, and its involved administrative issues [3]. Previously, we have successfully co-developed an infrastructure, which has been adopted by a commercial entity (Varian Learning Portal, Varian Medical Systems, Palo Alto, CA, USA). This system has been successful for distributed machine learning on horizontally partitioned datasets, however is not flexible in terms of analysis tools used, or configuration within hospital infrastructures. In this work, we will continue on previous experience and developed several criteria for the newly developed infrastructure:

- Executing questions at a local institute should be operating system agnostic
- It should facilitate use of different (versions of) libraries
- Communication and computation should be separated
- The communication network should be as light-weight as possible
- IT administration and requirements on the client side should be as limited

One of the consequences of sending algorithms to the data, is that we cannot actually access (and “see”) the data, and have to rely on information given regarding used data structures and systems where they are stored. Previously, we have developed an open-source infrastructure to extract data from clinical systems into standardized formats [4,5].

Proof of Concept (PoC) setup

The developed infrastructure was tested as a PoC in a collaborative information exchange project between a university and the national statistics agency. Specifically, this collaboration targets the vertically partitioned data problem, analyzing data from both participating. In the current PoC, we simulated two datasets:

- At the university: personal identifier and age
- At the statistics agency: personal identifier and income

The datasets were unbalanced in terms of number of patients, where the statistics agency has a large dataset, and the university dataset contains a small subset of patients. This resembles the actual situation, where the statistics agency has more data in comparison to the university. The goal of this PoC was to develop an automated system to plot the relationship between age and income.

Results

Identification of complementary data analytics methods

The final tree of complementary data analytics approaches, as a result from the brainstorm sessions, is shown in Figure 1. The tree in this figure also resulted in a development and validation flow; by starting with data transport and patient identifiers, we will have a validation method for more challenging approaches. In the current PoC, we will use the setup using a trusted third party (TTP) for linking datasets and performing the actual analysis. Using this TTP and appropriate encryption methods, the chances of one party pertaining all datasets and being able to decrypt them are limited (as the TTP cannot retrieve the original patient IDs).

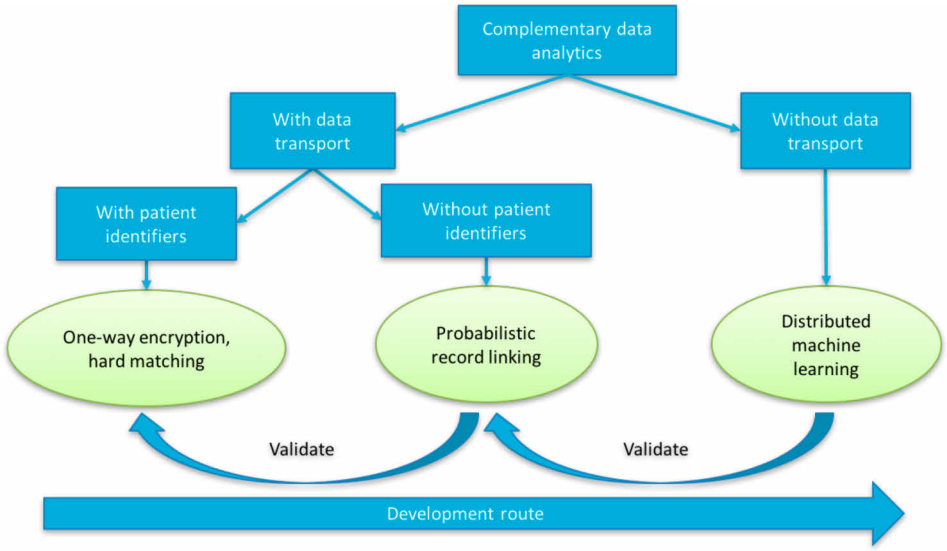


Figure 1 - Approaches for complementary data analytics and the chosen development and validation workflows

Development of the Personal Health Train infrastructure

In our current PoC, we split the Personal Health Train infrastructure into a client-server architecture, connected using the internet (HTTPS). These two developed applications are: a) central message dispatcher and b) client execution application. In this infrastructure, the client execution application (CEA) registers itself at the central message dispatcher (CMD). When successfully registered, the CEA will ask the CMD whether it needs to execute new tasks at regular time intervals. These tasks identify the execution of specific Docker images; hence a task only specifies the docker image identifier and optional (additional) input parameters, stored in a text-based format. The CEA will retrieve the Docker image form the central repository, will append several properties regarding access to the local data source (e.g. intranet URL of the data source), and executes the Docker container. The output of this container should be a text-based result, and is sent back to the CMD. This infrastructure is publicly available at https://github.com/PersonalHealthTrain/PyTaskManager.

Implementation of proof-of-concept

Based on the result of section 3.3, we simulated all involved parties: both institutes, and a TTP. Both institutes installed a CMD, to receive algorithms and work with the simulated data available. At the TTP, a modified version of the CEA was installed to fulfill the role of data receiver.

Three Docker containers were developed: a) for data extraction and encryption of the data at both institutes and b) for decryption at the TTP. The containers sent to the institutes (a) contained queries on the FAIR data sources. The personal identifiers would then be hashed with an agreed-upon salt at both sites. Afterwards the complete dataset would be signed, encrypted and signed, before sending the data to the TTP CEA. Encryption of the dataset was performed using symmetric encryption for performance reasons (symmetric keys are faster for large datasets, in comparison to public key encryption). The symmetric keys were exchanged separately using public key encryption. Finally, when encrypted data was sent, the container produced a positive (message: "OK") result to the CMD that it performed the given task at hand.

After both centers had given a positive result, the final task (container b) was sent to the TTP CEA. First, this container would retrieve the signed and encrypted data from the CEA, and verify-decrypt-verify the data using the securely provided verification and symmetrical decryption key. The first verification was to verify the encryption of the data, and the second signature was to verify the actual dataset. Afterwards, it would perform the actual analysis (merging both datasets), resulting in a scatterplot of age and income of the matched patients in both datasets. Plots can be retrieved manually from the TTP, to ensure a manual validation of anonymity of the results. Docker containers including data were removed by the CEA when execution finished for all locations (data providers & TTP).

Conclusion and Discussion

We have successfully shown that the developed infrastructure and proof-of-concept worked with simulation data. Although the proof-of-concept implementation only shows one case example, the infrastructure can be reused for different questions; both for horizontally and vertically partitioned data.

Limitations of the current work pertain the limited scope of variables in the PoC, and the use of simulation data. Furthermore, the infrastructure will need more security enhancements before actually being implemented in practice. Ethical, legal and societal issues (ELSI) are also of importance in such an infrastructure, however were not scoped in the current prototype. The discussion between ELSI, technical and scientific challenges is a continuous debate among different stakeholders, and evolves over time. Hence, we developed this PoC as input for ELSI discussions, and to show the technical possibilities to the scientific field. Furthermore, our example relies on researchers/analysts which can develop algorithms without actually accessing the data directly. This was not an issue with simulated data in our PoC, however will be addressed using the FAIR principles [6,7] in future work.

Future work will include the discussion and development of an ELSI framework, where the different approaches for complementary data analytics will be discussed, from multiple stakeholder perspectives. In example, some scientific questions can only

be answered with specific technical methods, which have certain ELSI requirements in terms of consent and privacy/security aspects. Likewise, ELSI insights may result in different technical opportunities or scientific directions.

From a technical perspective, future work will pertain further development of this reference infrastructure (e.g. security measures on executing applications), and case examples to use this network. FAIR descriptions of datasets will be part of these case examples, as well as measures to define of FAIR principles.

References

1. Beeler GW. HL7 Version 3—An object-oriented methodology for collaborative standards development. *Int J Med Inf* 1998;48(1):151–161
2. Koops BJ. The trouble with European data protection law. *Int Data Priv Law* 2014;4(4):250–261. doi:10.1093/idpl/ipu023
3. van Soest JA, Dekker AAJ, Roelofs E, Nalbantov G. Application of Machine Learning for Multicenter Learning. In: El Naqa I, Li R, Murphy MJ, editors. *Machine Learning in Radiation Oncology*, Springer International Publishing; 2015
4. van Soest J. *Data-Integration-Tutorial: SWAT4LS Data Integration Tutorial*. 2016
5. Van Soest J, Lustberg T, Grittner D, Marshall MS, Persoon L, Nijsten B, *et al.* Towards a semantic PACS: Using Semantic Web technology to represent imaging data. *Stud Health Technol Inform* 2014;205:166–170
6. Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, *et al.* Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clin Transl Radiat Oncol* 2017;4:24–31. doi:10.1016/j.ctro.2016.12.004
7. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3:160018. doi:10.1038/sdata.2016.18

Chapter 14

General discussion

Analysis of information in Radiation Oncology, and more generic in medicine, requires interaction between the clinical scientific question and technology. This thesis is an example of such interaction, by proposing and implementing an IT (data) architecture (Chapters 3-5), developing and testing components of this architecture (Chapters 6-8), and performing clinical data analysis by using (components of) the developed architecture (chapters 9-12). Although ontologies and explicit semantics are needed for more efficient and reliable research, we have to remind ourselves that explicit recording of all information is a utopian view. It is not feasible for healthcare professionals to record all the requested research data and semantics. Furthermore, for single efforts (may it be from clinical trials or practice), implicit semantics are fine to start with, however this will lead to issues in the longer run (e.g. when workflows, procedures or interventions change in clinical practice). Data will become more diffuse and hard to interpret when merging with other (internal/external) data sources, as seen in Chapter 11. Hence, it means that data quality is in the eye of the beholder, as data quality is defined by the specific use case (e.g. a clinical trial, or re-use of clinical trial data). As a result, clinical data science should not limit itself to data analysis, however should include the infrastructure, tools and provenance of the recorded information. Only when this information is explicitly provided, can we fully understand the output of the conducted data analysis.

Towards a standardized IT architecture for secondary use

To store information with explicit context, an IT architecture facilitating such information is necessary to cover multiple aspects, as introduced in Chapter 3, elaborated upon in Chapter 4 (ontologies) and 5 (technical infrastructure).

