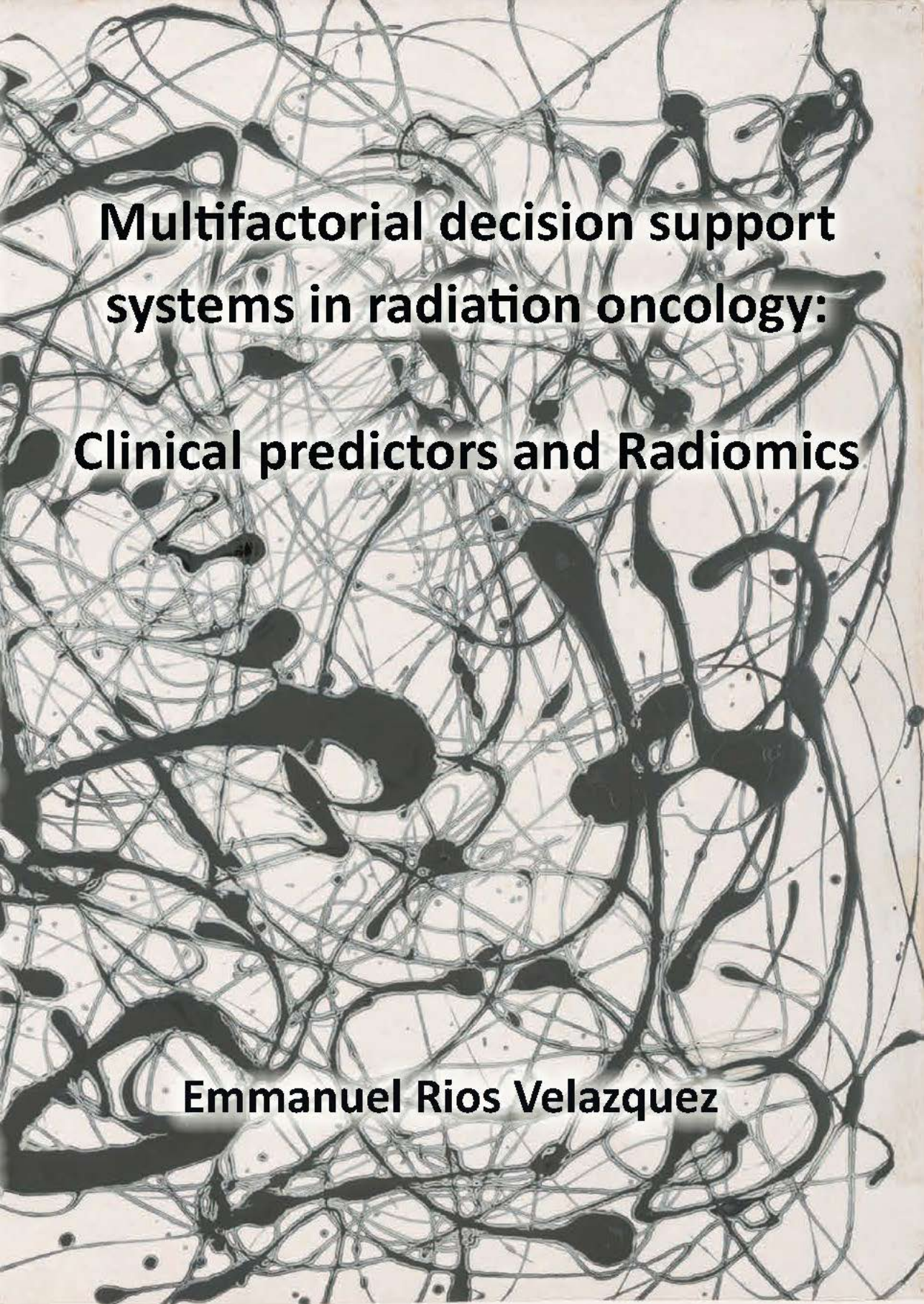
Multifactorial decision support systems in radiation oncology : clinical predictors and radiomics


Multifactorial decision support systems in radiation oncology:
Clinical predictors and Radiomics
Emmanuel Rios Velazquez

Cover illustration:

Yale University Art Gallery, New Haven, Connecticut 06520
1980.13.74
Artist Jackson Pollock
Number 14: Gray
1948
Enamel over gesso on paper
57 x 78.5 cm (22 7/16 x 30 7/8 in.), framed: 75.9 x 96.2 x 4.8 cm
(29 7/8 x 37 7/8 x 1 7/8 in.)

Yale University Art Gallery
Katharine Ordway Collection

Print: Datawyse | Universitaire Pers Maastricht
ISBN 978-94-6159-371-9

Multifactorial decision support systems in radiation oncology

Clinical predictors and Radiomics

DISSERTATION

to obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus Prof.dr. L.L.G. Soete,
in accordance with the decision of the Board of Deans,
to be defended in public on
Friday October 24th, 2014 at 14:00 hours

by

Emmanuel Rios Velazquez

UNIVERSITAIRE PERS MAASTRICHT

Promotor

Prof. Dr. Ph. Lambin

Co-promotores

Dr. Ir. H.J.W.L. Aerts (Harvard University)
Dr. Ir. A.L.A.J. Dekker
Dr. F.J.P. Hoebers

Assessment committee

Prof. Dr. F.C.S. Ramaekers, voorzitter
Prof. Dr. G.L. Beets
Prof. Dr. J.H.A.M. Kaanders (Radboud University)
Dr. M.E. Kooi

The work presented in this thesis is made possible by the financial support of: CTMM framework (AIRFORCE project, grant 030-103), euroCAT (IVA Interreg - www.eurocat.info) and the Dutch Cancer Society (KWF UM 2011-5020, KWF UM 2009-4454).

CONTENTS

INTRODUCTION

CHAPTER 1

General introduction and outline of the thesis

INTRODUCTION

A recent report from the World Health Organization shows that despite enormous scientific and technological advances in understanding and managing cancer, we cannot find our way out of the cancer problem yet. In 2012, there were 14.1 million new cancer cases, 8.2 million cancer deaths and 32.6 million people living with cancer [1]. It is expected that annual cancer cases will rise to 22 million within the next two decades.

Therefore, technological advances in fundamental research and clinical management of cancer as well as health care policies towards cancer prevention and early diagnosis are constantly being pushed forward. Over the past decade, many new diagnostic and treatment modalities have become available, generating an ever increasing amount of patient specific information. Novel disease markers are being identified in large numbers nowadays, including genomics, proteomics and non-invasive imaging [2]. These markers hold the promise of improving patient's prognostic information and moving personalized medicine a step closer (Figure 1).

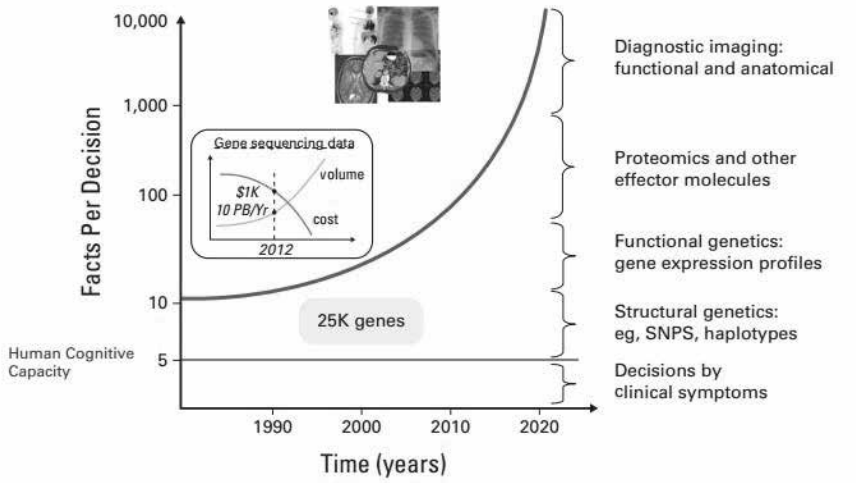


Figure 1

Technological advances in basic research lead to an explosion of factors available for medical decision making. The increase of factors required for medical decision making is shown relative to the human cognitive capacity [3].

However, an immediate consequence is that for each patient, the clinician needs to take into account a large number of factors while making a medical decision, for instance different treatment modalities, patient specific information coming from non-invasive imaging, blood markers, genomic essays, to name a few. This implies a cognitive overload, because the human cognitive capacity is limited to approximately five factors per decision [3]. It also poses the challenge of converting a myriad of raw data into digested medical

knowledge: the process of integrating diverse information (clinical, imaging, biological) to provide patient specific clinical predictions that can accurately estimate patients outcome and facilitate clinical decision making.

This highlights the growing need for validated clinical decision support systems (DSS's). The process of developing DSSs has been named rapid-learning [3, 4], and involves the use of historical data from routine clinical care, to derive knowledge that can be applied while making decisions concerning new patients. It is also coupled to the concept of personalized medicine, which involves the use of patient specific characteristics to tailor treatment, in contrast to applying the standard treatment that would be offered to a group of patients with similar characteristics.

It is expected that medical knowledge derived from the implementation of rapid-learning approaches will enable the application and validation of decision-support systems, which, in turn, will enable advances in shared decision making, in this case, in radiation oncology.

OBJECTIVE OF THE THESIS

The development of decision support systems to predict patients outcome in radiation oncology is needed to facilitate clinical shared decision making. Therefore, this thesis investigates the development of prognostic models for lung and head and neck cancer patients to identify patients at different risk levels before treatment. There were two specific research goals:

I) the integration of patient clinical and treatment characteristics in lung and head and neck cancer to derive prognostic models of outcome.

II) the use of advanced quantitative imaging features, extracted from conventional medical imaging, many of which are not currently used, to improve patients prognostic information in lung and head and neck cancer.

OUTLINE OF THE THESIS

The work of this thesis is presented in two parts, the first one on prognostic models using clinical patient characteristics and the second part on the use of Radiomic approaches for outcome prediction. To begin with, in this **Chapter 1**, a general introduction of the work presented in this thesis is given. We also introduce the concept of decision support systems in radiation oncology.

As a general introduction to Clinical Decision Support Systems in radiation oncology, **Chapter 2** provides an overview of the factors that have been associated with outcomes in radiation oncology, and discusses the methodology needed for the development of prediction models through the multiple stages involved.

Part 1: Clinical predictors

This part of the thesis describes the developed predictive models using patient clinical and treatment characteristics. For this purpose, relevant questions are investigated as to what clinical factors are associated with patient's outcome? Can prognostic models be validated in external datasets? How do prognostic models compare with traditional decision making aids such as the TNM staging system?

Chapter 3 presents a study in which we evaluated the most important patient, tumor and treatment factors associated with residual metabolic activity after treatment. Metabolic response assessment has been associated with survival and treatment failure. This study was performed in a MAASTRO dataset of 101 NSCLC patients.

Chapter 4 describes the development of a prognostic nomogram for the prediction of overall survival and local control in laryngeal carcinoma patients. It also shows the validation of the same prognostic tool in four external datasets.

Chapter 5 presents a study to develop a prognostic nomogram for oropharyngeal carcinoma patients for prediction of overall survival and progression-free survival. It also shows a comparison of the developed nomogram with TNM and an important prognostic factor in these patients: HPV status.

Part 2: Radiomics: Extracting more information from medical images using advanced feature analysis

This second part of the thesis deals with the potential of extracting advanced quantitative features from medical images for outcome prediction. In this section we ask whether there is more information in medical imaging than what is currently used. Are descriptors of tumor shape or texture useful for outcome prediction?
In specific, we proposed a methodology for high-throughput extraction of quantitative imaging parameters, evaluated methods for robust target definition and assessed the prognostic value of these imaging parameters in lung and head and neck cancer cohorts.

Chapter 6, puts forward the concept of Radiomics: the high-throughput extraction of large amounts of image features from radiographic images. This review addresses the hypothesis, methodological aspects, and challenges underlying the Radiomics approach.

Towards the extraction of reproducible quantitative imaging features, **Chapter 7**, evaluates the relevance of a semiautomatic CT-based segmentation method, by comparing it to manual delineations made by radiation oncologists and to pathological tumor measurements considered as “gold standard” in NSCLC patients.

Chapter 8, evaluates an open source, freely available method for lung tumors segmentation, and evaluates its usefulness by comparing it again, against the gold standard pathological measurements and radiation oncologists delineations, and a step further, examining whether its use reduces variability during tumor segmentation.

Following these results, **Chapter 9** evaluates whether quantitative imaging features extracted from semi-automatically segmented tumors have lower variability and are more robust compared to features extracted from manual tumor delineations. This study analyzes the robustness of imaging features derived from semi-automatically and manually segmented primary NSCLC tumors in twenty patients.

Chapter 10, presents an analysis of 440 quantitative imaging features quantifying phenotypic differences based on tumor appearance, i.e., shape, intensity and texture, in CT images of more than 1000 patients with lung or head and neck cancer.