Although we would like to think that information standards (for both data and communication) are set in stone, practice learns that standards are organic. They can evolve over time by adding incremental changes to the standard (e.g. DICOM amendments), or are disruptively evolving (e.g. FHIR in comparison to HL7) [1]. Both methods have their place in the medical domain, for clinical *and* secondary use. In the latter scenario, we find that most researchers develop their own (internal) data standards. Especially when receiving datasets from multiple outside sources, the data will likely be in different formats and represented using different terminologies which need to be aligned. This was the case in most clinical data analysis chapters (9-11), and hence a challenge in this thesis.

This brings us to the standardization of medical data, which is a research field in itself. Over time, different syntactical data communication and/or representation standards have been developed both for clinical (e.g. HL7v2, HL7v3, HL7 FHIR) and research purposes (e.g. CDISC-ODM, i2b2, OMOP-CDM) [2,1,3,4]. The clinical standards have always

attempted to be comprehensive for the first and/or second line of clinical care and are designed for rigidity to ensure information is not lost or changed upon transfer [5]. Due to this nature, they have always limited their focus on the actual clinical workflow, and embedded this workflow in these standards (e.g. concepts of appointments, findings and treatments). Hence, the *semantics* in these standards are mostly related to diagnostic, anatomical and/or treatment related terminologies (e.g. International Classification of Diseases or the Systematized Nomenclature of Medicine) [6–10]. For research purposes, the previously mentioned terminologies can be used as well, however research needs more freedom in terms of terminologies used and supported data structures. This freedom creates problems when defining data structures for research purposes. Common data formats, such as CDISC-ODM, i2b2, and OMOP-CDM have specific structures to serve specific research groups or questions. However, when adding unintended data sources, the structures become sub-optimal. Resulting in additional (custom) extensions or plugins to address specific needs [11]. Especially in clinical research – as a form of secondary use of clinical data – standards can be reused but are more and more appended with experimental research information. This experimental information is by definition not standardized, which imposes a flexibility requirement which is not available in clinical standards.

Standardized data representation

This thesis attempts to overcome these challenges by using Semantic Web technologies. By representing both clinical standards and additional (experimental) data using the Resource Description Framework (RDF), we attempt to bridge the data used in clinical practice and (experimental) research. RDF only describes a meta-structure how nodes and edges of a graph are stored, and is therefore domain agnostic. Hence, it allows the inclusion of different data representations. Whether this graph is used to build tables, tree structures, or other (a)cyclic graphs is to be determined by the application developer or data architect. This allows exploratory research to define (part of) the data structure, while being able to represent standardized data as well.

Chapter 6 shows the advantages of this approach, by developing our own Radiation Oncology Ontology (ROO). This ontology reused the International Classification of Diseases (ICD) terminology for disease definitions (e.g. ICD code C20 to define “Neoplastic process of the Rectum”) and the National Cancer Institute’s Thesaurus (NCIT) [12] as a popular and comprehensive terminology in the area of oncology. Although these standards define terminologies, they do not semantically define the relationship between *instances* of these terms in specific situations, which makes a terminology an *ontology* [13]. Hence, this is where we defined our own definitions of relationships in the ROO. Furthermore, the ROO facilitates in defining concepts which are experimental (and hitherto not in clinical vocabularies/terminologies); enabling us to use this experimental information and suggesting where it should fit with commonly known terminologies.

One of the examples is the definition of treatment protocols. Using these treatment protocols, we were able to link patients to trials with similar treatment protocols in Bio2RDF – ClinicalTrials.gov [14,15], and subsequently search for scientific literature of these trials (Chapter 6). Although this system couldn't perform natural language processing or other AI or machine learning techniques, the data representation itself allowed us to connect private and public data sources. Hence, this shows that we need to answer the following question before using/implementing specific technologies: to solve the problem at hand, do we need (computationally intensive) machine learning resources, are traditional approaches sufficient, or do we need a combination of both?

Linking multiple standardized and experimental data sources

One of the data sources which does not fit the clinical and/or research data representation standards is medical imaging information, and its derived information. The ubiquitous standard for medical imaging is Digital Imaging and Communications in Medicine (DICOM), which has evolved over time [16]. Although the core of this standard hasn't changed, new additions have been made over time (e.g. DICOMweb or WADO) to support new generic technologies, such as the world wide web [17]. In this thesis, we proposed additional tooling to convert DICOM into RDF, enabling querying possibilities based on all available metadata, using our developed Semantic DICOM (SeDI) ontology which closely matches the DICOM standard itself (Chapter 7). Although this work mostly provides a syntactical conversion of information, it enables the use of linked data, also for institutional purposes. By converting DICOM metadata into RDF (linked data), we can link the imaging metadata with clinical information, and execute the question "which CT scanning parameters were used for the development of treatment plans, for patients treated in the last two years?" in one query. Before, this question would result in multiple queries to multiple different source systems, using DICOM and SQL syntaxes and dialects.

Based on Chapter 7 we developed a workshop tutorial which shows the benefit of using linked data among data sources *within* a hospital (department) and the ability of the semantic web to incorporate multiple (graph) data structures ad-hoc [18,19]. This workshop also highlighted the possibility to link experimental datasets (e.g. calculations on images) to standardized data, represented as linked RDF data. For example, researchers might calculate experimental data on their own machines, and link it to standardized institutional linked data; creating different levels of data within institutes as well. As an example, experimental radiomics calculations on medical images can be represented in RDF (see https://bioportal.bioontology.org/ontologies/RO) on a researcher's own workstation, and linked to an institutional RDF source holding a representation of clinical (standardized) data. This setup creates a natural distinction between a central and standardized dataset, while being able to extend this dataset with experimental information. Furthermore, as the experimental data resides on a research workstation, the

Semantic Web mantra "Everyone can say Anything about Anything" applies here as well. Experimental data providing additional data about patients in the central database can be used to create new insights, without directly defining them as hard facts in the central database. As soon as these data elements and insights become facts, this information can be transferred to the central database for general use.

What is a standard?

By developing a new ontology (and introducing new terms), we did contribute to the overgrowth of terminological standards. However, we can argue whether developing a new data standard or structure for *internal* purposes is ill-founded. When data structures facilitate more efficient use of computational resources, the new structure might become a standard for specific purposes [20]. Sharing data among different entities (e.g. researchers or institutes) can also be seen as a specific purpose for which explicit data representation is of higher importance than computational resources. Hence, we can argue whether different data standards serve different needs. When explicit data representation is prioritized, aligning with already available terminological systems has an advantage as shown in Chapter 7. The triple-A mantra of the Semantic Web ("Anyone can say Anything about Anything") does not prohibit development of new standards, however, it does pronounce usefulness of reusing existing popular standards to be compatible to others [21–23]. This can be compared to the use of social media; it is more likely we will find a personal page or post on Facebook or LinkedIn than someone using his/her own personal webpage or blog. We can even speculate whether standards are set by popularity within a community, instead of the sophistication and elegance of a standard. Currently, we decided to use existing terminologies while only adding specific concepts which are missing, resulting in the inception of the ROO and SeDI ontologies [24,25] (Chapters 6 and 7, respectively). Due to their inheritance of existing terminologies, the developed ontologies need maintenance. New and popular standards might emerge, where alignment (or conversion to) these new standards might become a necessity. Hence, a standard can be defined when more than one group/entity is using the same definitions, and is bound to a specific purpose.

Clinical impact

Although explicit semantics and data standards create a transparent view of the data at hand, it does not directly contribute to better scientific results or clearer clinical answers. However, it creates the opportunity to investigate the cause of these results or answers.

Semantics in validation

The need for explicit semantics and data standardization becomes apparent when validating (existing) prediction models in other centers (e.g. when validating a model in a geographical different region; Chapter 9). Although the TRIPOD statement [26] provides a guideline what elements of the dataset and methods should be (at minimum) described in a scientific publication, it's only a first step towards transparent reporting of scientific results. Beyond the reporting requirements, there is the question what we should evaluate in specific situations. For example, what do we actually measure when we have validated a model in a different cohort? Using traditional validation approaches, it only shows if the model performs in a different setting. Only when we analyze the differences in datasets (with clear and explicit semantics), we have a better contextual perspective if (or why) a predictive/prognostic model performs (not) well. This method was developed by Debray et al. [27] and was further described and applied for (clinical) use in Chapter 10. This method gives a starting point to investigate which variables are different among cohorts, and subsequently hints towards the underlying (implicit/explicit) semantics in terms of causes for in cohorts and variables. Hence, in combination with more explicit information such as measurement acquisition methods and tools, this method could identify specific issues which usually go undetected and are sometimes blamed unfairly on a "bad" prognostic/predictive model.

As stated in Chapter 10, this validation is not limited to the testing phase of prediction/prognostic models. These methods can also be applied when implementing models in clinical practice, as we need to assure that the real-world clinical dataset shows enough similar characteristics in comparison to the training set, before actual use in clinical practice [28]. As this method uses both training and current data cohorts, it also implies that training data of a prediction/prognostic model should become available to the public, with proper licenses and/or strict terms of use.

Data and models in clinical decision support

Although we can develop prediction models and publish them, this does not mean they will automatically find their way to clinical practice. Historically, systems have been built to implement these models in a knowledge base, and attempt to embed it into the clinical systems and workflows [5]. These Clinical Decision Support (CDS) systems should also use terminological systems, as generalizability of these systems would otherwise be limited [5]. This is another validation aspect when implementing models in clinical practice, however is outside the scope of this thesis: "how many patients in clinical practice have the required information elements to execute a prediction model?" and "how do we embed prediction models into clinical practice?". Especially the latter question is the hardest, as it touches the actual prediction model development. Before development, we should consider carefully how a model would be used in clinical practice, and which

information elements are possibly available at that timepoint? Furthermore, there is the question which *actionable* information elements can we include (e.g. treatments administered *after* the timepoint)? This latter question determines if a model is prognostic (no actionable variables) or predictive (with actionable variables), and subsequently its use and value as well [28].

Although prediction/prognostic models are one of the possibilities to apply machine learning in clinical practice, other techniques can improve clinical practice as well. For example, Chapter 12 shows that there is a benefit of using deep learning for auto-contouring CT scans for clinical practice. This deep learning algorithm is a *black box* solution, where humans actually do not completely comprehend the algorithm due to the cognitive complexity. However, as the computer does not make a direct treatment decision, we do not need to understand the full details. It can already be applied to support clinical practice in reducing workload by automatically performing a time-intensive task (manual delineation of CT scans). The clinical expert still manually examines the automated contours, and can decide to discard or correct the contours manually. In this scenario, the automated delineations can save time on a laborious task, where this (accumulated) time can now be used for other tasks, where human clinical expertise is of higher importance. Hence, we can differentiate machine learning in suggestion or prediction, in addition to using machine learning to try and *explain* relationships in data [29].

Rectal cancer prediction models

When looking more into the application and results of the methods described above, this thesis shows that we are still improving model development and validation methods and measures. Chapter 9 acknowledges the accuracy and generalizability components of model validation as described by Justice et al. [30], and applied these in the external validation of previously developed local recurrence, distant metastases and overall survival prediction models in rectal cancer [31]. This paper showed reasonable performance in terms of discrimination and accuracy on the external validation set, which was geographically different in the context of generalizability and transportability testing. Chapter 10 performed validation in previously developed pathologic complete response (pCR) prediction models for rectal cancer patients [32], and applied more advanced validation measures as mentioned above on a new dataset which was collected after the initial model development. Therefore, this analysis tested reproducibility or transferability on a temporal interval. Here we found that only the model using clinical variables (e.g. from the EMR or TPS system) showed reasonable performance. This was due to sample sizes during training, influenced by PET-CT imaging, which is not routinely performed in clinical practice at all timepoints. Hence, to make final statements on the validity of these PET-CT enabled prediction models would require a larger dataset and repeating the analysis on a larger dataset. Within the limitation of the small da-

taset, our work support previous work by van Stiphout et al. [33] which claimed that the value of the intermediate treatment outcome (pCR) is limited, in comparison to the more final outcomes measured during long-term follow-up. We can argue that predicting on a shorter future time interval is also easier, as models using post-treatment PET scans show better model performance resulting in a smaller time window for exposure to (external) events/influences. Furthermore, pCR is also influenced by the delayed biological effect of radiation therapy. Macchia et al. showed that the time between final fraction of radiotherapy and surgery, influences pCR rates (longer waiting time results in higher percentage of pCR) [34]. Although this study did not show an upper boundary waiting time, it suggested that pCR is time-dependent and hence not a stable variable for outcome prediction when the radiotherapy-surgery interval varies. Hence, when different treatment protocols are in place (with different time intervals between end of radiotherapy and surgery), the incidence of pCR might vary. This results in the problem that current pCR prediction models are aiming to "shoot at a moving target".

Based on these insights, chapter 11 continues on the development of prediction models in an even larger set of rectal cancer patients, collected from a larger number of trials. In this chapter, we found that data quality in trials differ as what the data collection is mandated by the requirements of each trails. If variables were not deemed relevant for a given trial, this trial would not collect these variables. Therefore, our analysis was confronted with many missing values and had to deal with implicit knowledge about the trials. Especially this implicit trial knowledge made it difficult during analysis. For example, when missing units on continuous variables (e.g. a tumor length ranging from 0.1 to 150 within one trial) one cannot easily deduce whether the units are centimeters or millimeters (as 0.1 mm would be too small to measure/detect and suggests centimeters, 150 cm would be too large and would suggest millimeters). Only additional information from the paper or contacting the authors could solve this issue. Another common, however complex, example occurred when information was only available in the publication. For example, the type of neo-adjuvant chemotherapy for a several given trials was implicitly encoded in the randomization arm. Unfortunately, this randomization arm was coded as "0" or "1", meaning that we could not deduce the chemotherapy type. The only option was to check the number of patients for every randomization arm, assuming that there is a difference in randomization arm sizes and datasets exactly correspond to the actual publications. For both examples, not deducing and addressing these problems of implicit knowledge about the data would hamper results of any kind of analysis.

After correcting for this implicit knowledge where possible, we decided to start building prediction models on the largest common denominator: including only variables and trials where treatment protocols were deemed similar enough based on clinical expertise. One of the most common challenges here is the comparison of patients receiving different radiotherapy treatment protocols; some trials investigated the use of administering 25 Gray (Gy) of radiotherapy dose in 5 treatment events (5 Gy per frac-

tion/event), in comparison to the regular treatment dose of 45-50 Gy in 25 treatment events (1.8 Gy per fraction/event). If we would incorporate the radiotherapy dose as an independent variable, we would find that 25Gy result in more local tumor control in comparison to 45-50Gy. This is counter-intuitive to evidence in radiobiology, which invariably finds that a higher radiation dose will induce more cell kill. The key problem here is the nature of a pooled set where variables are assumed independent of the trial they were collected in. By combining many different trials (and their respective inclusion criteria), together with not correcting for these trials/inclusion criteria, we would find these unexpected outcomes.

Unfortunately, this example is not the only one in this rectal cancer dataset. As oncology – and rectal cancer – is treated using many different disciplines and treatments, the interactions between radiotherapy and chemotherapy, the order of treatments and their specific characteristics (e.g. chemotherapy regiments) all influence such a pooled analysis. Hence, one of our conclusions in this work, is that the basic prediction model development tools (e.g. linear regressions, or SVM with linear kernels) cannot develop an elegant and easy-to-interpret prediction model. This conclusion can be extrapolated to analyzing routine clinical data (from multiple institutes) as well. The differences in treatment policy will influence the analysis and its results. To overcome this issue, explicit data provenance is needed for proper interpretation of data, and to perform sound analysis on (routine) clinical data. Incorporating provenance information into prediction models could result in better understanding the interaction of data provenance (including clinical treatment policies) and the actual clinical data. Hence, it would create an opportunity to investigate which machine learning tools could support this process, and give a better understanding of what we are actually measuring and predicting.

Centralized and distributed learning

As mentioned in the introduction, and in the results of Chapter 10, we need platforms to share information to increase the number of cases/patients to increase the statistical reliability of our analysis or prediction models developed.

Hence, we need tools for data extraction from clinical care systems for secondary purposes. Standards and terminologies should preferably be used in the clinical system, or should be added during the extraction process (Chapter 5). Only when we have this data available and represented in a semantically interoperable format, preferably standardized by the hospital itself, can we hope to use centralized or distributed learning. In the latter case, we are not sending individual patient data to a centralized location, however we communicate the output of a specific analysis to a centralized location. Hence, only requests to execute an analysis and the results (on an aggregated level, not individual patient level) are shared/transferred [35,36].

As mentioned in this thesis, distributed learning can mitigate issues regarding data sharing and privacy. Although these advantages are interesting for political and policy reasons, there are still challenges for widespread use. First, the technology stack (e.g. programming languages, tools, applications) needs to become more user friendly. Most of the distributed learning tools are command-line interfaces, and require computer science skills to understand how the underlying infrastructure operates. This could also be an interesting market for companies: developing software and services to hide the infrastructure complexity for distributed learning. Second, distributed learning currently requires a change in paradigm for data analysts and researchers. We have always developed our analysis scripts *with* the data (centralized learning). This meant we could always test if our code would work, and to correct for unexpected behavior. Taking short-cuts and developing custom scripts for custom datasets has become the *modus operandi*, hampering not only data interoperability but also interoperability and reusability of our analyses. With distributed learning, we cannot see the data at remote sites, and have to rely on (FAIR) data specifications. At first, this looks as a limitation of distributed learning: we cannot test on the actual data. However, in the end it would also mean that the interface between the data and the analysis becomes more standardized and therefore increasing the reusability of analysis scripts.

Third, distributed learning does not solve all our non-technical issues regarding political and administrative issues. As we are still using data from other hospitals (without transfer), regulations and interests for multiple stakeholders exists (e.g. rights to publish in scientific communities, rights to use data in commercial applications). Hence, agreements between parties should still be made before actually performing the distributed learning tasks. This challenge is mostly of a political or administrative nature, however by removing the need for physical data sharing the discussion does become easier.

Based on the limitations above, currently the easiest research-oriented use-case for distributed learning is the external validation of a previously developed model (Chapter 9), or learning previously developed (institutional) prediction models on a larger distributed dataset. The latter case has been tested by Deist et al. [35] and Jochems et al. [37]. The exploratory analysis (e.g. finding the variables of influence) is done outside the distributed learning infrastructure, and the network is only used to train and validate the final prediction model. However, as the distributed learning infrastructure builds on the transport and execution of applications, one could develop an application which retrieves the dataset characteristics of all participating centers or even perform a data-driven variable selection (e.g. lasso as implemented by Boyd et al. [38]).

Future of distributed learning: a broader context

Most of the examples in this thesis are addressing the problem of horizontal partitioning: all the same variables are recorded in different hospitals on different patients. This works for in-depth investigation for a medical specialty (e.g. radiation oncology), but only when all the relevant information is recorded in the source systems (e.g. EMR, TPS, RIS or LIS). The problem with data extraction from clinical systems is a chicken-and-egg one, where information is not recorded as it is medically (based on scientific results) irrelevant, however we cannot investigate relevance as this information is not recorded. Hence, it is of importance for research to link datasets with other information sources ranging from custom spreadsheets to wearables (e.g. devices recording personal activities) to national/governmental databases. These additional sources contain information which e.g. relates to socioeconomic, regional, or lifestyle characteristics, which modulate a patient's health and clinical wellbeing. To address this vertical partitioned data problem (adding additional information sources on the same data subjects) using the technologies for distributed learning, we developed a proof-of-concept for vertical data partitioning using a privacy-preserving approach (Chapter 13). This chapter delivers the groundwork for the further development of infrastructure and applications, and serves as a validation approach for more advanced distributed learning methods on vertically partitioned data. At the moment, Chapter 13 focuses on bringing together relevant subsets of data from multiple sources in a neutral 3rd party. Although transferring data to a 3rd party is controversial for a privacy-preserving infrastructure, this method can act as a gold standard to validate more advanced methods (with varying accuracy) which do not require this 3rd party.