General discussion and future perspectives

Chapter 11 provides a general discussion of the results presented in this thesis and its future outlook and perspectives.

REFERENCES

1. World Health Organization. Globocan 2012, IARC.
2. Fraass, B. A. & Moran, J. M. Quality, technology and outcomes: evolution and evaluation of new treatments and/or new technology. *Semin. Radiat. Oncol.* 22, 3–10 (2012).
3. Abernethy, A. P. *et al.* Rapid-learning system for cancer care. *J. Clin. Oncol.* 28, 4268–4274 (2010).
4. Lambin P, Roelofs E, Reymen B, et al. 'Rapid Learning health care in oncology' - An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol* 2013; **109**(1): 159-64.

CHAPTER 2

Predicting outcomes in radiation oncology – multifactorial decision support systems

Published in: Nature Reviews Clinical Oncology 10:1 (2013) 27–40

Predicting outcomes in radiation oncology —multifactorial decision support systems

Philippe Lambin, Ruud G. P. M. van Stiphout, Maud H. W. Starmans, Emmanuel Rios Velazquez, Georgi Nalbantov, Hugo J. W. L. Aerts, Erik Roelofs, Wouter van Elmpt, Paul C. Boutros, Pierluigi Granone,Vincenzo Valentini, Adrian C. Begg, Dirk De Ruysscher and Andre Dekker

ABSTRACT

With the emergence of individualized medicine and the increasing amount and complexity of available medical data, a growing need exists for the development of clinical decision-support systems based on prediction models of treatment outcome. In radiation oncology, these models combine both predictive and prognostic data factors from clinical, imaging, molecular and other sources to achieve the highest accuracy to predict tumour response and follow-up event rates. In this Review, we provide an overview of the factors that are correlated with outcome including survival, recurrence patterns and toxicity—in radiation oncology and discuss the methodology behind the development of prediction models, which is a multistage process. Even after initial development and clinical introduction, a truly useful predictive model will be continuously re-evaluated on different patient datasets from different regions to ensure its population-specific strength. In the future, validated decision-support systems will be fully integrated in the clinic, with data and knowledge being shared in a standardized, instant and global manner.

INTRODUCTION

Over the past decade we have witnessed advances in cancer care, with many new diagnostic methods and treatment modalities[1] becoming available, including advances in radiation oncology.[2] The abundance of new options and individualized medicine has, however, created new challenges. For example, achieving level I evidence is increasingly difficult given the numerous disease and patient parameters that have been discovered, resulting in an ever-diminishing number of 'homogeneous' patients.[3] This reality contrasts to a certain extent with classic evidence-based medicine, whereby randomized trials are designed for large populations of patients. Thus, new strategies are needed to find evidence for subpopulations on the basis of patient and disease characteristics.[4]

For each patient, the clinician needs to consider state-of-the-art imaging, blood tests, new drugs, improved modalities for radiotherapy planning and, in the near future, genomic data. Medical decisions must also consider quality of life, patient preferences and, in many health-care systems, cost efficiency. This combination of factors renders clinical decision making a dauntingly complex, and perhaps inhuman, task because human cognitive capacity is limited to approximately five factors per decision.[3]

Furthermore, dramatic genetic[5], transcriptomic[6], histological[7] and micro-environmental[8] heterogeneity exists within individual tumors, and even greater heterogeneity between patients.[9] Despite these complexities, individualized cancer treatment is inevitable. Indeed, intratumoural and intertumoural variability might be leveraged advantageously to maximize the therapeutic index by increasing the effects of radiotherapy on the tumour and decreasing those effects on normal tissues.[10-12]

The central challenge, however, is how to integrate diverse, multimodal information (clinical, imaging and molecular data) in a quantitative manner to provide specific clinical predictions that accurately and robustly estimate patient outcomes as a function of the possible decisions. Currently, many prediction models are being published that consider factors related to disease and treatment, but without standardized assessments of their robustness, reproducibility or clinical utility.[13] Consequently, these prediction models might not be suitable for clinical decision-support systems for routine care.

In this Review, we highlight prognostic and predictive models in radiation oncology, with a focus on the methodological aspects of prediction model development. Some characteristic prognostic and predictive factors and their challenges are discussed in relation to clinical, treatment, imaging and molecular factors. We also enumerate the steps that will be required to present these models to clinical professionals and to integrate them into clinical decision-support systems (CDSSs).

METHODOLOGICAL ASPECTS

Factors for prediction

The overall aim of developing a prediction model for a CDSS is to find a combination of factors that accurately anticipate an individual patient's outcome.[14] These factors include, but are not limited to, patient demographics as well the results of imaging, pathology, proteomic and genomic testing, the presence of key biomarkers and, crucially, the treatment undertaken. 'Outcome' can be defined as tumour response to radiotherapy, toxicity evolution during follow up, rates of local recurrence, evolution to metastatic disease, survival or a combination of these end points. Although predictive factors (that is, factors that influence the response to a specific treatment) are necessary for decision support, prognostic factors (that is, factors that influence response in the absence of treatment)[15] are equally important in revealing the complex relationship with outcome. Herein, we refer to both of these terms generically as 'features' because, for a predictive model, correlation with outcome must be demonstrable.

Model development stages

The procedure for finding a combination of features correlated with outcome is analogous to the development of biomarker assays.[16] In that framework, we can distinguish qualification and validation. Qualification demonstrates that the data are indicative or predictive of an end point, whereas validation is a formalized process used to demonstrate that a combination of features is both reliable and suitable for the intended purpose. That is, we need to identify features, test whether they are predictive in independent datasets and then determine whether treatment decisions made using these features improve outcome. The complete cycle of model development entails several stages (Figure 1).

In the hypothesis-generation stage, one must consider the end point to predict, the timing of the treatment decision and the available data at these time points. In the data-selection step, a review of potential features is first conducted, ideally by an expert panel. A practical inventory of the available data and sample-size calculations are recommended, especially for the validation phase.[17,18] Data from both clinical trials (high quality, low quantity, controlled, biased selection) and clinical practice (low quality, high quantity, unbiased selection) are useful, but selection biases must be identified in both cases and the inclusion criteria should be equivalent. For all features, including the characteristics of the treatment decision, data heterogeneity is a requirement to identify predictive features and to have the freedom to tailor treatment.

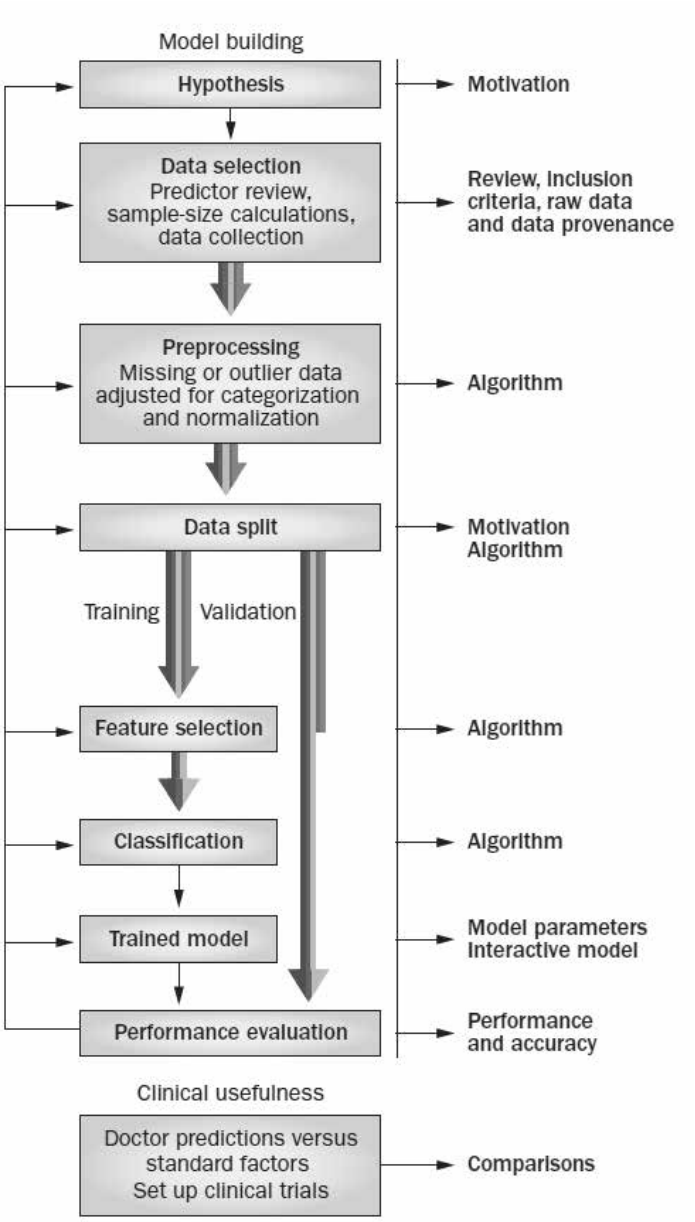


Figure 1

Schematic overview of methodological processes in clinical decision-support system development, describing model development, assessment of clinical usefulness and what ideally to publish. The coloured, parallel lines represent heterogeneous data, which have been split early for independent validation (but without internal cross-validation).

Next, performance measures for models are determined, and these measures include the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity and c-index of censored data.[19] AUC, which has values between 0 and 1 (with 1 denoting the best model and 0.5 randomness), is the most commonly used performance measure. However, for time-to-event models, the c-index and hazard ratio are more appropriate because both can handle censored data.

The preprocessing stage deals with missing data (imputation strategies; that is, replacing missing values by calculated estimates),[20] identifying incorrectly measured or entered data [21], as well as discretizing (if applicable) and normalizing data to avoid sensitivity for different orders of data scales.[22] If an external, independent dataset is not available for validation, the available data must be split (in a separate stage) into a model-training dataset and a validation set, the latter of which is used subsequently in the validation step. In the feature-selection stage, the ratio of the number of evaluated features to number of outcome events must be kept as low as possible to avoid overfitting. When a model is overfitted, it is specifically and exclusively trained for the training data (including its data noise) and, as a result, performs poorly on new data. Data-driven preselection of features is, therefore, recommended.[23] Univariate analyses are commonly used to prioritize the features—that is, testing each feature individually and ranking them on their strength of correlation with outcome.

Predicting outcomes

In the next stage, the input data are fed into a model that can classify all possible patient outcomes. Traditional statistical[24] and machine-learning models[25] can be considered. For two or more classes (for example, response versus no response), one might consider logistic regression, support vector machines, decision trees, Bayesian networks or Naive Bayes algorithms.[26,27] For time-to-event outcomes, whether censored or not, Cox proportional hazards models28 or the Fine and Gray model[28] of competing risks are most common. The choice of model depends on the type of outcome (for example, logistic regression for two or more outcomes, or Cox regression for survival-type data) and the type of input data (for example, Bayesian networks require categorized data, whereas support-vector machines can deal with continuous data). In general, several models with similar properties can be tested to find the optimal model for the available data. A simple model is, however, preferred because it is expected to be robust to a wider range of data than a more complex model.

Performance on the training dataset is upwards-biased because the features were selected. Thus, external validation data must be used, which can be derived from a separate institute or independent trial. When data are limited, internal validation can be considered using random split, temporal split or k-fold cross-validation techniques.[29] The developed model should have a benefit over standard decision making, and must be assessed prospectively in the clinic in the penultimate stage of development. Models must

be compared against predictions by clinicians[30,31] and to standard prognostic and predictive factors.[32] Critically, to demonstrate the improvement of patient outcome, quality of life and/or reduced toxicity,[33] clinical trials must be conducted whereby the random assignment of patients is based on the prediction model output. Fulfilling this requirement will generate the final evidence that the model is improving health care by comparing, in a controlled way, the tailored treatments with standard treatments in the clinic.

Finally, the prediction models and data can be published, enabling the wider oncological community to evaluate them. Full transparency on the data and methodology is the key towards global implementation of the model into CDSSs. This suggestion is similar to clinical 'omics' publications for which the raw data, the code used to derive the results from the raw data, evidence for data provenance (the process that led to a piece of data) and a written description of nonscriptable analysis steps are routinely made available.[34]
In practice, this cycle of development usually begins by identifying clinical parameters, because these are widely and instantly available in patient information systems and clinical trials. These clinical variables also form the basis for extending prediction models with imaging or molecular data.