Developing standards for information exchange in primary or secondary clinical use settings is not new. It has been a topic for medical informatics departments for many years [5]. However, most emphasis has been on transferring and querying data. In our opinion, the above developed infrastructure could also be used to ask clinical *questions*, encoded in applications. In this view, when institute A requests information from institute B, an application will travel from A to B and will perform the query at institute B. In this scenario, there is an opportunity to include logic, reasoning or other transformations in the application, before the result is sent back to institute A. The underlying mechanism (sending algorithms instead of data, and executing algorithms at the requested hospital) is equal to distributed learning; however, such an infrastructure is not limited anymore to information exchange. It can also enable *interaction* with institutional information systems. For example, patients can send a request (consenting) to make specific information available for a research project, or insert additional information into the hospital EMR system. This results in institutional information systems becoming (FAIR) data stations, which are connected using an infrastructure where applications can be transferred among data stations to ask clinical and research questions,

or to add/modify certain information elements. As a metaphor, we can talk about (FAIR data) stations, (infrastructure) tracks and (application) trains; which is captured in the objectives of the Personal Health Train (PHT) [39].
By reducing data duplication, data stewardship and provenance could become a smaller administrative burden. Sending applications to hospitals increases the chances of research reproducibility [40]. As we cannot see the data and have to implement standards and clean research code, chances are higher to be able to run studies again at a later point in time to include longer follow-up times, or to include more patients. Furthermore, retracting consent for future studies, which is currently hard to implement due to issues of tracking down data storage locations, would be automatically included in the infrastructure. As soon as a patient gives the instruction to withdraw consent, the next research application will not be able to query the information in the FAIR data station.

Although this future looks promising, there is still a lot to be developed and to mature. Chapter 13 is only a starting point for such an infrastructure. Many different research projects both nationally (e.g. Limburg Meet PHT, VWData) [41] and internationally [42] have started on these topics with varying focus. Hence, we have to see how (and if) these initiatives will converge into a national or global, societal accepted infrastructure for information exchange, both for clinical and secondary use of healthcare related information.

References

1. Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, IEEE; 2013
2. Beeler GW. HL7 Version 3—An object-oriented methodology for collaborative standards development. *Int J Med Inf* 1998;48(1):151–161
3. Meineke FA, Stäubert S, Löbe M, Winter A. A comprehensive clinical research database based on CDISC ODM and i2b2. *MIE*, 2014
4. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015;216:574–578
5. Shortliffe EH, Cimino JJ, editors. *Biomedical Informatics*. London: Springer London; 2014. doi:10.1007/978-1-4471-4474-8
6. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med* 1998;37(4–5):394
7. World Health Organization, editor. *International statistical classification of diseases and related health problems*. 10th revision, 2nd edition. Geneva: World Health Organization; 2004
8. Rothwell DJ. SNOMED-based knowledge representation. *Methods Inf Med* 1995;34(1–2):209–213
9. Cornet R, de Keizer N. Forty years of SNOMED: a literature review. *BMC Med Inform Decis Mak* 2008;8(Suppl 1):S2. doi:10.1186/1472-6947-8-S1-S2
10. de Keizer NF, Abu-Hanna A. Understanding terminological systems II: Experience with conceptual and formal representation of structure. *Methods Inf Med* 2000;39(1):22–29
11. Murphy SN, Avillach P, Bellazzi R, Phillips L, Gabetta M, Eran A, *et al.* Combining clinical and genomics queries using i2b2 – Three methods. *PLoS ONE* 2017;12(4). doi:10.1371/journal.pone.0172187
12. Fieschi M. NCI Thesaurus: Using Science-Based Terminology to Integrate Cancer Research Results Sherri de Coronado", Margaret W. Haberb, Nicholas Sioutosc, Mark S. Tuttle'1, Lawrence W. Wright. *Medinfo 2004: Proceedings of the 11th World Congress on Medical Informatics*, Ios Pr Inc; 2004
13. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems I: Terminology and typology. *Methods Inf Med* 2000;39(1):16–21
14. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008;41(5):706–716. doi:10.1016/j.jbi.2008.03.004
15. Dumontier M, Callahan A, Cruz-Toledo J, Ansell P, Emonet V, Belleau F, *et al.* Bio2RDF release 3: a larger connected network of linked data for the life sciences. *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, CEUR-WS. org; 2014
16. Mildenberger P, Eichelberg M, Martin E. Introduction to the DICOM standard. *Eur Radiol* 2002;12(4):920–927. doi:10.1007/s003300101100
17. Lipton P, Nagy P, Sevinc G. Leveraging Internet Technologies with DICOM WADO. *J Digit Imaging* 2012;25(5):646–652. doi:10.1007/s10278-012-9469-3
18. van Soest J, Lustberg T, Marshall MS, Dekker A. Horizontal and vertical medical data federation: Linking clinical and DICOM data using Semantic Web technologies, *SWAT4LS*. Amsterdam: 2016
19. van Soest J. *Data-Integration-Tutorial: SWAT4LS Data Integration Tutorial*. 2016
20. Roelofs E, Persoon L, Nijsten S, Wiessler W, Dekker A, Lambin P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. *Radiother Oncol* 2013;108(1):174–179. doi:10.1016/j.radonc.2012.09.019
21. Berners-Lee T, Hendler J, Lassila O. The semantic web. *Sci Am* 2001;284(5):28–37
22. Shadbolt N, Berners-Lee T, Hall W. The semantic web revisited. *IEEE Intell Syst* 2006;21(3):96–101
23. Allemang D, Hendler J. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. 1 edition. Amsterdam ; Boston: Morgan Kaufmann; 2008
24. Dekker A, Van Soest J, Traverso, Alberto. *Radiation Oncology Ontology*. 2017

25. Grittner D, Van Soest J, Lustberg T, Marshall MS, Dekker A. Semantic DICOM Ontology 2015. http://bioportal.bioontology.org/ontologies/SEDI [accessed March 27, 2018]
26. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015;13(1). doi:10.1186/s12916-014-0241-z
27. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;68(3):279–289. doi:10.1016/j.jclinepi.2014.06.018
28. 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* 2013;10(1):27–40. doi:10.1038/nrclinonc.2012.196
29. Shmueli G. To Explain or to Predict? *Stat Sci* 2010;25(3):289–310. doi:10.1214/10-STS330
30. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130(6):515–524
31. Valentini V, Van Stiphout RG, 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* 2011;29(23):3163–3172
32. 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* 2011;98(1):126–133. doi:10.1016/j.radonc.2010.12.002
33. van Stiphout RGPM, Valentini V, Buijsen J, Lammering G, Meldolesi E, van Soest J, *et al.* Nomogram predicting response after chemoradiotherapy in rectal cancer using sequential PETCT imaging: A multicentric prospective study with external validation. *Radiother Oncol* 2014;113(2):215–222. doi:10.1016/j.radonc.2014.11.002
34. Macchia G, Gambacorta MA, Masciocchi C, Chiloiro G, Mantello G, di Benedetto M, *et al.* Time to surgery and pathologic complete response after neoadjuvant chemoradiation in rectal cancer: A population study on 2094 patients. *Clin Transl Radiat Oncol* 2017;4:8–14. doi:10.1016/j.ctro.2017.04.004
35. Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, *et al.* Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clin Transl Radiat Oncol* 2017;4:24–31. doi:10.1016/j.ctro.2016.12.004
36. Damiani A, Vallati M, Gatta R, Dinapoli N, Jochems A, Deist T, *et al.* Distributed Learning to Protect Privacy in Multi-centric Clinical Studies. In: Holmes JH, Bellazzi R, Sacchi L, Peek N, editors. *Artificial Intelligence in Medicine*, Springer International Publishing; 2015
37. Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, *et al.* Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. *Radiother Oncol* 2016. doi:10.1016/j.radonc.2016.10.002
38. Boyd S. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found Trends® Mach Learn* 2010;3(1):1–122. doi:10.1561/2200000016
39. Dekker, Andre, 't Hoen, Peter-Bram. Personal Health Train. *Dutch Techcentre Life Sci* 2017. https://www.dtls.nl/fair-data/personal-health-train/ [accessed January 5, 2018]
40. Gentleman R, Temple Lang D. Statistical Analyses and Reproducible Research. *J Comput Graph Stat* 2007;16(1):1–23. doi:10.1198/106186007X178663
41. De Thema's – Limburg Meet 2017. https://www.limeproject.nl/de-themas/#!/thema2 [accessed January 5, 2018]
42. Gesundheitscloud 2017. https://hpi.de/open-campus/hpi-initiativen/gesundheitscloud.html [accessed January 5, 2018]

Summary

The amount of data recorded in clinical practice is growing. To facilitate the primary (clinical practice) and secondary use (e.g. scientific research) of this data, we need technical infrastructures which are flexible in terms of data types (e.g. images and structured databases) and scalable in terms of the amount of data (Chapter 2). This thesis mainly focused on the development of such an infrastructure, and the application of several components in this infrastructure for the analysis of treatment effects on outcomes in rectal cancer patients.

Like other fields of engineering, an IT infrastructure needs an architecture where the envisioned infrastructure is described. The high-level concept of such an infrastructure is presented in Chapter 3, where we describe the key components needed. Next to these technical components, a common method to collect data (when to record which variables? How to code these variables? By whom and how recorded?) needs to be defined and is presented in an "umbrella protocol" (Chapter 4). Furthermore, a detailed description of how this "umbrella protocol" should be used in combination with the proposed infrastructure is given Chapter 5. This includes semi-automatic data collection/extraction from clinical data sources (such as an electronic medical record), and the possible application of (distributed) machine learning algorithms.

Although the infrastructure itself is important, specific focus has been targeted towards data representation and interoperability. In this thesis, several chapters have been dedicated to development and (re)use of available data and terminologies in radiotherapy; for example, by linking multiple (institutional) data sources. We have shown that private (patient-specific) data using standardized terminologies can be linked to public knowledge: by using information about the patient's specific treatment protocol, we were able to find scientific articles explaining the results in patients with similar treatment protocols. This specific example was defined in one question (query), by connecting multiple data sources and therefore enriching private data with public knowledge (Chapter 6). Next to public knowledge, different institutional data sources (e.g. medical images and image derived information) can be linked as well. Linking these additional data sources enables more advanced questions for exploratory and investigational analysis (Chapter 7).

Although the need for standardized terminologies is commonly recognized, in practice this only happens at the institutional level. Resulting into issues during automated (statistical) analysis of data across different institutes. To overcome this issue, we developed an application to add standardized terminological concepts to local naming conventions by storing them in an institutional repository (Chapter 8).