CLINICAL FEATURES

Decision making in radiotherapy is mainly based on clinical features, such as the patient performance status, organ function and grade and extent of the tumour (for example, as defined by the TNM system). In almost all studies, such features have been found to be prognostic for survival and development of toxicity.[35-37] Consequently, these features should be evaluated in building robust and clinically acceptable radiotherapy prognostic and predictive models. Moreover, measurement of some clinical variables, such as performance status, can be captured with minimal effort. Even the simplest questionnaire, however, should be validated as is the case for laboratory measurements of organ function or parameters measured from blood.[38,39]

Furthermore, a standardized protocol should be available to ensure that comparisons are possible between centres and questionnaires over time.[40] Moreover, why specific features were chosen for measurement should be clearly explained. For example, if haemoglobin measurements were only taken in patients with fatigue, the resulting bias would demand caution when including and interpreting the measurements. Only when clinical parameters are recorded prospectively with the same scrutiny as laboratory measurements will observational studies become as reliable as randomized trials.[41,42]

Toxicity measurements and scoring should also build on validated scoring systems, such as the Common Terminology Criteria for Adverse Events (CTCAE), which can be scored by the physician or patient.[43,44] Indeed, a meta-analysis showed that high-quality toxicity assessments from observational trials are similar to those of randomized trials.[45,46]

However, a prospective protocol must clarify which scoring system was used and how changes in toxicity score were dealt with over time with respect to treatment.

Finally, to ensure a standardized interpretation, the reporting of clinical and toxicity data and their analyses should be performed in line with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) statement for observational studies and genetic-association studies, which is represented as checklists of items that should be addressed in reports to facilitate the critical appraisal and interpretation of these type of studies.[47,48]

TREATMENT FEATURES

Currently, image-guided radiotherapy (IGRT) is a highly accurate cancer treatment modality in delivering its agent (radiation) to the tumour.[49] Furthermore, very accurate knowledge of the effects of radiation on normal tissue has been obtained[50]. With modern radiotherapy techniques—such as intensity-modulated radiotherapy, volumetric arc therapy or particle-beam therapy—the treatment dose can be sculpted around the target volume with dosimetric accuracy of a few percentage points. IGRT ensures millimetre precision to spare the organs at risk as much as possible.[51]

For prediction modelling, recording features that are derived from planned spatial and temporal distribution of the radiotherapy dose is crucial. Additionally, features must be recorded that describe the efforts undertaken during treatment to ensure that the dose is delivered as planned (that is, *in vivo* dosimetry*)*; a delicate balance exists between tumour control and treatment-related toxicity.[52] Additional therapies, such as (concurrent) chemotherapy, targeted agents and surgery, and their features must also be recorded because these have various effects on outcome.[32,53] An example is the difference between concurrent versus sequential chemoradiation, which has a major influence on the occurrence of acute oesophagitis that induces dysphagia.[54] With respect to the spatial dimension of radiotherapy, how to combine information about the spatially variable dose distribution for every subvolume of the target tumour (or organ) with the global effect to the tumour or adjacent normal tissue remains indeterminate. Dose-response relationships for tumour tissues are often reported in terms of mean (biologically equivalent) dose, although voxel-based measures have also been reported.[55]

Mean doses or doses to a prescription point inside the tumour are easily determined and reported and can suffice for many applications. However, spatial characteristics might be more relevant in personalized approaches to ensure radioresistant areas of the tumour receive higher doses.[55] For normal tissue toxicity, dose features—including the mean and maximum dosage, as well as the volume of the normal tissue receiving a certain dose—are important. For example, V20 <35% is a common threshold to prevent lung toxicity.[56]

Clinical dose–volume histogram analysis for pneumonitis after the 3D treatment for non-small-cell lung cancer was first described in 1991.[57] In 2010, a series of detailed re-

views of all frequently irradiated organs (the QUANTEC project) was described,[50] showing that, as for the tumour, care must be taken when assessing dose at the organ level. For example, in some organs, the volume receiving a certain dose is important (such as the oesophagus or lung) because of their proximity to other vital structures, whereas the maximum dose to a small region of other organs might be most important (such as for the spinal cord) because preserving its post-treatment function is crucial. Predicting complications to normal tissue is an active research area in ongoing, large, prospective multicentre projects, including ALLEGRO[58] and others.[59-61]

Although important, in general, one must be careful about relying completely on planned-radiotherapy dose-based predictions because patients display wide variability in toxicity development. The reasons for this variability include many known clinical and molecular-based features as well as the quality of the treatment execution. The focus on the planned radiotherapy dose distribution as the prime determinant of outcome is perhaps the most common pitfall in prediction models because deviations from the original plan during the time of treatment frequently occur.[62] The accuracy of prediction models is expected to increase when measured dose is used, as this measure reflects the effect of radiotherapy most accurately. Figure 2 shows an example of these variations in a patient with prostate cancer. Dose reconstructions (2D and 3D), Gamma-Index calculations and dose–volume histograms during treatment can help in identifying increasingly accurate dose-related features,[63,64] such as radiation pneumonitis[65] and oesophagitis.[66]

The temporal aspect of fractionated radiotherapy is also an active area of research. The fact that higher radiation doses are required to control a tumour when treatment is prolonged is well-known, and increasing evidence suggests that accelerated regimens giving the same physical dose can improve outcome.[67,68] A multicentre analysis of patients with head-and-neck cancer treated with radiotherapy alone showed that the potential doubling time of the tumour before treatment was not a predictor for local control.[69] Alongside the classic explanation of accelerated repopulation,[70] changes in cell loss, hypoxia and selection of radioresistant stem cells have each been suggested as underlying causes of this observation, the possible implications of which include shorter overall treatment times with higher doses per fraction and the avoidance of breaks during treatment.[71,72] Overall, treatment time is an accessible feature that is correlated with local failure in several tumour sites.[73,74]

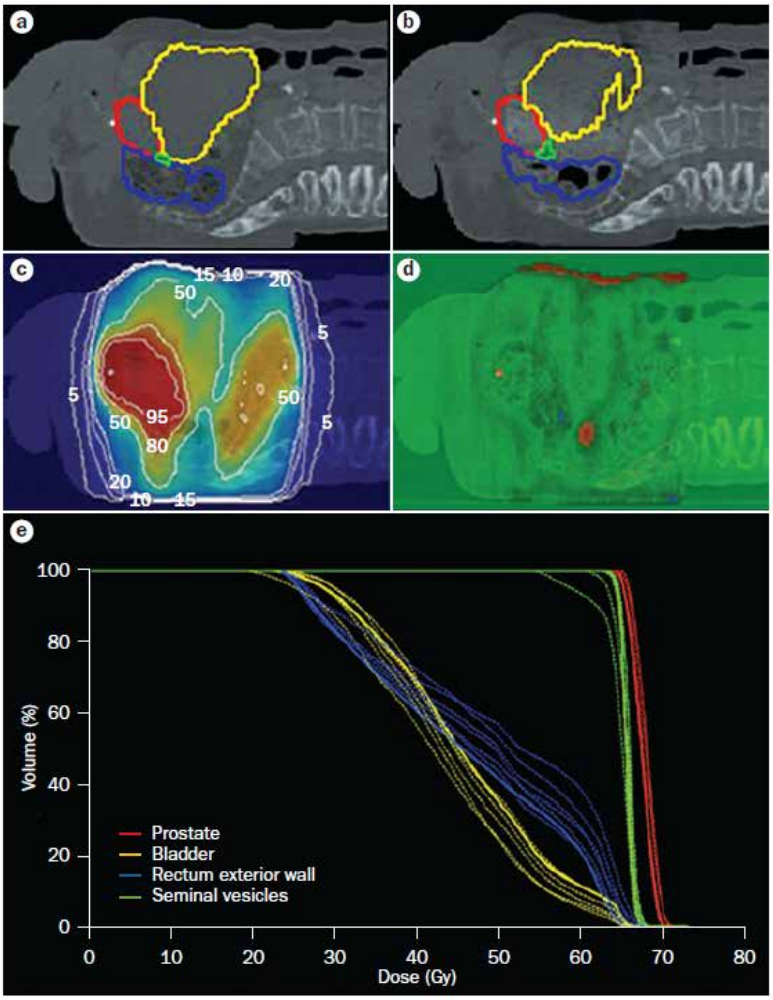


Figure 2

The importance of considering measured dose for outcome prediction for a patient with prostate cancer. ***a*** *| Original planning CT scan that includes contours of the prostate (red), bladder (yellow), exterior wall of the rectum (blue) and seminal vesicles (green).* ***b*** *| Contoured CT scan after 16 fractions of radiotherapy.* ***c*** *| Reconstructed 3D dose after 16 fractions of radiotherapy.* ***d*** *| Calculated dose differences (expressed as a 3D Gamma Index) after 16 fractions of radiotherapy.* ***e*** *| Dose–volume histograms at fractions 1, 6, 11, 16, 21 and 26 (dashed lines) as well as pretreatment histograms (solid lines). Clear deviations are visible from the planned dose–volume histogram for the rectum and bladder.*

Ideally, the spatial and temporal dimensions of radiotherapy would be exploited by showing a fractional dose distribution in a tumour radioresistance (and normal-tissue radiosensitivity) map that is continuously updated during treatment. However, such an image of radioresistance does not yet exist. If it did, CDSSs would guide the planning and modifica-

tion of the spatial and temporal distribution of radiation in such a way as to maintain or improve the balance between tumour control and the probability of normal tissue complications continuously during treatment, instead of the current approach that delivers radiation as planned with an identical dose to the tumour as a whole.

IMAGING FEATURES

Medical imaging has a fundamental role in radiation oncology, particularly for treatment planning and response monitoring.[75,76] Technological advances in noninvasive imaging—including improved temporal and spatial resolution, faster scanners and protocol standardization—have enabled the field to move towards the identification of quantitative noninvasive imaging biomarkers.[77–79]

Metrics based on tumour size and volume are the most commonly used image-based predictors of tumour response to therapy and survival,[80–87] and rely on CT and MRI technology for 3D measurement.[88–90] Although used in clinical practice, tumour size and volume measurements are subject to inter-observer variability that can be attributed to differences in tumour delineations.[85–87, 91,92] Moreover, the optimal measurement technique and definitions of appropriate response criteria, in terms of changes in tumour size, are unclear.[93] Additionally, tumour motion and image artefacts are additional sources of variability.[94,95] To overcome these issues, automated tumour delineation methods have been introduced[96–99] on the basis of, for example, the selection of ranges of Hounsfield units (which represent the linear attenuation coefficient of the X-ray beam by the tissue) on CT that define a certain tissue type, or calculation of the gradient of an image (mathematical filter) to reveal the borders between tissue types. Extensive evaluation, however, is needed before these methods can be used routinely in the clinic.[100–102]

A commonly used probe for the metabolic uptake of the tumour is 18F-fluorodeoxyglucose (FDG) for PET imaging.[103,104] The pre-treatment maximum standardized uptake value (SUV, which is the normalized FDG uptake for an injected dose according to the patient's body weight) is strongly associated with overall survival and tumour recurrence in a range of tumour sites, including the lung, head and neck, rectum, oesophagus and cervix.[105–111] Furthermore, several studies have shown that changes in SUV during and after treatment are early predictors of tumour recurrence.[112–115] FDG–PET measurements, however, are dependent on a number of factors, including injected dose, baseline glucose concentration, FDG clearance, image reconstruction methods used and partial-volume effects.[116,117] Standardization of these factors across institutions is, therefore, fundamental to enable comparisons and validation of data from FDG–PET imaging.[118,119]

Multiple studies have shown that diffusion-weighted MRI parameters, such as the apparent diffusion coefficient (ADC), which is a measure of water mobility in tissues, can accurately predict response and survival in multiple tumour sites.[120–124] However, lack of reproducibility of ADC measurements—due to lack of standardization of instruments be-

tween vendors and of internationally accepted calibration protocols—remains a bottleneck in these types of studies.[125] Evaluations of different time points in dynamic contrast-enhanced MRI have also been used to describe tumour perfusion.[90,126–128] Indeed, hypothesis-driven preclinical[129] and xenograft studies[130,131] support these clinical studies. For example, assessment of the correlation of features from imaging (such as lactate level and the extent of reoxygenation) with tumour control is possible.[130,131]

Increasingly advanced image-based features are currently being investigated. For example, routine clinical imaging can capture both tumour heterogeneity and post-treatment changes, which can be analyzed to identify functional biomarkers (Figure 3).

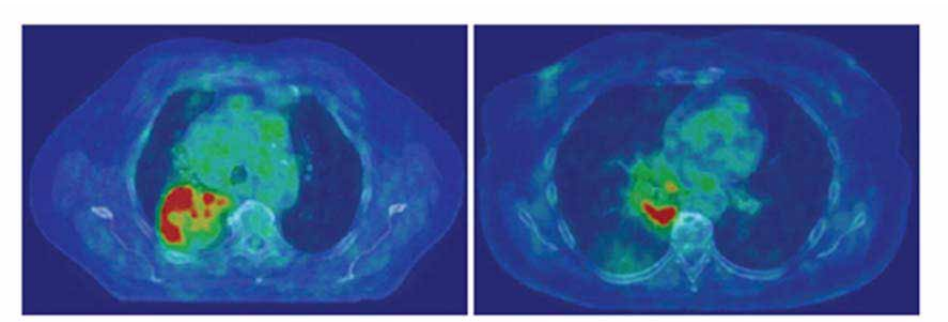


Figure 3

Axial FDG–PET and CT images of two different patients with NSCLC. Tumour imaging biomarkers describing, for example, textural heterogeneity, FDG uptake and tumour size can be assessed noninvasively before, during and after radiotherapy and associated with treatment outcome. Abbreviations: FDG, ^{18}F-fluorodeoxyglucose; NSCLC, non-small-cell lung cancer.

Changes in Hounsfield units in contrast-enhanced CT are directly proportional to the quantity of contrast agent present in the tissue and have been used as a surrogate for tumour perfusion.[132,133] Indeed, reductions of Hounsfield units following treatment have been used to evaluate treatment response in rectal, hepatic and pulmonary cancers.[134,135] Standardizing the extraction and quantification of a large number of traits derived from diagnostic imaging are now being considered in new imaging marker approaches.[79] Through advanced image-analysis methods, we can quantify descriptors of tumour heterogeneity (such as variance or entropy of the voxel values) and the relationship of the tumour with adjacent tissues.[136–138] These analytical methods enable high-throughput evaluation of imaging parameters that can be correlated with treatment outcome and, potentially, with biological data. Indeed, qualitative imaging parameters on CT and MRI scans have been used to predict mRNA abundance variation in hepatocellular carcinomas and brain tumours.[139–141] Furthermore, a combination of anatomical, functional and metabolic imaging techniques might be used to capture pathophysiological and morphological tumour characteristics in a noninvasive manner, including apparent intratumoural heterogeneity.[142]

MOLECULAR FEATURES

Biological markers are also valuable clinical decision-support features; these include prognostic and predictive factors for outcomes, such as tumour response and normal-tissue tolerance. Despite these strengths, trials of molecular biomarkers are prone to experimental variability; for this reason standardizing assay criteria, trial design and analysis are imperative if multiple molecular markers are to be used in predictive modelling.[16]

Tumour response

Next to tumour size, tumour control after radiotherapy is largely determined by three criteria: intrinsic radiosensitivity, cell proliferation and the extent of hypoxia.[143] In addition, large tumours intuitively require higher doses of radiation than small tumours because there are simply more cells to kill—this requirement is true even if intrinsic radiosensitivity, hypoxia and repopulation rates are equal. Several approaches have been developed to measure these additional three parameters to predict tumour response to radiotherapy.

Intrinsic radiosensitivity

Malignant tumours display wide variation in intrinsic radiosensitivity, even between tumours of similar origin and histological type.[144] Attempts to assess the radiosensitivity of human tumours have relied on determining the *ex vivo* tumour survival fraction.[145] Those studies and others have shown that tumour cell radiosensitivity is a significant prognostic factor for radiotherapy outcome in both cervical[146] and head-and-neck[147] carcinomas. However, these colony assays suffer from technical disadvantages that include a low success rate (<70%) for human tumours and the time needed to produce data, which can be up to several weeks.

Other studies have included assessments of chromosome damage, DNA damage, glutathione levels and apoptosis.[148] Indeed, some clinical studies using such assays have shown correlations with radiotherapy outcome, whereas others have not.[149] However, these cell-based functional assays only have limited clinical utility as predictive assays, despite being useful in confirming a mechanism that underlies differences in the response of tumours to radiotherapy. For example, some studies have provided encouraging data showing that immunohistochemical staining for γ-histone H2AX, a marker of DNA damage, might be a useful way to assess intrinsic radiosensitivity very early after the start of treatment.[150, 151] Double-stranded breaks are generated when cells are exposed to ionizing radiation or DNA-damaging chemotherapeutic agents, which rapidly results in the phosphorylation of γ-histone H2AX. γ-Histone H2AX is the most sensitive marker that can be used to examine the DNA damage and its subsequent repair, and it can be detected by immunoblotting and immunostaining using microscopic or flow cytometric detection. Clinically, two biopsies (one before and one after treatment) are needed to assess the γ-histone H2AX status, which is not always easy to implement in practice.

Hypoxia

Tumour hypoxia is the key factor involved in determining resistance to treatment and malignant progression; it is a negative prognostic factor after treatment with radiotherapy, chemotherapy and surgery.[152, 153] Indeed, some data show that hypoxia promotes both angiogenesis and metastasis and, therefore, has a key role in tumour progression.[154] Although a good correlation has been demonstrated between pimonidazole (a chemical probe of hypoxia) staining and outcome after radiotherapy in head-and-neck cancer,[155] the same relationship has not been found in cervical cancer.[156] In light of these contrasting results, one of the hypotheses put forward to explain this is that hypoxia tolerance is more important than hypoxia itself.[157]

The use of fluorinated derivatives of such chemical probes also enables their detection by noninvasive PET.[158–160] Although this approach requires administration of a drug, it does benefit from sampling the whole tumour and not just a small part of it. Another possible surrogate marker of hypoxia is tumour vasculature; the prognostic significance of tumour vascularity has been measured as both intercapillary distance (thought to reflect tumour oxygenation) and microvessel density (the 'hotspot' method that provides a histological assessment of tumour angiogenesis). Some studies have found positive correlations with outcome—mainly using microvessel density in cervical cancer—whereas others have shown negative correlations.[161] Some concerns have been raised about the extent to which biopsies taken randomly truly represent the usually large, heterogeneous tumours.

Proliferation

If the overall radiotherapy treatment time is prolonged, for example, for technical reasons (breakdown of a linear particle accelerator) or because of poor tolerance by the patient to the treatment, higher doses of radiation are required for tumour control—clearly indicating that the influence of tumour proliferation is important.[162] Although proliferation during fractionated radiotherapy is clearly an important factor in determining outcome, reliable measurement methods are not yet available. To understand why radiation leads to an accelerated repopulation response in some tumours and not in others, a greater understanding of the response at both the cellular and molecular level is required.

Normal-tissue tolerance

Inherent differences in cellular radiosensitivity among patients dominate normal-tissue reactions more than other contributing factors.[164] That is, the radiation doses given to most patients might in actuality be too low for an optimal cure because 5% of patients are very sensitive; these 5% of patients are so sensitive that they skew what is 'optimal' radiotherapy to the lower end of the spectrum, to the detriment of the majority of patients who are not as sensitive. Future CDSSs should be able to distinguish such overly sensitive

patients and classify them separately so they receive different treatments to the less-sensitive patients.

Several small[165] and large[166] *in vitro* studies found a correlation between radiosensitivity and severity of late effects, namely radiation-induced fibrosis of the breast, but these findings were not consistent because no standardized quality assurance exists for radiotherapy *in vivo*.[167, 168] Similar discrepancies were later found using rapid assays that measure chromosomal damage,[169] DNA damage[170] and clonogenic cell survival.[171] For example, the lymphocyte apoptosis assay has been used in a prospective trial as a stratification factor to assess late toxicity using letrozole as radiosensitizer in patients with breast cancer.[172] Cytokines such as TGF-β, which influences fibroblast proliferation and differentiation, are known to have a central role in fibrosis and senescence.[173,174] Currently, the relationships between the lymphocyte predictive assay, TGF-β and late complications are purely correlative and a clear molecular explanation is lacking. Genome-wide association studies (GWAS) and the analysis of single nucleotide polymorphisms (SNPs) in candidate genes have also shown promise in identifying normal-tissue tolerance,[175,176] although these do not often validate results from independent studies.[177] In general, the problem with all these studies has been the wide experimental variability rather than interindividual differences in radiosensitivity. Normal-tissue tolerance is the dose-limiting factor for the administration of radiotherapy, therefore, any CDSS should be based on predictors of tumour control and the probability of complications.

REPRESENTATION OF PREDICTIONS

Although the decisions made in the process of developing predictive models will determine the characteristics of a multivariate model (for example, which features are selected and the overall prediction accuracy), the success of the model depends on other factors, such as its availability and interactivity, which increases the acceptability. Even models based on large patient populations, with proper external validation, can fail to be accepted within the health-care community if the model and its output are not easily interpretable, if there is a lack of opportunity to apply the model or if the clinical usefulness is not proven or reported.[178]

Although some models, such as decision trees, implicitly have a visual representation that is somewhat interpretable, most models do not. One highly interpretable representation of a set of features is the nomogram.[179] The nomogram was originally used in the early 20th century to make approximate graphical computations of mathematical equations. In medicine, nomograms have experienced a revival, reflected by the increasing number of studies reporting them.[180–184] Figure 4 shows an example of a published clinical nomogram of local control in larynx cancer in which values for the selected features directly relate to a prediction score. The sum of these scores corresponds to a probability of local control within 2 or 5 years.[181]

Another idea for increasing acceptability of computer-assisted personalized medicine is to make prediction models available on the internet. If interactive, peer-reviewed models are provided with sufficient background information, clinicians can test them using their own patient data. Such a system would provide retrospective validation of the multiple features by the wider community, as well as provide an indication on the clinical usefulness of the methodology. The best-known website with interactive clinical prediction tools is Adjuvant! Online.[185] This website provides decision support for adjuvant therapy (for example, chemotherapy and hormone therapy) after surgery for patients with early-stage cancer. Many researchers have evaluated the models available on this prediction website, thereby refining them with additional predictors and updated external validations.[186,187] A prediction website that focuses on decision support for radiotherapy was recently established.[188] The aim of this website is to let users work with and validate the interactive models developed for patients with cancer treated with radiotherapy, which contributes to CDSS development in general by demonstrating the potential of these predictions and raising the awareness of their existence and limitations.

FUTURE PROSPECTS

The major focus of this Review, thus far, has been model development, validation and presentation (including the features from different domains that might be considered as predictive and prognostic). Although an accurate outcome prediction model forms the basis of a CDSS, additional considerations must be made before a new CDSS can be used in daily radiation oncology practice.

First, any decision a patient or physician makes is based on a balance between its benefits (survival, local control and quality of life) and harms (toxic effects, complications, quality of life and financial cost). For example, an increased radiation dose usually results in both a higher probability of tumour control, but a concomitant higher probability of normal-tissue complications. Identifying the right balance between harm and benefit is a deeply personal choice that can vary substantially among patients.

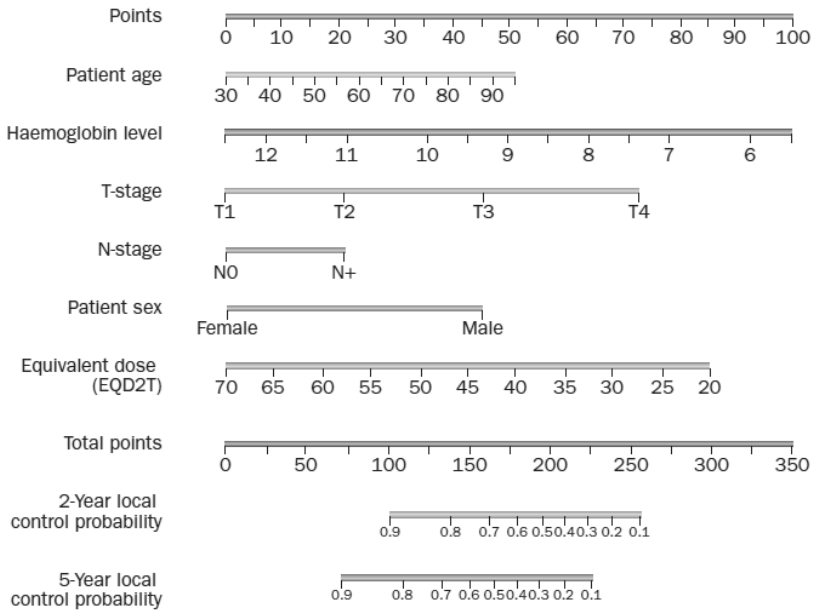


Figure 4

A published nomogram for local control in patients with cancer of the larynx treated with radiotherapy. Clinical and treatment variables are associated with local control status at follow-up durations of 2 and 5 years. The predictors are age of the patient (in years), haemoglobin level (in mmol/l), clinical tumour stage (T-stage), clinical nodal stage (N-stage), patient's sex and equivalent dose (in Gy). A probability for local control can be calculated by drawing a vertical line from each predictor value to the score scale at the top—'points'. After manually summing up the scores, the 'total points' correspond to the probability of local control, which are estimated by drawing a vertical line from this value to the bottom scales to estimate local control.[181]

Thus, a CDSS should simultaneously predict local control, survival, treatment toxicity, quality of life and cost. The system should represent these predictions and the balance between them in a way that is not only clear to the physician, but also to the patient, to achieve shared decision making.

Additionally, any prediction using a CDSS should be accompanied by a confidence interval. Accurately evaluating the confidence interval is an active and challenging area of research because uncertainties in the input features, missing features, size and quality of the training set and the intrinsic uncertainty of cancer must be incorporated to specify the uncertainty in the prediction for an individual patient. Without knowing if two possible decisions have a statistically significant and clinically meaningful difference in outcome, clinical decision support is difficult. Always sharing the data on which the model was based is a crucial prerequisite for this effort.