Next to handling data, several chapters use components of this infrastructure for clinical data analysis purposes. This thesis describes the validation of three models predicting local recurrence, distant metastases and overall survival in locally-advanced rectal cancer (Chapter 9). These models have been trained on European trials (and thus European

patients), where external validation was performed on the clinical dataset of a Chinese hospital. Hence this validation showed whether the previously developed models were applicable to a geographically different area, including a different kind of data collection (routine clinical data, instead of manually curated trial data used during training). The validation above hypothesized a difference between training and validation datasets due to geographical and population differences, although these differences are hard to identify.

To address this identification issue, and to include more background into validation results, Chapter 10 applies a novel method for detecting (subtle) differences between cohorts. This chapter explains how to measure and interpret the results of an external validation and applied this method to previously developed models predicting for pathologic complete response in rectal cancer patients. Pathologic complete response means that there are no tumor cells found in the pathologic specimen after preoperative radiotherapy. Prediction models could serve as an indication whether surgery is needed, or a more intensive follow-up would be sufficient and would eliminate risks of surgical complications. However, only one previously developed model showed sufficient performance during validation due to small sample sizes and related variability in cohort characteristics.

Based on the validation results in Chapter 9, we conducted a new pooled analysis of rectal cancer trial datasets, including longer follow-up time and more trials (Chapter 11). This chapter shows many aspects related to clinical data science, including semantic interoperability, implicit knowledge in datasets and the question what improves the statistical performance of a prediction model (more data, more variables, or better analysis techniques?). This study showed no significant improvement in model performance (e.g. better prediction of local or distant recurrence after many years) after application of strict inclusion criteria. Hence, we hypothesize that only more advanced machine learning techniques or including additional variables could improve model performance.

Although prediction models are mainly targeted towards prediction of clinical outcomes (e.g. survival, recurrence or development of toxicities), they can be applied for other purposes. For example, Chapter 12 reports on the evaluation of an automatic contouring application of medical images, where a certain type of machine learning model is used to contour specific organs on CT scans. This study showed a reduction in time for medical personnel to contour these images. In the discussion (Chapter 14), we hypothesize whether prediction models not directly interfering with the decision-making process are more easily accepted in clinical practice.

The proposed architecture and infrastructure as presented in chapter 3, is reused in Chapter 13. This chapter describes and presents a method and application to safely combine sensitive (patient) data when strictly needed. This chapter serves as a future

perspective on the proposed infrastructure, suggesting use for specific (research) questions or needs (e.g. temporarily combining data from different institutes regarding the same subject), while safeguarding participants' privacy.

However, to combine these different data sources, we need to make sure terminologies among the different institutes do align. Based on the chapters before, we hypothesize in the general discussion (Chapter 14) whether the medical domain can adhere to one standard. As this level of standardization is currently not present, we suggest how to use FAIR (findable, accessible, interoperable, reusable) principles to address this issue when developing, evaluating or implementing prediction models in for clinical practice.

Samenvatting

In de klinische praktijk wordt een grote en groeiende hoeveelheid data geregistreerd en opgeslagen. Voor primair (klinische zorg) en secundair gebruik (bijv. onderzoek) van deze data, zijn technische infrastructuren nodig die flexibel omgaan met de variëteit aan data (bijv. afbeeldingen en gestructureerde databases), en mee kan groeien met de hoeveelheid data (Hoofdstuk 2). Deze thesis richt zich dan ook op de ontwikkeling van een dergelijke infrastructuur, inclusief de ontwikkeling en toepassing van verschillende componenten binnen deze infrastructuur. Verschillende componenten worden dan ook gebruikt bij de analyse van behandelresultaten in patiënten met endeldarmkanker.

Zoals in vele technische beroepen heeft een IT-infrastructuur ook een architectuur nodig waar de beoogde infrastructuur wordt beschreven. Deze architectuur is dan ook beschreven in Hoofdstuk 3, waarbij een aantal componenten worden uitgelicht. Naast deze technische componenten is er ook een algemene methodiek nodig voor het registreren van gegevens (wanneer worden specifieke gegevens opgeslagen, hoe worden deze gecodeerd, en door wie/hoe?). Dit is beschreven in een "paraplu protocol" (Hoofdstuk 4). Een gedetailleerde beschrijving hoe dit "paraplu protocol" de infrastructuur faciliteert, wordt gegeven in Hoofdstuk 5. Hierbij wordt ook aangegeven hoe specifieke informatie uit klinische databronnen (zoals onder andere een elektronisch patiëntendossier) geëxtraheerd kan worden, waarna het gebruikt kan worden voor (gedistribueerde) data-analyse.

Naast deze technische infrastructuur speelt de data zelf een grote rol: specifiek hoe deze wordt opgeslagen, en de bijbehorende definities van deze data. Een aantal hoofdstukken in deze thesis zijn dan ook gewijd aan het ontwikkelen en hergebruik van beschikbare terminologieën binnen radiotherapie, en het koppelen van verschillende intramurale (binnen bijvoorbeeld een ziekenhuis) en transmurale databronnen. In deze thesis wordt onder andere de mogelijkheid getoond om persoonlijke data te koppelen aan publieke kennis, met gebruik van gestandaardiseerde terminologieën. Hoofdstuk 6 is hiervan een voorbeeld, waarbij de mogelijkheid wordt getoond om op basis van behandelprotocollen de bijbehorende wetenschappelijke artikelen te zoeken, voor individuele patiënten. Dit specifieke voorbeeld toont aan dat het mogelijk is om in één vraagstelling (query) meerdere databronnen te raadplegen, en hierdoor de persoonlijke data te verrijken met publiek beschikbare kennis (Hoofdstuk 6). Naast publieke kennis kan deze techniek worden gebruikt om verschillende intramurale databronnen te koppelen. Het koppelen van deze databronnen betekend dan ook dat meer verschillende data beschikbaar komt, en dus meer geavanceerde vragen en (exploratieve) analyses beantwoord kunnen worden (Hoofdstuk 7).

Hoewel iedereen het eens is dat standaard terminologieën nodig zijn, wordt dit nog niet tussen verschillende instellingen goed vormgegeven. Naast de primaire gezondheidszorg levert dit ook problemen tijdens geautomatiseerde (statistische) analyse van data tussen verschillende centra. Deze thesis beschrijft dan ook de ontwikkeling van een

applicatie die een deel van dit probleem aanpakt. Deze applicatie kan lokale terminologieën omzetten in standaard terminologieën, en hiervoor een database per instelling creëren (Hoofdstuk 8).

Naast de extractie en omgang met data beschrijft deze thesis ook het gebruik van een aantal infrastructuur componenten voor klinische data-analyse. Deze thesis beschrijft dan ook de validatie van drie modellen die lokale recidieven, metastases en overleving voorspellen voor patiënten met endeldarmkanker (Hoofdstuk 9). Deze modellen zijn geleerd op Europese klinische studies (en dus overwegend Europese patiënten), waarbij een externe validatie is uitgevoerd op een dataset uit een Chinees ziekenhuis. Deze validatie toonde dan ook aan of modellen toepasbaar zouden zijn in een ander geografisch gebied, en een andere manier van data verzamelen (data uit de klinische routine, in plaats van handmatig gecontroleerde studie data). De hypothese was dan ook dat er een verschil zou zijn tussen de Europese studie data waarop was geleerd, en de Chinese klinische routine data waarop was gevalideerd; al zijn deze verschillen lastig te identificeren. Dit wordt dan ook in Hoofdstuk 10 aangepakt, waarbij een nieuwe methodiek wordt toegepast om (subtiele) verschillen in cohorten van patiënten te detecteren. In dit hoofdstuk wordt beschreven hoe deze verschillen gekwantificeerd worden, en hoe deze resultaten geïnterpreteerd kunnen worden. Deze methodiek wordt dan ook toegepast op het valideren van voorheen ontwikkelde modellen die pathologisch complete remissie voorspellen bij patiënten met endeldarmkanker. Pathologisch complete remissie betekent dat er, na radiotherapie, geen tumorcellen worden gevonden in het chirurgisch verwijderde endeldarm weefsel. Deze voorspellende modellen kunnen een indicatie geven of chirurgie in bepaalde groepen patiënten nodig is, of een meer intensieve controle na behandeling volstaat. Onze analyse toonde aan dat één voorspellend model een redelijke accuraatheid had. De afname in accuraatheid van de andere twee modellen werd veroorzaakt door een klein aantal patiënten geïncludeerd in deze studie, en de variabiliteit in cohort (en patiënt) karakteristieken.

Gebaseerd op de resultaten in Hoofdstuk 9 is er een nieuwe gecombineerde analyse van verschillende endeldarmkankerstudies uitgevoerd in Hoofdstuk 11. Deze bevat meer trials (en dus meer patiënten), en een langere opvolging van patiënten voor de studies geïncludeerd in Hoofdstuk 9. Dit hoofdstuk laat dan ook meerdere aspecten van klinische data-analyse zien; onder andere semantische interoperabiliteit, impliciete kennis over data(sets) en de vraag hoe een voorspellend model verbeterd kan worden (meer data, meer variabelen of betere analysetechnieken)? De resultaten laten zien dat er geen significante verbetering was in de accuraatheid van het model (bijvoorbeeld een betere voorspelling van lokale recidieven of metastasen 5 jaar na behandeling) na het toepassen van specifieke inclusiecriteria. Hieruit suggereren we dat verbetering in accuraatheid alleen kan worden behaald door het toepassen van meer geavanceerde “machine learning” technieken, of door het includeren van additionele patiënt informatie (variabelen).

Naast het gebruik van voorspellende modellen voor klinische uitkomsten (zoals overleving, recidieven of ontwikkeling van toxiciteiten/bijwerkingen) kunnen voorspellende modellen ook voor andere doeleinden ingezet worden. Hoofdstuk 12 is hiervan een voorbeeld, waarbij intekeningen van organen op medische beelden door een computer automatisch worden uitgevoerd. Deze studie toonde aan dat er een tijdsreductie mogelijk is door het inzetten van deze automatische intekening modellen en software. In de discussie (Hoofdstuk 14) kom dit dan ook terug waarbij een hypothese wordt gevormd of voorspellende modellen die niet direct het beslissingsproces beïnvloeden makkelijker worden geaccepteerd in de klinische praktijk.

De architectuur zoals gepresenteerd in hoofdstuk 3, wordt dan ook verder gebruikt in Hoofdstuk 13. Dit hoofdstuk beschrijft een methode en de toepassing voor het veilig combineren van gevoelige (patiënt)data, in situaties waar het strikt noodzakelijk is. Dit hoofdstuk dient als een toekomstblik op toepassingen die op de voorgestelde infrastructuur uitgevoerd kunnen worden. Een voorbeeld hiervan is het tijdelijk combineren van data vanuit verschillende instellingen over dezelfde persoon, waarbij de privacy van deelnemers wordt gewaarborgd.

Gebaseerd op de hoofdstukken voorafgaand wordt er in Hoofdstuk 14 (algemene discussie) de vraag gesteld of het medisch domein zich kan conformeren aan één terminologie standaard. In een infrastructuur waar we niet de directe patiënt data uitwisselen, maar juist de applicaties die patiënt data verwerken en alleen resultaten terugsturen, is dit van groot belang. In het geval het niet mogelijk is om tot één standaard te komen, geeft dit hoofdstuk een suggestie hoe FAIR (vertaald as Vindbaar, Bereikbaar, Interoperabel en herbruikbaar) principes toegepast kunnen worden om dit probleem te omzeilen. Specifiek tijdens de ontwikkeling, evaluatie en implementatie van voorspellende modellen.

Valorization addendum

The activities related to valorization can belong to several (overlapping) themes: (a) Knowledge dissemination, (b) Societal exploitation and (c) Economic exploitation. In this chapter, we will bring this thesis into perspective of valorization, categorized into these three themes.

Knowledge dissemination

Next to scientific publications, we can perform knowledge dissemination among (academic) peers by different methods. Software products developed in this thesis are in general online available, and ready to try for other researchers. These software products are:

- The Radiation Oncology Ontology (ROO): developed as part of **Chapter 6**, with public examples available (https://github.com/RadiationOncologyOntology/ROO).
- The Semantic DICOM (SeDI) ontology: developed as part of **Chapter 7** to extract metadata from DICOM images, and make the complete metadata available for querying (http://bioportal.bioontology.org/ontologies/SEDI)
- The DataFAIRifier: an open-source platform for data extraction from clinical databases and DICOM images. This platform is based on the insights from **Chapters 4-8** (https://github.com/maastroclinic/DataFAIRifier).
- The PyTaskManager: an open-source infrastructure for distributed learning, and as being used in **Chapter 13** (https://github.com/PersonalHealthTrain/PyTaskManager).
- The SparqlExampleR repository: a codebase showing how to develop distributed learning applications in the statistical programming language R. Currently this application retrieves univariate distributions for all variables available in a cohort, in relation to **Chapter 5** (https://github.com/DistributedRapidLearning/SparqlExampleR)

Although distribution of software applications is a logical consequence of scientific publications, we also need tutorials and trainings to teach peers how to use these software components. One of these examples is a pre-meeting course organized and conducted at the Practical Big Data Workshop 2018 in Ann Arbor, Michigan, United states. During this one-day course, we trained participants how to use the DataFAIRifier to convert clinical and DICOM data (typically used within radiotherapy) into a format useful for scientific research. At the end of course, we were able to perform a distributed learning run among 8 participants. The contents of this workshop are available at https://github.com/jvsoest/PBDW2018_hackathon, whereas the distributed learning was performed using the PyTaskManager, in combination with the code publicly available to perform distributed learning of a support vector model (SVM; a specific type of machine learning model see https://github.com/DistributedRapidLearning/DistributedSVM/tree/feature/docker_dl).

Societal exploitation

Translational research

Although knowledge dissemination is valuable to drive the academic fields of clinical data science and radiotherapy forward, it does not exempt us from applying these techniques in clinical practice. This is also known by the "Valley of Death" of medical innovation, where many software products are developed in research, however never make it to clinical practice. One of the reasons for this "Valley of Death" is the growing divide between researchers (without clinical activities) and clinical practice (without research activities). An umbrella term to perform research to investigate tools for clinical practice is called translational research. In this thesis, we attempted to address translational research two times, by performing clinical evaluations of research (software) products to support clinical practice. For example, Chapter 12 evaluates the time reduction of clinical experts to delineate organs and tissues in medical images, resulting in more time for other tasks. In this example, the direct affected stakeholder is the clinical expert, however indirect the patient and society as well. By saving time on delineations, more time would become available for other critical tasks in the chain of tasks needed for patient treatment.

Societal education

Next to the practical and translational aspect, tools for distributed learning can be a back-end infrastructure to feed decision aids with the most accurate numbers for treatment outcomes. For example, based on the specific patient characteristics, we developed a dashboard to compare current patient characteristics to previous treated patients in multiple hospitals. When results are presented in an understandable representation in terms of treatment (adverse) effects, clinicians can use this dashboard to retrieve a prognosis for a new patient.

Furthermore, various clinical decision aids are in development to educate patients and relatives about the disease at hand, and what treatment options are available, including their consequences (for example, see http://www.beslissamen.nl). This information is currently filled by information from scientific articles, however could in the future be regularly updated with information from a network of hospitals participating in a distributed learning effort.

Societal expectation management

Next to the societal exploitation, we need to consider expectation management of the society as well. Information Technology has changed our society in the last 15-20 years, and it is expected to work for medicine as well. Unfortunately, we are not (yet) able to transfer all patient data between healthcare institutes. This is mainly due to the nature

of medical data. Medical images, produced by a scanner and software, are transferable between institutes. This is due to the available DICOM standard (although with variations in implementations), and the little interference possible or needed by humans. Unfortunately, such a standard does not exist on a detailed level for (structured) electronic medical record data, or departmental information systems (such as an oncology/radiotherapy information system). Every hospital and department can decide how to structure their own data, resulting in incompatible data structures and terminologies. Many initiatives are attempting to address this issue; however, these initiatives are complex and need to consider stakeholders on the executive, medical and technical level, both for internal and transmural alignment.

Economical exploitation

This thesis can be related to four different economical exploitations. Varying customers from hospitals, research institutes to other businesses as well.

First, the clinical evaluation in Chapter 12 was conducted to test the performance of a deep learning contouring tool, developed by a commercial company (MIRADA Medical Ltd., United Kingdom). After this evaluation, this company has brought the product to market. Hence, hospitals can purchase this software, and therefore contributes to the economical theme of valorization.

Second, we co-developed the software to extract metadata from medical (DICOM) images with a commercial company (Sohard Software GmbH, Germany). Enabling us to query the complete metadata at once for a specific patient as described in Chapter 7. This tool is further used in research projects and is marketed as a business-to-business software product, which can be implemented by hospitals, or providers of hospital software (e.g. a Picture Archiving and Communication System software vendor) to enable more advanced questions on imaging metadata.

Third, we developed algorithms (see https://github.com/DistributedRapidLearning) to run on a distributed learning infrastructure developed by an international commercial company (Varian Medical Systems). This infrastructure is publicly available as a free-for-use infrastructure, however requires a “first right to negotiate” to buy/license intellectual property (IP) on the knowledge gathered by researchers using this infrastructure. Hence, it can be an economical exploitation for universities, to finance further scientific research. At the same time, the company (being one of the largest radiotherapy device manufacturers) can catalyze the use of the gained IP, however with the risk of vendor lock-in.

Fourth and maybe foremost, this thesis contributed to the start of a spin-off company (Medical Data Works B.V.), focusing on support, maintenance and hosting of medical

(research) software. Next to the knowledge dissemination described above, healthcare institutes want support contracts for software, to guarantee uptime and operational continuity. Universities in general do not have a core-business of delivering support and maintenance for developed software. This would impact the academic output over time (as researchers are slowed down by the time needed for support/maintenance) and would add a risk of liability for this software. Hence, in this spin-off company we are performing these tasks, relieving the academic institute from this non-scientific burden.

Curriculum vitae

Johan van Soest was born on the 30th of October 1987 in Melick, the Netherlands. After finishing secondary education at Stedelijk Lyceum in Roermond, he studied computer science at Fontys Hogeschool of applied sciences in Eindhoven (graduation 2009). Afterwards, he followed the master's program in Medical Informatics at the University of Amsterdam (graduation 2012). His master thesis focused on scoring system optimization and prediction of in-hospital survival of intensive care patients, conducted at the Amsterdam Medical Centre. This thesis laid the groundwork for his expertise in clinical data science, focusing on genetic algorithms for optimization purposes, and prediction models to support clinical decisions. After his graduation, he started working at MAASTRO Clinic as a technical application manager for the Translational IT (TraIT) imaging work package. In this setting, he helped to setup, maintain and further develop the national imaging archive, which is still operational (http://www.ctmm-trait.nl). After a year, he started in August 2013 the PhD project to develop a distributed infrastructure for rectal cancer patient data analysis. The goal of this infrastructure was to perform clinical data science using an infrastructure where only algorithms and statistical results are shared, instead of data. He contributed to different research projects focusing on data extraction, semantic interoperability and analysis of clinical data collected in radiotherapy. Both for central and distributed analysis of data. From September 2017 he is working as postdoctoral researcher at Maastricht University, focusing on further extending the infrastructure and possibilities of distributed data analysis (the Personal Health Train initiative).

List of publications

E. Meldolesi, J. van Soest, N. Dinapoli, A. Dekker, A. Damiani, M.A. Gambacorta, V. Valentini, An umbrella protocol for standardized data collection (SDC) in rectal cancer: a prospective uniform naming and procedure convention to support personalized medicine, Radiotherapy and Oncology. 112 (2014) 59–62.

J. van Soest, T. Lustberg, D. Grittner, M.S. Marshall, L. Persoon, B. Nijsten, P. Feltens, A. Dekker, Towards a semantic PACS: Using Semantic Web technology to represent imaging data, Stud Health Technol Inform. 205 (2014) 166–170.

E. Meldolesi, J. van Soest, A.R. Alitto, R. Autorino, N. Dinapoli, A. Dekker, M.A. Gambacorta, R. Gatta, L. Tagliaferri, A. Damiani, V. Valentini, VATE: VAlidation of high TEchnology based on large database analysis by learning machine, Colorectal Cancer. 3 (2014) 435–450. doi:10.2217/crc.14.34.