Current prediction models for decision support can only assist in distinguishing very high-level decisions—such as palliative versus curative treatment, sequential versus concurrent

chemoradiation, surgery versus a watch-and-wait approach. The radiation oncology community, however, is probably more interested in decisions such as intensity-modulated radiotherapy versus 3D-conformal radiotherapy or accelerated versus nonaccelerated treatment. The current prediction models are simply not trained on datasets with these detailed subgroups and are not, therefore, accurate enough to support these decisions. Whether learning from increasingly diverse patient groups and adding other features will sufficiently improve the current models is unclear. As a result, tightly controlled studies using evidence-based medicine approaches are still crucial to guide clinical practice.

Finally, CDSSs should be seen as medical devices that require stringent acceptance, commissioning and quality assurance by the local institute. The key part of the commissioning and subsequent quality assurance is to validate the accuracy of the prediction model in the local patient population. Indeed, local patient data should be collected and the predicted outcomes compared with actual outcomes to convince local physicians that the support system works in their local setting. This 'local validation' should be done at the commissioning stage, but should be repeated to ensure the decision support remains valid, despite changes in local practice. Validation studies need to indicate what will be the required commission frequency.

This required quality assurance also enables the improvement of the system as more patient data becomes available. Using routine patient data to extract knowledge and apply that knowledge immediately is called 'rapid learning'.[3, 189] Rapid learning via continuously updated CDSSs offers a way to quickly learn from retrospective data and include new data sets (such as randomized controlled trial results) to adapt treatment protocols and deliver personalized decision support.

As a data-driven discipline with well-established standards, such as DICOM–RT (digital imaging and communications in medicine in radiotherapy), radiotherapy offers an excellent starting point for adopting these rapid-learning principles (Figure 5). Aside from the importance of local data capture, which is still often lacking for (patient-reported) outcome and toxicity in particular, the quantity and heterogeneity of data that is necessary for rapid learning requires the pooling of data in a multi-institutional, international fashion.[190,191] One method of pooling data is to replicate routine clinical data sources in a distributed de-identified data warehouse, such as what is done in an international Computer-Aided Theragnostics network.[192] Examples of initiatives that create large centralized data and tissue infrastructures for routine radiation oncology patients are GENEPI,[193] the Radiogenomics Consortium,[194] ALLEGRO[58] and ULICE.[195] These initiatives also facilitate studies for external validation, reproducibility and hypothesis generation.[190]

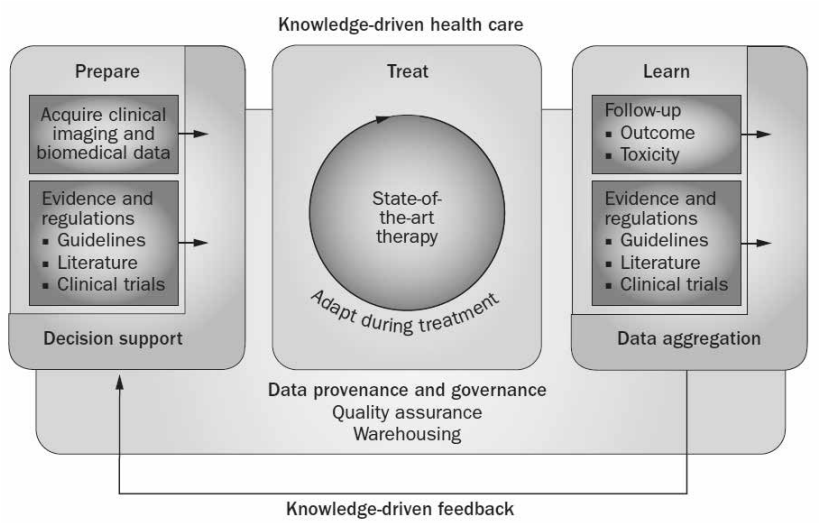


Figure 5

Knowledge-driven health-care principles using a clinical decision-support system in conjunction with standard evidence and regulations to choose the optimal treatment. In learning from follow-up data, knowledge is fed back to improve the clinical decision-support system and adapt regulations.

As datasets become larger (both in number of patients and in number of features per patient) high-throughput methods, both molecular[196–201] and imaging-based,[79] can produce large numbers of features that correlate with outcome.[68,70,202–204] A limited application of these techniques has already transformed our understanding of radiotherapy response. For example, GWAS have associated SNPs with radiation toxicity.[205,206] Similarly, mRNA-abundance microarrays have been used to predict tumour response and normal-tissue toxicity in both patient and *in-vitro* studies,[207–211] as well as to create markers that reflect biological phenotypes that are important for radiation response, such as hypoxia[212,213] and proliferation.[214] Both the data analysis and validation are important but challenging aspects of model development.[196,215] For example, the studies described above[207–214] suffer from the substantial multiple-testing problem (that is, a large number of measured features compared with the sample number), which renders their results preliminary. Human input and large, robust validation studies are, therefore, needed before features from high-throughput techniques can be included in CDSSs.[216–218]

Although studies on a single feature can be informative, only its combination into multimodal, multivariate models can be expected to provide a more holistic view of the response to radiation. By combining events at different levels using systems-biology-like approaches, creating tumour-specific and patient-specific models of the effects and implications of radiation therapy should become possible (Figure 6). Indeed, future studies will not only need to identify the individual components related to radiation response, but will also need to establish the interactions and relations amongst them.[219] Although this approach has not yet been applied to model radiotherapy responses, at least one study has

demonstrated that combining multiple high-throughput data types can be used to map molecular cancer characteristics.[220] Combining models at different levels (societal, patient, whole tumour or organ, local tumour or organ, and cellular) is expected to lead to an increasingly holistic and accurate CDSS for the individual patient. Evidence that longitudinal data have added value to predicting outcome in, for example, repeated PET-imaging[221] and tumour-perfusion[222] studies is growing, implying that this data need to be taken into account as candidates for future CDSSs.

Despite the challenges that remain, the vision of predictive models leading to CDSSs that are continuously updated via rapid learning on large datasets is clear, and numerous steps have already been taken. These include universal data-quality assurance programmes and semantic interoperability issues.[223] However, we believe that this truly innovative journey will lead to necessary improvement of healthcare effectiveness and efficiency. Indeed, investments are being made in research and innovation for health-informatics systems, with an emphasis on interoperability and standards for secured data transfer, which shows that 'eHealth' will be among the largest health-care innovations of the coming decade.[223,22]

CONCLUSIONS

Accurate, externally validated prediction models are being rapidly developed, whereby multiple features related to the patient's disease are combined into an integrated prediction. The key, however, is standardization—mainly in data acquisition across all areas, including molecular-based and imaging-based assays, patient preferences and possible treatments. Standardization requires harmonized clinical guidelines, regulated image acquisition and analysis parameters, validated biomarker assay criteria and data-sharing methods that use identical ontologies.

Assessing the clinical usefulness of any CDSS is just as important as standardizing the development of externally validated accurate prediction models with high-quality data, preferably by standardizing the design of clinical trials. These crucial steps are the basis of validating a CDSS, which, in turn, will stimulate developments in rapid-learning healthcare and will enable the next major advances in shared decision making.

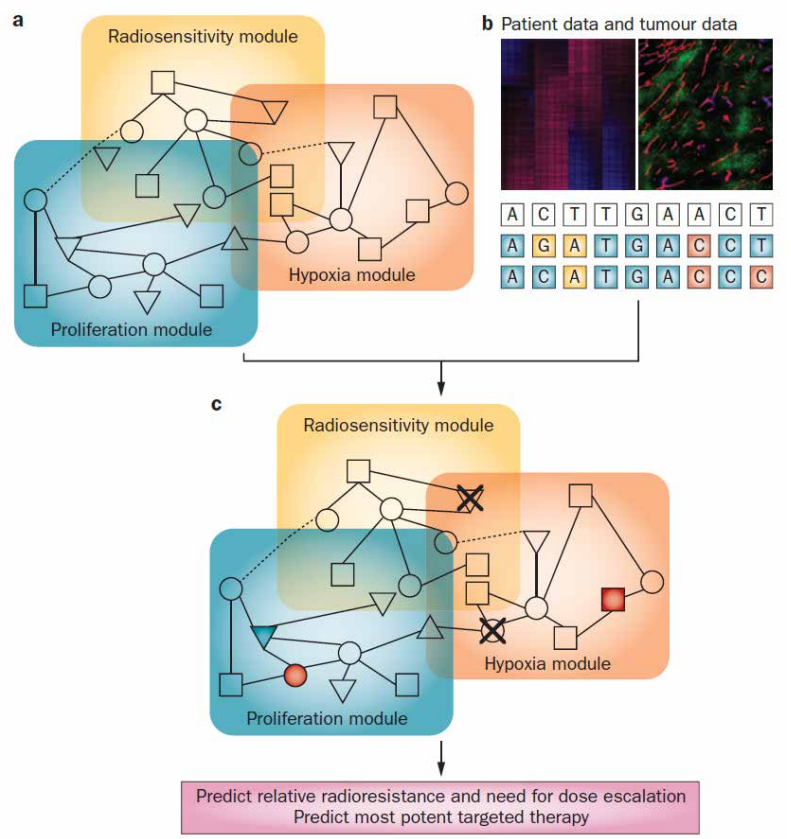


Figure 6

A simplified schematic representation of systems biology applied to radiotherapy. ***a*** *| On the basis of in-vitro, in-vivo and patient data, modules representing the three biological categories (gene expression, immunohistochemical data and mutation data) important for radiotherapy response can be created.* ***b*** *| For an individual patient, appropriate molecular data will be accumulated.* ***c*** *| Combining the individual patient data with the modules will provide knowledge on specific module alterations (such as a deletion [X], upregulation [red] or downregulation [blue]), which can be translated to information on relative radioresistance and the molecular 'weak' spots of the tumour. This information will subsequently indicate whether dose escalation is necessary and which targeted drug is most effective for the patient. Part b used with permission from the National Academy of Sciences © Duboisa, L. J. Proc. Natl Acad. Sci. USA 108, 14620–14625 (2011).*

ACKNOWLEDGMENTS

We acknowledge financial support from the Center for Translational Molecular Medicine framework (AIR FORCE), European Union sixth and seventh framework programme (ART-FORCE and METOXIA), INTERREG (www.eurocat.info), QuIC-ConCePT (funded by the Inno-

vative Medicine Initiative Joint Undertaking) and the Dutch Cancer Society (KWF UM 2011-5020 and KWF UM 2009-4454).

REFERENCES

1. Vogelzang, N. J. *et al.* Clinical cancer advances 2011: annual report on progress against cancer from the American Society of Clinical Oncology. *J. Clin. Oncol.* 30, 88–109 (2012).
2. Fraass, B. A. & Moran, J. M. Quality, technology and outcomes: evolution and evaluation of new treatments and/or new technology. *Semin. Radiat. Oncol.* 22, 3–10 (2012).
3. Abernethy, A. P. *et al.* Rapid-learning system for cancer care. *J. Clin. Oncol.* 28, 4268–4274 (2010).
4. Maitland, M. L. & Schilsky, R. L. Clinical trials in the era of personalized oncology. *CA Cancer J. Clin.* 61, 365–381 (2011).
5. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366, 883–892 (2012).
6. Bachtiary, B. *et al.* Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity. *Clin. Cancer Res.* 12, 5632–5640 (2006).
7. Boyd, C. A., Benarroch-Gampel, J., Sheffield, K. M., Cooksley, C. D. & Riall, T. S. 415 patients with adenosquamous carcinoma of the pancreas: a population-based analysis of prognosis and survival. *J. Surg. Res.* 174, 12–19 (2012).
8. Milosevic, M. F. *et al.* Interstitial fluid pressure in cervical carcinoma: within tumor heterogeneity, and relation to oxygen tension. *Cancer* 82, 2418–2426 (1998).
9. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 468, 346–352 (2012).
10. Suit, H., Skates, S., Taghian, A., Okunieff, P. & Efird, J. T. Clinical implications of heterogeneity of tumor response to radiation therapy. *Radiother. Oncol.* 25, 251–260 (1992).
11. Aerts, H. J. *et al.* Identification of residual metabolic-active areas within NSCLC tumours using a pre-radiotherapy FDG-PET-CT scan: a prospective validation. *Lung Cancer* 75, 73–76 (2012).
12. Aerts, H. J. *et al.* Identification of residual metabolic-active areas within individual NSCLC tumours using a pre-radiotherapy (18)Fluorodeoxyglucose–PET–CT scan. *Radiother. Oncol.* 91, 386–392 (2009).
13. Vickers, A. J. Prediction models: revolutionary in principle, but do they do more good than harm? *J. Clin. Oncol.* 29, 2951–2952 (2011).
14. Bright, T. J. *et al.* Effect of clinical decision-support systems: a systematic review. *Ann. Intern. Med.* 157 29–43 (2012).
15. Clark, G. M. Prognostic factors versus predictive factors: examples from a clinical trial of erlotinib. *Mol. Oncol.* 1, 406–412 (2008).
16. Dancey, J. E. *et al.* Guidelines for the development and incorporation of biomarker studies in early clinical trials of novel agents. *Clin. Cancer Res.* 16, 1745–1755 (2010).
17. Peek, N., Arts, D. G., Bosman, R. J., van der Voort, P. H. & de Keizer, N. F. External validation of prognostic models for critically ill patients required substantial sample sizes. *J. Clin. Epidemiol.* 60, 491–501 (2007).
18. Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. & Habbema, J. D. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J. Clin. Epidemiol.* 58, 475–483 (2005).
19. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 21, 128–138 (2010).
20. Aittokallio, T. Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Brief. Bioinform.* 11, 253–264 (2010).
21. Ludbrook, J. Outlying observations and missing values: how should they be handled? *Clin. Exp. Pharmacol. Physiol.* 35, 670–678 (2008).

22. Jayalakshmi, T. & Santhakumaran, A. Statistical normalization and back propagation for classification. *Int. J. Comput. Theory Eng.* 3, 89–93 (2011).
23. Huan, L. & Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining* (Kluwer Academic Publishers, Norwell, MA, 1998).
24. Harrell, F. E. *Regression Modeling Strategies* (Springer, New York, 2001).
25. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, New York, 2007).
26. Lee, S. M. & Abbott, P. A. Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers. *J. Biomed. Inform.* 36, 389–399 (2003).
27. Cruz, J. A. & Wishart, D. S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* 2, 59–77 (2007).
28. Putter, H., Fiocco, M. & Geskus, R. B. Tutorial in biostatistics: competing risks and multi-state models. *Stat. Med.* 26, 2389–2430 (2007).
29. Moons, K. G. *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 98, 691–698 (2012).
30. Dehing-Oberije, C. *et al.* Development, external validation and clinical usefulness of a practical prediction model for radiation-induced dysphagia in lung cancer patients. *Radiother. Oncol.* 97, 455–461 (2010).
31. Specht, M. C., Kattan, M. W., Gonen, M., Fey, J. & Van Zee, K. J. Predicting nonsentinel node status after positive sentinel lymph biopsy for breast cancer: clinicians versus nomogram. *Ann. Surg. Oncol.* 12, 654–659 (2005).
32. Dehing-Oberije, C. *et al.* Tumor volume combined with number of positive lymph node stations is a more important prognostic factor than TNM stage for survival of non-small-cell lung cancer patients treated with (chemo)radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.* 70, 1039–1044 (2008).
33. Vickers, A. J., Kramer, B. S. & Baker, S. G. Selecting patients for randomized trials: a systematic approach based on risk group. *Trials* 7, 30 (2006).
34. Baggerly, K. A. & Coombes, K. R. What information should be required to support clinical "omics" publications? *Clin. Chem.* 57, 688–690 (2011).
35. Klopp, A. H. & Eifel, P. J. Biological predictors of cervical cancer response to radiation therapy. *Semin. Radiat. Oncol.* 22, 143–150 (2012).
36. Kristiansen, G. Diagnostic and prognostic molecular biomarkers for prostate cancer. *Histopathology* 60, 125–141 (2012).
37. Dehing-Oberije, C. *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. Biol. Phys.* 74, 355–362 (2009).
38. Ang, C. S., Phung, J. & Nice, E. C. The discovery and validation of colorectal cancer biomarkers. *Biomed. Chromatogr.* 25, 82–99 (2011).
39. Schmidt, M. E. & Steindorf, K. Statistical methods for the validation of questionnaires--discrepancy between theory and practice. *Methods Inf. Med.* 45, 409–413 (2006).
40. Garrido-Laguna, I. *et al.* Validation of the Royal Marsden Hospital prognostic score in patients treated in the Phase I Clinical Trials Program at the MD Anderson Cancer Center. *Cancer* 118, 1422–1428 (2012).
41. Shrier, I. *et al.* Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *Am. J. Epidemiol.* 166, 1203–1209 (2007).
42. Tzoulaki, I., Siontis, K. C. & Ioannidis, J. P. Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. *BMJ* 343, d6829 (2011).
43. Trotti, A., Colevas, A. D., Setser, A. & Basch, E. Patient-reported outcomes and the evolution of adverse event reporting in oncology. *J. Clin. Oncol.* 25, 5121–5127 (2007).
44. Trotti, A. *et al.* CTCAE v3.0: development of a comprehensive grading system for the adverse effects of cancer treatment. *Semin. Radiat. Oncol.* 13, 176–181 (2003).
45. Golder, S., Loke, Y. K. & Bland, M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. *PLoS Med.* 8, e1001026 (2011).

46. Steg, P. G. *et al.* External validity of clinical trials in acute myocardial infarction. *Arch. Intern. Med.* 167, 68–73 (2007).
47. Little, J. *et al.* Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE statement. *Ann. Intern. Med.* 150, 206–215 (2009).
48. von Elm, E. *et al.* The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* 370, 1453–1457 (2007).
49. Dawson, L. A. & Sharpe, M. B. Image-guided radiotherapy: rationale, benefits, and limitations. *Lancet Oncol.* 7, 848–858 (2006).
50. Bentzen, S. M. *et al.* Quantitative analyses of normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues. *Int. J. Radiat. Oncol. Biol. Phys.* 76 (Suppl. 3), 3–9 (2010).
51. Verellen, D. *et al.* Innovations in image-guided radiotherapy. *Nat. Rev. Cancer* 7, 949–960 (2007).
52. Holthusen, H. Erfahrungen über die Verträglichkeitsgrenze für Röntgenstrahlen und deren nutzanwendung zur verhütung von schäden [German]. *Strahlentherapie* 57, 254–269 (1936).
53. Valentini, V. *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.* 29, 3163–3172 (2011).
54. Belderbos, J. *et al.* Randomised trial of sequential versus concurrent chemo-radiotherapy in patients with inoperable non-small cell lung cancer (EORTC 08972–22973). *Eur. J. Cancer* 43, 114–121 (2007).
55. Lambin, P. *et al.* The ESTRO Breur Lecture 2009. From population to voxel-based radiotherapy: exploiting intra-tumour and intra-organ heterogeneity for advanced treatment of non-small cell lung cancer. *Radiother. Oncol.* 96, 145–152 (2010).
56. Graham, M. V. *et al.* Clinical dose-volume histogram analysis for pneumonitis after 3D treatment for non-small cell lung cancer (NSCLC). *Int. J. Radiat. Oncol. Biol. Phys.* 45, 323–329 (1999).
57. Emami, B. *et al.* Tolerance of normal tissue to therapeutic irradiation. *Int. J. Radiat. Oncol. Biol. Phys.* 21, 109–122 (1991).
58. Ottolenghi, A., Smyth, V. & Trott, K. R. The risks to healthy tissues from the use of existing and emerging techniques for radiation therapy. *Radiat. Prot. Dosimetry* 143, 533–535 (2011).
59. Beetz, I. *et al.* NTCP models for patient-rated xerostomia and sticky saliva after treatment with intensity modulated radiotherapy for head and neck cancer: the role of dosimetric and clinical factors. *Radiother. Oncol.* http://dx.doi.org/10.1016/j.radonc.2012.03.004.
60. van der Schaaf, A. *et al.* Multivariate modeling of complications with data driven variable selection: guarding against overfitting and effects of data set size. *Radiother. Oncol.* http://dx.doi.org/10.1016/j.radonc.2011.12.006.
61. Xu, C.-J., van der Schaaf, A., van't Veld, A. A., Langendijk, J. A. & Schilstra, C. Statistical validation of normal tissue complication probability models. *Int. J. Radiat. Oncol. Biol. Phys.* 84, e123–e129 (2012).
62. Nijsten, S. M., Mijnheer, B. J., Dekker, A. L., Lambin, P. & Minken, A. W. Routine individualised patient dosimetry using electronic portal imaging devices. *Radiother. Oncol.* 83, 65–75 (2007).
63. van Elmpt, W., Petit, S., De Ruysscher, D., Lambin, P. & Dekker, A. 3D dose delivery verification using repeated cone-beam imaging and EPID dosimetry for stereotactic body radiotherapy of non-small cell lung cancer. *Radiother. Oncol.* 94, 188–194 (2010).
64. van Elmpt, W. *et al.* 3D *in vivo* dosimetry using megavoltage cone-beam CT and EPID dosimetry. *Int. J. Radiat. Oncol. Biol. Phys.* 73, 1580–1587 (2009).
65. Rodrigues, G., Lock, M., D'Souza, D., Yu, E. & Van Dyk, J. Prediction of radiation pneumonitis by dose–volume histogram parameters in lung cancer—a systematic review. *Radiother. Oncol.* 71, 127–138 (2004).
66. Werner-Wasik, M., Yorke, E., Deasy, J., Nam, J. & Marks, L. B. Radiation dose-volume effects in the esophagus. *Int. J. Radiat. Oncol. Biol. Phys.* 76 (Suppl. 3), S86–S93 (2010).
67. Saunders, M., Rojas, A. M. & Dische, S. CHART revisited: a conservative approach for advanced head and neck cancer. *Clin. Oncol. (R. Coll. Radiol.)* 20, 127–133 (2008).
68. Turner, N. *et al.* Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets. *Oncogene* 29, 2013–2023 (2010).

69. Begg, A. C. *et al.* The value of pretreatment cell kinetic parameters as predictors for radiotherapy outcome in head and neck cancer: a multicenter analysis. *Radiother. Oncol.* 50, 13–23 (1999).
70. Taguchi, F. *et al.* Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: a multicohort cross-institutional study. *J. Natl Cancer Inst.* 99, 838–846 (2007).
71. Hessel, F. *et al.* Impact of increased cell loss on the repopulation rate during fractionated irradiation in human FaDu squamous cell carcinoma growing in nude mice. *Int. J. Radiat. Biol.* 79, 479–486 (2003).
72. Baumann, M., Krause, M. & Hill, R. Exploring the role of cancer stem cells in radioresistance. *Nat. Rev. Cancer* 8, 545–554 (2008).
73. Ben-Josef, E. *et al.* Impact of overall treatment time on survival and local control in patients with anal cancer: a pooled data analysis of Radiation Therapy Oncology Group trials 87–04 and 98–11. *J. Clin. Oncol.* 28, 5061–5066 (2010).
74. Thames, H. D. *et al.* The role of overall treatment time in the outcome of radiotherapy of prostate cancer: an analysis of biochemical failure in 4839 men treated between 1987 and 1995. *Radiother. Oncol.* 96, 6–12 (2010).
75. Fass, L. Imaging and cancer: A review. *Mol. Oncol.* 2, 115–152 (2008).
76. Torigian, D. A., Huang, S. S., Houseni, M. & Alavi, A. Functional imaging of cancer with emphasis on molecular techniques. *CA Cancer J. Clin.* 57, 206–224 (2007).
77. Eadie, L. H., Taylor, P. & Gibson, A. P. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. *Eur. J. Radiol.* 81, e70–e76 (2012).
78. Gillies, R. J., Anderson, A. R., Gatenby, R. A. & Morse, D. L. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clin. Radiol.* 65, 517–521 (2010).
79. Lambin, P. *et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* 48, 441–446 (2012).
80. Velazquez, E. R., Aerts, H. J., Oberije, C., De Ruysscher, D. & Lambin, P. Prediction of residual metabolic activity after treatment in NSCLC patients. *Acta Oncol.* 49, 1033–1039 (2010).
81. Cangir, A. K. *et al.* Prognostic value of tumor size in non-small cell lung cancer larger than five centimeters in diameter. *Lung Cancer* 46, 325–331 (2004).
82. Lam, J. S. *et al.* Prognostic relevance of tumour size in T3a renal cell carcinoma: a multicentre experience. *Eur. Urol.* 52, 155–162 (2007).
83. Pitson, G. *et al.* Tumor size and oxygenation are independent predictors of nodal diseases in patients with cervix cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 51, 699–703 (2001).
84. Thomas, F. *et al.* Radical radiotherapy alone in non-operable breast cancer: The major impact of tumor size and histological grade on prognosis. *Radiother. Oncol.* 13, 267–276 (1988).
85. Steenbakkers, R. J. *et al.* Observer variation in target volume delineation of lung cancer related to radiation oncologist-computer interaction: a 'Big Brother' evaluation. *Radiother. Oncol.* 77, 182–190 (2005).
86. Greco, C., Rosenzweig, K., Cascini, G. L. & Tamburrini, O. Current status of PET/CT for tumour volume definition in radiotherapy treatment planning for non-small cell lung cancer (NSCLC). *Lung Cancer* 57, 125–134 (2007).
87. Caldwell, C. B. *et al.* Observer variation in contouring gross tumor volume in patients with poorly defined non-small-cell lung tumors on CT: the impact of 18FDG-hybrid PET fusion. *Int. J. Radiat. Oncol. Biol. Phys.* 51, 923–931 (2001).
88. Bowden, P. *et al.* Measurement of lung tumor volumes using three-dimensional computer planning software. *Int. J. Radiat. Oncol. Biol. Phys.* 53, 566–573 (2002).
89. Nishino, M. *et al.* CT tumor volume measurement in advanced non-small-cell lung cancer: performance characteristics of an emerging clinical tool. *Acad. Radiol.* 18, 54–62 (2011).
90. Marcus, C. D. *et al.* Imaging techniques to evaluate the response to treatment in oncology: current standards and perspectives. *Crit. Rev. Oncol. Hematol.* 72, 217–238 (2009).
91. Schwartz, L. H., Mazumdar, M., Brown, W., Smith, A. & Panicek, D. M. Variability in response assessment in solid tumors: effect of number of lesions chosen for measurement. *Clin. Cancer Res.* 9, 4318–4323 (2003).