R.G.P.M. van Stiphout, V. Valentini, J. Buijsen, G. Lammering, E. Meldolesi, J. van Soest, L. Leccisotti, A. Giordano, M.A. Gambacorta, A. Dekker, P. Lambin, Nomogram predicting response after chemoradiotherapy in rectal cancer using sequential PETCT imaging: A multicentric prospective study with external validation, Radiotherapy and Oncology. 113 (2014) 215–222. doi:10.1016/j.radonc.2014.11.002.

A. Damiani, M. Vallati, R. Gatta, N. Dinapoli, A. Jochems, T. Deist, J. van Soest, A. Dekker, V. Valentini, Distributed Learning to Protect Privacy in Multi-centric Clinical Studies, in: J.H. Holmes, R. Bellazzi, L. Sacchi, N. Peek (Eds.), Artificial Intelligence in Medicine, Springer International Publishing, 2015: pp. 65–75. http://link.springer.com/ chapter/10.1007/978-3-319-19551-3_8.

J. van Soest, A. Dekker, E. Roelofs, G. Nalbantov, Application of Machine Learning for Multicenter Learning, in: I. El Naqa, R. Li, M.J. Murphy (Eds.), Machine Learning in Radiation Oncology, Springer International Publishing, 2015: pp. 71–97. http://dx.doi.org/10.1007/978-3-319-18305-3_6.

E. Meldolesi, J. van Soest, N. Dinapoli, A. Dekker, A. Damiani, M.A. Gambacorta, V. Valentini, Medicine is a science of uncertainty and an art of probability (Sir W. Osler), Radiotherapy and Oncology. 114 (2015) 132–134. doi:10.1016/j.radonc.2014.12.013.

L. Shen, J. van Soest, J. Wang, J. Yu, W. Hu, Y.U.T. Gong, V. Valentini, Y. Xiao, A. Dekker, Z. Zhang, Validation of a rectal cancer outcome prediction model with a cohort of Chinese patients, Oncotarget. (2015).

T. Lustberg, J. van Soest, A. Jochems, T. Deist, Y. van Wijk, S. Walsh, P. Lambin, A. Dekker, Big Data in Radiation Therapy: challenges and opportunities, The British Journal of Radiology. (2016).

P. Lambin, J. Zindler, B.G.L. Vanneste, L.V. De Voorde, D. Eekers, I. Compter, K.M. Panth, J. Peerlings, R.T.H.M. Larue, T.M. Deist, A. Jochems, T. Lustberg, J. van Soest, E.E.C. de Jong, A.J.G. Even, B. Reymen, N. Rekers, M. van Gisbergen, E. Roelofs, S. Carvalho, R.T.H.

Leijenaar, C.M.L. Zegers, M. Jacobs, J. van Timmeren, P. Brouwers, J.A. Lal, L. Dubois, A. Yaromina, E.J. Van Limbergen, M. Berbee, W. van Elmpt, C. Oberije, B. Ramaekers, A. Dekker, L.J. Boersma, F. Hoebers, K.M. Smits, A.J. Berlanga, S. Walsh, Decision support systems for personalized and participative radiation oncology, Advanced Drug Delivery Reviews. (2016). doi:10.1016/j.addr.2016.01.006.

E. Meldolesi, J. van Soest, A. Damiani, A. Dekker, A.R. Alitto, M. Campitelli, N. Dinapoli, R. Gatta, M.A. Gambacorta, V. Lanzotti, P. Lambin, V. Valentini, Standardized data collection to build prediction models in oncology: a prototype for rectal cancer, Future Oncology. 12 (2016) 119–136. doi:10.2217/fon.15.295.

Q. Cheng, E. Roelofs, B.L.T. Ramaekers, D. Eekers, J. van Soest, T. Lustberg, T. Hendriks, F. Hoebers, H.P. van der Laan, E.W. Korevaar, A. Dekker, J.A. Langendijk, P. Lambin, 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, Radiotherapy and Oncology. 118 (2016) 281–285. doi:10.1016/j.radonc.2015.12.029.

A. Jochems, T.M. Deist, J. van Soest, M. Eble, P. Bulens, P. Coucke, W. Dries, P. Lambin, A. Dekker, Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept, Radiotherapy and Oncology. (2016). doi:10.1016/j.radonc.2016.10.002.

J. van Soest, E. Meldolesi, R. Van Stiphout, R. Gatta, A. Damiani, V. Valentini, P. Lambin, A. Dekker, Prospective validation of pathologic complete response models in rectal cancer: transferability and reproducibility, Medical Physics. (2017).

T.M. Deist, A. Jochems, J. van Soest, G. Nalbantov, C. Oberije, S. Walsh, M. Eble, P. Bulens, P. Coucke, W. Dries, A. Dekker, P. Lambin, Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT, Clinical and Translational Radiation Oncology. 4 (2017) 24–31. doi:10.1016/j.ctro.2016.12.004.

A. Jochems, T.M. Deist, I. El Naqa, M. Kessler, C. Mayo, J. Reeves, S. Jolly, M. Matuszak, R. Ten Haken, J. van Soest, C. Oberije, C. Faivre-Finn, G. Price, D. de Ruysscher, P. Lambin, A. Dekker, Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries, International Journal of Radiation Oncology*Biology*Physics. 99 (2017) 344–352. doi:10.1016/j.ijrobp.2017.04.021.

P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. de Jong, J. van Timmeren, S. Sanduleanu, R.T.H.M. Larue, A.J.G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F.M. Mottaghy, J.E. Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine, Nature Reviews Clinical Oncology. 14 (2017) 749–762. doi:10.1038/nrclinonc.2017.141.

T. Lustberg, J. van Soest, M. Gooding, D. Peressutti, P. Aljabar, J. van der Stoep, W. van Elmpt, A. Dekker, Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer, Radiotherapy and Oncology. (2017). doi:10.1016/j.radonc.2017.11.012.

D. Cusumano, N. Dinapoli, L. Boldrini, G. Chiloiro, R. Gatta, C. Masciocchi, J. Lenkowicz, C. Casà, A. Damiani, L. Azario, J. van Soest, A. Dekker, P. Lambin, M. De Spirito, V. Valentini, Fractal-based radiomic approach to predict complete pathological response after chemo-radiotherapy in rectal cancer, La Radiologia Medica. (2017). doi:10.1007/s11547-017-0838-3.

A. Damiani, C. Masciocchi, L. Boldrini, R. Gatta, N. Dinapoli, J. Lenkowicz, G. Chiloiro, M.A. Gambacorta, L. Tagliaferri, R. Autorino, M. Maria, M.A. Blasi, J. van Soest, A. Dekker, V. Valentini, Preliminary Data Analysis in Healthcare Multicentric Data Mining: a Privacy-preserving Distributed Approach, 14 (2018) 11.

J. van Soest, C. Sun, O. Mussmann, M. Puts, B. van den Berg, A. Malic, C. van Oppen, D. Townend, A. Dekker, M. Dumontier, Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data, Studies in Health Technology and Informatics. (2018) 581–585. doi:10.3233/978-1-61499-852-5-581.

N. Dinapoli, B. Barbaro, R. Gatta, G. Chiloiro, C. Casà, C. Masciocchi, A. Damiani, L. Boldrini, M.A. Gambacorta, M. Dezio, G.C. Mattiucci, M. Balducci, J. van Soest, A. Dekker, P. Lambin, C. Fiorino, C. Sini, F. De Cobelli, N. Di Muzio, C. Gumina, P. Passoni, R. Manfredi, V. Valentini, Magnetic Resonance, Vendor-independent, Intensity Histogram Analysis Predicting Pathologic Complete Response After Radiochemotherapy of Rectal Cancer, International Journal of Radiation Oncology*Biology*Physics. (2018). doi:10.1016/j.ijrobp.2018.04.065.

T.M. Deist, F.J.W.M. Dankers, G. Valdes, R. Wijsman, I.-C. Hsu, C. Oberije, T. Lustberg, J. van Soest, F. Hoebers, A. Jochems, I.E. Naqa, L. Wee, O. Morin, D.R. Raleigh, W. Bots, J.H. Kaanders, J. Belderbos, M. Kwint, T. Solberg, R. Monshouwer, J. Bussink, A. Dekker, P. Lambin, Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers, Medical Physics. 45 (2018) 3449–3459. doi:10.1002/mp.12967.

A. Traverso[*], J. van Soest[*], L. Wee, A. Dekker, The radiation oncology ontology (ROO): Publishing linked data in radiation oncology using semantic web and ontology techniques, Medical Physics. (2018). doi:10.1002/mp.12879.

M.J. Gooding, A.J. Smith, M. Tariq, P. Aljabar, D. Peressutti, J. van der Stoep, B. Reymen, D. Emans, D. Hattu, J. van Loon, M. de Rooy, R. Wanders, S. Peeters, T. Lustberg, J. van Soest, A. Dekker, W. van Elmpt, Comparative evaluation of auto-contouring in clinical practice: a practical method using the Turing Test, Medical Physics. (2018). doi:10.1002/mp.13200.

[*] Authors contributed equally

Acknowledgements – Dankwoord

At first, I want to thank the members of the assessment and defense committee for their time and effort in assessing my dissertation thesis. I do realize the topics covered in this thesis span a broad domain and hence takes effort for proper assessment.

Beste Andre, ik wil je danken voor de mogelijkheid om mijn weg binnen het wetenschappelijk onderzoek voort te zetten. Naast het ontwikkelen van de “hard skills” heb ik ook veel geleerd in de “soft skills”. De manier hoe je (meerdere) projecten (tegelijk) aanpakt, het luisterend oor naar verschillende opinies, en de focus en drijfveer om tot resultaten te komen heb ik veel van geleerd. De mogelijkheid om op verschillende soorten projecten te mogen werken (onder andere TraIT, VATE, SAGE, SeDI, CloudAtlas) in internationale samenwerkingen heeft mede deze brede thesis tot stand laten komen. Ik hoop dat we in de komende jaren nog zeker samen kunnen blijven werken in context van clinical data science, en dat ik mijn steentje bij kan dragen in de groei van jouw onderzoekslijn.

Dear Vincenzo, thank you for the support and clinical input making this thesis possible. I will certainly remember the visits to Rome and late evening discussions regarding the purpose of the validated/developed prediction models. It has certainly broadened my mind how to look at clinical practice, and to bridge the gap between technical and clinical mindsets. I hope we can continue our collaboration in future projects.