92. Erasmus, J. J. *et al.* Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. *J. Clin. Oncol.* 21, 2574–2582 (2003).
93. Therasse, P. Measuring the clinical response. What does it mean? *Eur. J. Cancer* 38, 1817–1823 (2002).
94. Nehmeh, S. A. & Erdi, Y. E. Respiratory motion in positron emission tomography/computed tomography: a review. *Semin. Nucl. Med.* 38, 167–176 (2008).
95. Sonke, J. J. & Belderbos, J. Adaptive radiotherapy for lung cancer. *Semin. Radiat. Oncol.* 20, 94–106 (2010).
96. van Baardwijk, A. *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. Biol. Phys.* 68, 771–778 (2007).
97. Wu, K. *et al.* PET CT thresholds for radiotherapy target definition in non-small-cell lung cancer: how close are we to the pathologic findings? *Int. J. Radiat. Oncol. Biol. Phys.* 77, 699–706 (2010).
98. Wanet, M. *et al.* Gradient-based delineation of the primary GTV on FDG-PET in non-small cell lung cancer: a comparison with threshold-based approaches, CT and surgical specimens. *Radiother. Oncol.* 98, 117–125 (2011).
99. Strassmann, G. *et al.* Atlas-based semiautomatic target volume definition (CTV) for head-and-neck tumors. *Int. J. Radiat. Oncol. Biol. Phys.* 78, 1270–1276 (2010).
100. Nestle, U. *et al.* Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. *J. Nucl. Med.* 46, 1342–1348 (2005).
101. Daisne, J. F. *et al.* Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical specimen. *Radiology* 233, 93–100 (2004).
102. van Loon, J. *et al.* Therapeutic implications of molecular imaging with PET in the combined modality treatment of lung cancer. *Cancer Treat. Rev.* 37, 331–343 (2011).
103. Wood, K. A., Hoskin, P. J. & Saunders, M. I. Positron emission tomography in oncology: a review. *Clin. Oncol.* 19, 237–255 (2007).
104. O'Connor, J. P. *et al.* Quantitative imaging biomarkers in the clinical development of targeted therapeutics: current and future perspectives. *Lancet Oncol.* 9, 766–776 (2008).
105. van Baardwijk, A. *et al.* Time trends in the maximal uptake of FDG on PET scan during thoracic radiotherapy. A prospective study in locally advanced non-small cell lung cancer (NSCLC) patients. *Radiother. Oncol.* 82, 145–152 (2007).
106. Rodney, J. H. PET for therapeutic response monitoring in oncology. *PET Clinics* 3, 89–99 (2008).
107. Chung, H. H. *et al.* Prognostic value of metabolic tumor volume measured by FDG-PET/CT in patients with cervical cancer. *Gynecol. Oncol.* 120, 270–274 (2011).4. Garrido P, Gonzalez-Larriba JL, Insa A, et al., Long-term survival associated with complete resection after induction chemotherapy in stage IIIA (N2) and IIIB (T4N0-1) non small-cell lung cancer patients: the Spanish Lung Cancer Group Trial 9901, J Clin Oncol 25 (30), 4736-4742 (2007).
108. Borst, G. R. *et al.* Standardised FDG uptake: a prognostic factor for inoperable non-small cell lung cancer. *Eur. J. Cancer* 41, 1533–1541 (2005).
109. Mac Manus, M. P. *et al.* Metabolic (FDG–PET) response after radical radiotherapy/chemoradiotherapy for non-small cell lung cancer correlates with patterns of failure. *Lung Cancer* 49, 95–108 (2005).
110. Hoekstra, C. J. *et al.* Prognostic relevance of response evaluation using [18F]-2-fluoro-2-deoxy-d-glucose positron emission tomography in patients with locally advanced non-small-cell lung cancer. *J. Clin. Oncol.* 23, 8362–8370 (2005).
111. Soto, D. E., Kessler, M. L., Piert, M. & Eisbruch, A. Correlation between pretreatment FDG–PET biological target volume and anatomical location of failure after radiation therapy for head and neck cancers. *Radiother. Oncol.* 89, 13–18 (2008).
112. Lambrecht, M. *et al.* The use of FDG–PET/CT and diffusion-weighted magnetic resonance imaging for response prediction before, during and after preoperative chemoradiotherapy for rectal cancer. *Acta Oncol.* 49, 956–963 (2010).

113. Janssen, M. H. M. *et al.* Evaluation of early metabolic responses in rectal cancer during combined radiochemotherapy or radiotherapy alone: Sequential FDG–PET–CT findings. *Radiother. Oncol.* 94, 151–155 (2010).
114. Ceulemans, G. *et al.* Can 18-FDG-PET during radiotherapy replace post-therapy scanning for detection/demonstration of tumor response in head-and-neck cancer? *Int. J. Radiat. Oncol. Biol. Phys.* 81, 938–942 (2011).
115. van Loon, J. *et al.* Early CT and FDG-metabolic tumour volume changes show a significant correlation with survival in stage I-III small cell lung cancer: a hypothesis generating study. *Radiother. Oncol.* 99, 172–175 (2011).
116. Bussink, J., Kaanders, J. H., van der Graaf, W. T. & Oyen, W. J. PET–CT for radiotherapy treatment planning and response monitoring in solid tumors. *Nat. Rev. Clin. Oncol.* 8, 233–242 (2011).
117. Boellaard, R. Need for standardization of 18F-FDG PET/CT for treatment response assessments. *J. Nucl. Med.* 52 (Suppl. 2), 93–100 (2011).
118. Boellaard, R. *et al.* The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials. *Eur. J. Nucl. Med. Mol. Imaging* 35, 2320–2333 (2008).
119. Boellaard, R. *et al.* FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur. J. Nucl. Med. Mol. Imaging* 37, 181–200 (2010).
120. Bayouth, J. E. *et al.* Image-based biomarkers in clinical practice. *Semin. Radiat. Oncol.* 21, 157–166 (2011).
121. Harry, V. N., Semple, S. I., Parkin, D. E. & Gilbert, F. J. Use of new imaging techniques to predict tumour response to therapy. *Lancet Oncol.* 11, 92–102 (2010).
122. Heijmen, L. *et al.* Tumour response prediction by diffusion-weighted MR imaging: ready for clinical use? *Crit. Rev. Oncol. Hematol.* 83, 194–207 (2012).
123. Lambrecht, M. *et al.* The prognostic value of pretherapeutic diffusion-weighted MRI in oropharyngeal carcinoma treated with (chemo-)radiotherapy. *Cancer Imaging* 11, S112–S113 (2011).
124. Vandecaveye, V. *et al.* Diffusion-weighted magnetic resonance imaging early after chemoradiotherapy to monitor treatment response in head-and-neck squamous cell carcinoma. *Int. J. Radiat. Oncol. Biol. Phys.* 82, 1098–1107 (2012).
125. Kim, S. Y. *et al.* Malignant hepatic tumors: short-term reproducibility of apparent diffusion coefficients with breath-hold and respiratory-triggered diffusion-weighted MR imaging. *Radiology* 255, 815–823 (2010).
126. Sinkus, R., Van Beers, B. E., Vilgrain, V., Desouza, N. & Waterton, J. C. Apparent diffusion coefficient from magnetic resonance imaging as a biomarker in oncology drug development. *Eur. J. Cancer* 48, 425–431 (2012).
127. Kierkels, R. G. *et al.* Comparison between perfusion computed tomography and dynamic contrast-enhanced magnetic resonance imaging in rectal cancer. *Int. J. Radiat. Oncol. Biol. Phys.* 77, 400–408 (2010).
128. Shukla-Dave, A. *et al.* Dynamic contrast-enhanced magnetic resonance imaging as a predictor of outcome in head and neck squamous cell carcinoma patients with nodal metastases. *Int. J. Radiat. Oncol. Biol. Phys.* 82, 1837–1844 (2012).
129. Yaromina, A. *et al.* Co-localisation of hypoxia and perfusion markers with parameters of glucose metabolism in human squamous cell carcinoma (hSCC) xenografts. *Int. J. Radiat. Biol.* 85, 972–980 (2009).
130. Mörchel, P. *et al.* Correlating quantitative MR measurements of standardized tumor lines with histological parameters and tumor control dose. *Radiother. Oncol.* 96, 123–130 (2010).
131. Quennet, V. *et al.* Tumor lactate content predicts for response to fractionated irradiation of human squamous cell carcinomas in nude mice. *Radiother. Oncol.* 81, 130–135 (2006).
132. Kim, Y. I. *et al.* Multiphase contrast-enhanced CT imaging in hepatocellular carcinoma correlation with immunohistochemical angiogenic activities. *Acad. Radiol.* 14, 1084–1091 (2007).
133. Miles, K. A. Perfusion CT for the assessment of tumour vascularity: which protocol? *Br. J. Radiol.* 76, S36–S42 (2003).
134. Miles, K. A. Molecular imaging with dynamic contrast-enhanced computed tomography. *Clin. Radiol.* 65, 549–556 (2010).
135. Petralia, G. *et al.* CT perfusion in oncology: how to do it. *Cancer Imaging* 10, 8–19 (2010).

136. Asselin, M. C., O'Connor, J. P., Boellaard, R., Thacker, N. A. & Jackson, A. Quantifying heterogeneity in human tumours using MRI and PET. *Eur. J. Cancer* 48, 447–455 (2012).
137. Eary, J. F., O'Sullivan, F., O'Sullivan, J. & Conrad, E. U. Spatial heterogeneity in sarcoma 18F-FDG uptake as a predictor of patient outcome. *J. Nucl. Med.* 49, 1973–1979 (2008).
138. Tixier, F. *et al.* Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J. Nucl. Med.* 52, 369–378 (2011).
139. Diehn, M. *et al.* Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc. Natl Acad. Sci. USA* 105, 5213–5218 (2008).
140. Kuo, M. D., Gollub, J., Sirlin, C. B., Ooi, C. & Chen, X. Radiogenomic analysis to identify imaging phenotypes associated with drug response gene expression programs in hepatocellular carcinoma. *J. Vasc. Interv. Radiol.* 18, 821–831 (2007).
141. Segal, E. *et al.* Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat. Biotechnol.* 25, 675–680 (2007).
142. Rutman, A. M. & Kuo, M. D. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. *Eur. J. Radiol.* 70, 232–241 (2009).6. Zatloukal P, Petruzelka L, Zemanova M, et al., Concurrent versus sequential chemoradiotherapy with cisplatin and vinorelbine in locally advanced non-small cell lung cancer: a randomized study, Lung Cancer 46 (1), 87-98 (2004).
143. Lindegaard, J. C., Overgaard, J., Bentzen, S. M. & Pedersen, D. Is there a radiobiologic basis for improving the treatment of advanced stage cervical cancer? *J. Natl Cancer Inst. Monogr.* 105–112 (1996).
144. Slonina, D. & Gasin´ska, A. Intrinsic radiosensitivity of healthy donors and cancer patients as determined by the lymphocyte micronucleus assay. *Int. J. Radiat. Biol.* 72, 693–701 (1997).
145. Fertil, B. & Malaise, E. P. Intrinsic radiosensitivity of human cell lines is correlated with radioresponsiveness of human tumors: analysis of 101 published survival curves. *Int. J. Radiat. Oncol. Biol. Phys.* 11, 1699–1707 (1985).
146. West, C. M., Davidson, S. E., Roberts, S. A. & Hunter, R. D. The independence of intrinsic radiosensitivity as a prognostic factor for patient response to radiotherapy of carcinoma of the cervix. *Br. J. Cancer* 76, 1184–1190 (1997).
147. Björk-Eriksson, T., West, C., Karlsson, E. & Mercke, C. Tumor radiosensitivity (SF2) is a prognostic factor for local control in head and neck cancers. *Int. J. Radiat. Oncol. Biol. Phys.* 46, 13–19 (2000).
148. Bartelink, H. *et al.* Towards prediction and modulation of treatment response. *Radiother. Oncol.* 50, 1–11 (1999).
149. Begg, A. C. Predicting recurrence after radiotherapy in head and neck cancer. *Semin. Radiat. Oncol.* 22, 108–118 (2012).
150. Menegakis, A. *et al.* Prediction of clonogenic cell survival curves based on the number of residual DNA double strand breaks measured by γ-H2AX staining. *Int. J. Radiat. Biol.* 85, 1032–1041 (2009).
151. Olive, P. L. & Banáth, J. P. Phosphorylation of histone H2AX as a measure of radiosensitivity. *Int. J. Radiat. Oncol. Biol. Phys.* 58, 331–335 (2004).
152. Höckel, M. *et al.* Association between tumor hypoxia and malignant progression in advanced cancer of the uterine cervix. *Cancer Res.* 56, 4509–4515 (1996).
153. Vaupel, P. & Mayer, A. Hypoxia in cancer: significance and impact on clinical outcome. *Cancer Metastasis Rev.* 26, 225–239 (2007).
154. Chouaib, S. *et al.* Hypoxia promotes tumor growth in linking angiogenesis to immune escape. *Front. Immunol.* 3, 21 (2012).
155. Kaanders, J. H. *et al.* Pimonidazole binding and tumor vascularity predict for treatment outcome in head and neck cancer. *Cancer Res.* 62, 7066–7074 (2002).
156. Nordsmark, M. *et al.* The prognostic value of pimonidazole and tumour pO_2 in human cervix carcinomas after radiation therapy: a prospective international multi-center study. *Radiother. Oncol.* 80, 123–131 (2006).
157. Rouschop, K. M. A. *et al.* The unfolded protein response protects human tumor cells during hypoxia through regulation of the autophagy genes *MAP1LC3B* and *ATG5*. *J. Clin. Invest.* 120, 127–141 (2010).