Many thanks to all members of the Knowledge Engineering research group at MAASTRO. We’ve grown form a few outliers to a complete group which helps each other out. Special thanks to Tim and Timo as companion PhD students. Tim, many IT tasks we have done together, and you’ve shown your leadership in the software development team. I hope we can truly valorise this work in our spin-off company. Timo, I look back at many interesting discussions on why we’re doing this work, and mathematical/statistical discussions on problems at hand.

Alberto, Ananya, Anshu, Arthur, Biche, Cheryl, Frank, Hubert, Ivan, Jonathan, Leonard, Peter, Petros, Rianne, Roel, Sander, Scott, Sean, Sonia, Tim (Hendriks), Zhenwei, thank you!

Also, many thanks to the main collaborator in this thesis, the Knowledge-Based Oncology lab at the Gemelli hospital in Rome. Elisa, Carlotta and Roberto, it has been nice working with you. And the nice non-touristic guides for real authentic Roman experiences.

I’d like to thank all other international collaborators as well. It would become a too long list to thank everyone. However, do know that I’m grateful for all the efforts and opportunities provided.

Heel veel dank aan MAASTRO voor het faciliteren in vele aspecten. I&S (onder andere Joeri en Natascha) dank voor de ondersteuning en het meedenken. Ik denk dat er vaker vragen binnen zijn gekomen die niet zo gebruikelijk waren. Erik, dank voor het wegwijs

maken binnen MAASTRO in de eerste maanden, en het inwerken in de TraIT activiteiten. Wouter, dank voor het nuchtere voorbeeld en de ervaringen als postdoc en senior scientist. Waarschijnlijk ga ik daar de komende jaren nog veel aan hebben!

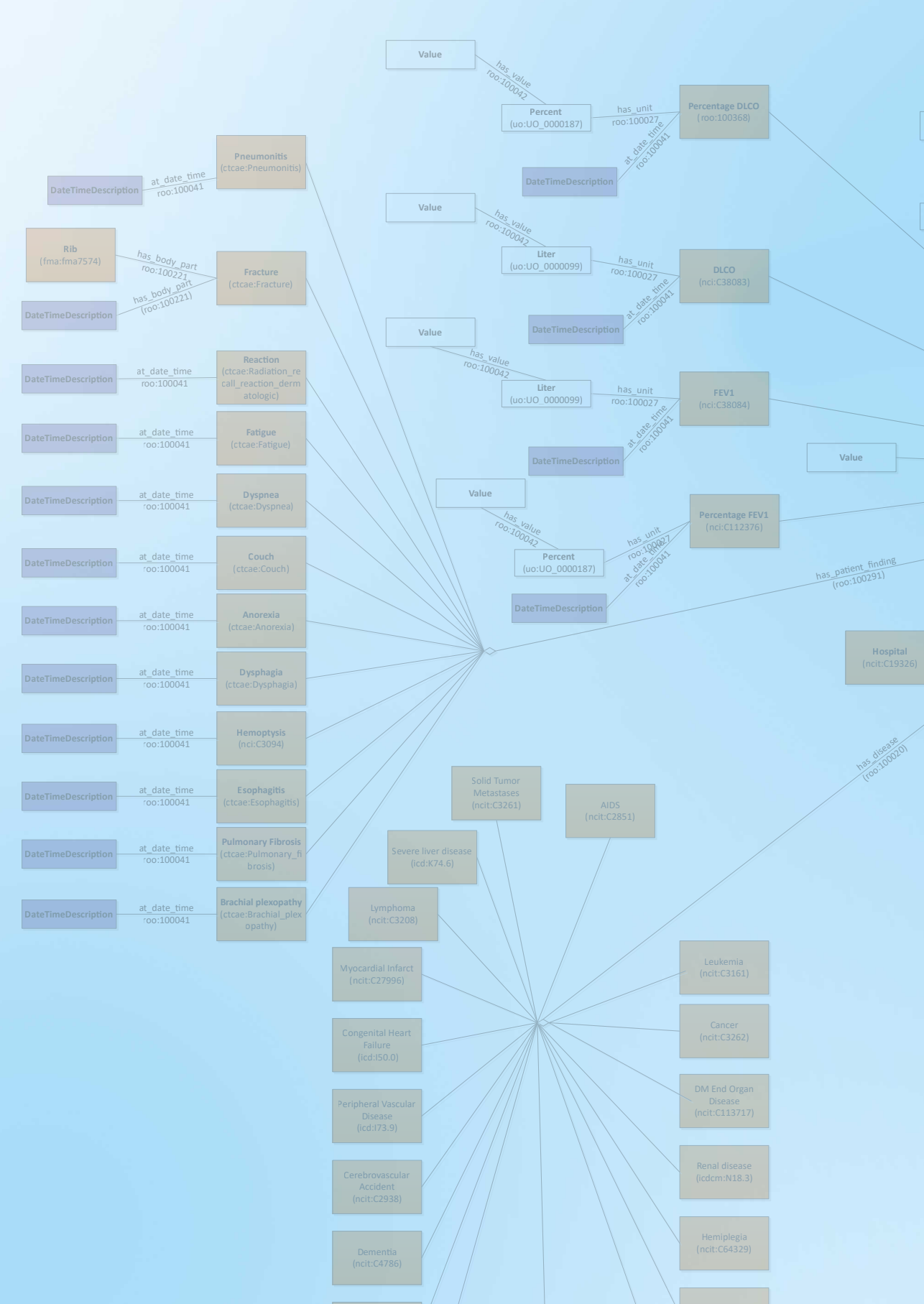
Sonia, dank voor alle hulp op de achtergrond. Niet alleen voor het organiseren van de thesis administratie, maar ook alle andere activiteiten en administratie in de afgelopen jaren (samen met de hele secretariële ondersteuning).

Research room 2nd floor at MAASTRO, thank you all as well! Although I haven't joined many of the social activities, it has been a nice and supportive group! Not only at Maastro, but also at the various conferences.

Lieve Esther, jij bent degene waar ik het meeste dank aan verschuldigd ben. Het is vaak voor gekomen dat ik – alweer – een aantal dagen van huis was. En als ik dan thuis was, waren er ook avonden dat ik heb gezegd "nog even een paar minuten dit afmaken" (dat meestal niet een paar minuten duurde). Ondanks dit, ben je altijd de steun en toeverlaat geweest, ook wanneer ik niet met beide benen op de grond stond. In dit proefschrift heb jij dan ook heel veel energie gestoken, en hoop dat je hier ook van kan genieten.

Verder wil ik ook mijn (schoon)familie en vrienden bedanken voor het meeleven. Pap en mam, Hans en Loes, Angela en Rob, Tim en Michelle, dank voor al het vertrouwen en steun! Siske en Roel, het is een mooie stap die jullie hebben gezet. En ondanks de afstand zullen we zeker contact blijven houden! De Outlaw Cycling Gang, dank voor het uitwaaien in het weekend. Al moeten we nog een hypothese beter onderzoeken (hoe meer alcohol de avond ervoor, hoe hoger de gemiddelde snelheid in de ochtend erna).

Zoals meerdere mensen hebben verteld, en ervaring leert, wordt dit hoofdstuk door bijna iedereen gelezen. Het is niet te doen om iedereen persoonlijk te bedanken, maar iedereen die interesse heeft getoond, of op een of andere manier heeft bijgedragen, heel erg bedankt!


Value
has_value
roo:100042
Percent
(uo:UO_0000187)
has_unit
roo:100027
Percentage DLCO
(roo:100368)
at_date_time
roo:100041
DateTimeDescription
Pneumonitis
(ctcae:Pneumonitis)
DateTimeDescription
at_date_time
roo:100041
Value
has_value
roo:100042
Liter
(uo:UO_0000099)
has_unit
roo:100027
DLCO
(nci:C38083)
Rib
(fma:fma7574)
has_body_part
roo:100221
Fracture
(ctcae:Fracture)
has_body_part
(roo:100221)
DateTimeDescription
at_date_time
roo:100041
DateTimeDescription
Value
has_value
roo:100042
Liter
(uo:UO_0000099)
has_unit
roo:100027
FEV1
(nci:C38084)
DateTimeDescription
at_date_time
roo:100041
DateTimeDescription
at_date_time
roo:100041
Reaction
(ctcae:Radiation_recall_reaction_dermatologic)
Value
Value
has_value
roo:100042
Percentage FEV1
(nci:C112376)
DateTimeDescription
at_date_time
roo:100041
Fatigue
(ctcae:Fatigue)
DateTimeDescription
at_date_time
roo:100041
Percent
(uo:UO_0000187)
has_unit
roo:100027
at_date_time
roo:100041
has_patient_finding
(roo:100291)
DateTimeDescription
at_date_time
roo:100041
Dyspnea
(ctcae:Dyspnea)
DateTimeDescription
at_date_time
roo:100041
Couch
(ctcae:Couch)
DateTimeDescription
at_date_time
roo:100041
Anorexia
(ctcae:Anorexia)
Hospital
(ncit:C19326)
DateTimeDescription
at_date_time
roo:100041
Dysphagia
(ctcae:Dysphagia)
DateTimeDescription
at_date_time
roo:100041
Hemoptysis
(nci:C3094)
has_disease
(roo:100020)
DateTimeDescription
at_date_time
roo:100041
Esophagitis
(ctcae:Esophagitis)
Solid Tumor
Metastases
(ncit:C3261)
AIDS
(ncit:C2851)
DateTimeDescription
at_date_time
roo:100041
Pulmonary Fibrosis
(ctcae:Pulmonary_fibrosis)
Severe liver disease
(icd:K74.6)
DateTimeDescription
at_date_time
roo:100041
Brachial plexopathy
(ctcae:Brachial_plexopathy)
Lymphoma
(ncit:C3208)
Myocardial Infarct
(ncit:C27996)
Leukemia
(ncit:C3161)
Congenital Heart
Failure
(icd:I50.0)
Cancer
(ncit:C3262)
Peripheral Vascular
Disease
(icd:I73.9)
DM End Organ
Disease
(ncit:C113717)
Cerebrovascular
Accident
(ncit:C2938)
Renal disease
(icdcm:N18.3)
Dementia
(ncit:C4786)
Hemiplegia
(ncit:C64329)