158. Krause, B. J., Beck, R., Souvatzoglou, M. & Piert, M. PET and PET/CT studies of tumor tissue oxygenation. *Q. J. Nucl. Med. Mol. Imaging* 50, 28–43 (2006).
159. Dubois, L. J. *et al.* Preclinical evaluation and validation of [18F]HX4, a promising hypoxia marker for PET imaging. *Proc. Natl Acad. Sci. USA* 108, 14620–14625 (2011).
160. van Loon, J. *et al.* Selective nodal irradiation on basis of (18)FDG–PET scans in limited-disease small-cell lung cancer: a prospective study. *Int. J. Radiat. Oncol. Biol. Phys.* 77, 329–336 (2010).
161. West, C. M., Cooper, R. A., Loncaster, J. A., Wilks, D. P. & Bromley, M. Tumor vascularity: a histological measure of angiogenesis and hypoxia. *Cancer Res.* 61, 2907–2910 (2001).
162. Maciejewski, B., Withers, H. R., Taylor, J. M. & Hliniak, A. Dose fractionation and regeneration in radiotherapy for cancer of the oral cavity and oropharynx: tumor dose-response and repopulation. *Int. J. Radiat. Oncol. Biol. Phys.* 16, 831–843 (1989).
163. Suzuki, Y. *et al.* Prognostic impact of mitotic index of proliferating cell populations in cervical cancer patients treated with carbon ion beam. *Cancer* 115, 1875–1882 (2009).
164. Turesson, I., Nyman, J., Holmberg, E. & Odén, A. Prognostic factors for acute and late skin reactions in radiotherapy patients. *Int. J. Radiat. Oncol. Biol. Phys.* 36, 1065–1075 (1996).
165. Johansen, J., Bentzen, S. M., Overgaard, J. & Overgaard, M. Evidence for a positive correlation between *in vitro* radiosensitivity of normal human skin fibroblasts and the occurrence of subcutaneous fibrosis after radiotherapy. *Int. J. Radiat. Biol.* 66, 407–412 (1994).
166. West, C. M. *et al.* Lymphocyte radiosensitivity is a significant prognostic factor for morbidity in carcinoma of the cervix. *Int. J. Radiat. Oncol. Biol. Phys.* 51, 10–15 (2001).
167. Peacock, J. *et al.* Cellular radiosensitivity and complication risk after curative radiotherapy. *Radiother. Oncol.* 55, 173–178 (2000).
168. Russell, N. S. *et al.* Low predictive value of intrinsic fibroblast radiosensitivity for fibrosis development following radiotherapy for breast cancer. *Int. J. Radiat. Biol.* 73, 661–670 (1998).
169. Russell, N. S., Arlett, C. F., Bartelink, H. & Begg, A. C. Use of fluorescence *in situ* hybridization to determine the relationship between chromosome aberrations and cell survival in eight human fibroblast strains. *Int. J. Radiat. Biol.* 68, 185–196 (1995).
170. Kiltie, A. E. *et al.* A correlation between residual radiation-induced DNA double-strand breaks in cultured fibroblasts and late radiotherapy reactions in breast cancer patients. *Radiother. Oncol.* 51, 55–65 (1999).
171. Dileto, C. L. & Travis, E. L. Fibroblast radiosensitivity *in vitro* and lung fibrosis *in vivo*: comparison between a fibrosis-prone and fibrosis-resistant mouse strain. *Radiat. Res.* 146, 61–67 (1996).
172. Azria, D. *et al.* Concurrent or sequential adjuvant letrozole and radiotherapy after conservative surgery for early-stage breast cancer (CO-HO-RT): a phase 2 randomised trial. *Lancet Oncol.* 11, 258–265 (2010).
173. Bentzen, S. M. Preventing or reducing late side effects of radiation therapy: radiobiology meets molecular pathology. *Nat. Rev. Cancer* 6, 702–713 (2006).
174. Rodemann, H. P. & Bamberg, M. Cellular basis of radiation-induced fibrosis. *Radiother. Oncol.* 35, 83–90 (1995).
175. Andreassen, C. N., Alsner, J., Overgaard, M., Sørensen, F. B. & Overgaard, J. Risk of radiation-induced subcutaneous fibrosis in relation to single nucleotide polymorphisms in *TGFB1*, *SOD2*, *XRCC1*, *XRCC3*, *APEX* and *ATM*—a study based on DNA from formalin fixed paraffin embedded tissue samples. *Int. J. Radiat. Biol.* 82, 577–586 (2006).
176. Chang-Claude, J. *et al.* Association between polymorphisms in the DNA repair genes, *XRCC1*, *APE1*, and *XPD* and acute side effects of radiotherapy in breast cancer patients. *Clin. Cancer Res.* 11, 4802–4809 (2005).
177. Barnett, G. C. *et al.* Independent validation of genes and polymorphisms reported to be associated with radiation toxicity: a prospective analysis study. *Lancet Oncol.* 13, 65–77 (2012).
178. Cammann, H., Jung, K., Meyer, H. A. & Stephan, C. Avoiding pitfalls in applying prediction models, as illustrated by the example of prostate cancer diagnosis. *Clin. Chem.* 57, 1490–1498 (2011).
179. Iasonos, A., Schrag, D., Raj, G. V. & Panageas, K. S. How to build and interpret a nomogram for cancer prognosis. *J. Clin. Oncol.* 26, 1364–1370 (2008).

180. Dehing-Oberije, C. *et al.* Development and validation of a prognostic model using blood biomarker information for prediction of survival of non-small-cell lung cancer patients treated with combined chemotherapy and radiation or radiotherapy alone (NCT00181519, NCT00573040, and NCT00572325). *Int. J. Radiat. Oncol. Biol. Phys.* 81, 360–368 (2011).
181. Egelmeer, A. G. *et al.* Development and validation of a nomogram for prediction of survival and local control in laryngeal carcinoma patients treated with radiotherapy alone: a cohort study based on 994 patients. *Radiother. Oncol.* 100, 108–115 (2011).
182. van Stiphout, R. G. *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.* 98, 126–133 (2011).
183. Marko, N. F., Xu, Z., Gao, T., Kattan, M. W. & Weil, R. J. Predicting survival in women with breast cancer and brain metastasis: a nomogram outperforms current survival prediction models. *Cancer* 118, 3749–3757 (2011).
184. Rudloff, U. *et al.* Nomogram for predicting the risk of local recurrence after breast-conserving surgery for ductal carcinoma *in situ. J. Clin. Oncol.* 28, 3762–3769 (2010).
185. Adjuvant! Inc. *Adjuvant! Online* [online], http://www.adjuvantonline.com (2011).
186. Hajage, D. *et al.* External validation of Adjuvant! Online breast cancer prognosis tool. Prioritising recommendations for improvement. *PLoS ONE* 6, e27446 (2011).
187. Kuo, Y. L., Chen, D. R. & Chang, T. W. Accuracy validation of Adjuvant! Online in Taiwanese breast cancer patients—a 10-year analysis. *BMC Med. Inform. Decis. Mak.* 12, 108 (2012).
188. MAASTRO Clinic. *Cancer Prediction Models* [online], http://www.predictcancer.org (2012).
189. Ginsburg, G. S., Staples, J. & Abernethy, A. P. Academic medical centers: ripe for rapid-learning personalized health care. *Sci. Transl. Med.* 3, 101cm127 (2011).
190. Deasy, J. O. *et al.* Improving normal tissue complication probability models: the need to adopt a "data-pooling" culture. *Int. J. Radiat. Oncol. Biol. Phys.* 76 (Suppl. 3), S151–S154 (2010).
191. Roelofs, E. *et al.* Design of and technical challenges involved in a framework for multicentric radiotherapy treatment planning studies. *Radiother. Oncol.* 97, 567–571 (2010).
192. *Euregional Computer Assisted Theragnostics (EuroCAT) project* [online], http://www.eurocat.info (2012).
193. De Ruysscher, D. *et al.* First report on the patient database for the identification of the genetic pathways involved in patients over-reacting to radiotherapy: GENEPI–II. *Radiother. Oncol.* 97, 36–39 (2010).
194. West, C. *et al.* Establishment of a radiogenomics consortium. *Int. J. Radiat. Oncol. Biol. Phys.* 76, 1295–1296 (2010).
195. Kessel, K. A. *et al.* Connection of European particle therapy centers and generation of a common particle database system within the European ULICE-framework. *Radiat. Oncol.* 7, 115 (2012).
196. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* 11, 685–696 (2010).
197. Wulfkuhle, J. D., Liotta, L. A. & Petricoin, E. F. Proteomic applications for the early detection of cancer. *Nat. Rev. Cancer* 3, 267–275 (2003).
198. Pinkel, D. & Albertson, D. G. Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.* 37 (Suppl.), S11–S17 (2005)9. Bradley J, A review of radiation dose escalation trials for non-small cell lung cancer within the Radiation Therapy Oncology Group, Semin Oncol 32 (2 Suppl 3), S111-113 (2005).

PART 1

Clinical predictors

CHAPTER 3

Prediction of residual metabolic activity after treatment in NSCLC patients

Published in: *Acta Oncologica,* 2010; 49: 1033–1039

Prediction of residual metabolic activity after treatment in NSCLC patients

Emmanuel Rios Velazquez, Hugo J.W.L. Aerts, Cary Oberije, Dirk de Ruysscher and Philippe Lambin

ABSTRACT

Purpose

Metabolic response assessment is often used as a surrogate of local failure and survival. Early identification of patients with residual metabolic activity is essential as this enables selection of patients who could potentially benefit from additional therapy. We report on the development of a pre-treatment prediction model for metabolic response using patient, tumor and treatment factors.

Methods

One-hundred and one patients with inoperable NSCLC (stage I-IV), treated with 3-D conformal radical (chemo)-radiotherapy were retrospectively included in this study. All patients received a pre and post-radiotherapy fluorodeoxyglucose positron emission tomography-computed tomography FDG-PET-CT scan. The electronic medical record system and the medical patient charts were reviewed to obtain demographic, clinical, tumour and treatment data. Primary outcome measure was examined using a metabolic response assessment on a post-radiotherapy FDG-PET-CT scan. Radiotherapy was delivered in fractions of 1.8 Gy, twice a day, with a median prescribed dose of 60 Gy.

Results

Overall survival was worse in patients with residual metabolic active areas compared with the patients with a complete metabolic response ($p = 0.0001$). In univariate analysis, three variables were significantly associated with residual disease: larger primary gross tumor volume ($GTV_{primary}$, $p = 0.002$), higher pre-treatment maximum standardized uptake value (SUV_{max}, $p = 0.0005$) in the primary tumor and shorter overall treatment time (OTT, $p = 0.046$). A multivariate model including $GTV_{primary}$, SUV_{max}, equivalent radiation dose at 2 Gy corrected for time ($EQD_{2,\ T}$) and OTT yielded an area under the curve assessed by the leave-one-out cross validation of 0.71 (95 % CI, 0.65-0.76).

Conclusions

Our results confirmed the validity of metabolic response assessment as a surrogate of survival. We developed a multivariate model that is able to identify patients at risk of residual disease. These patients may benefit from an individualized and more adequate therapeutic approach, thereby improving local control and survival.

INTRODUCTION

Lung cancer is an important cause of cancer-related deaths worldwide [1]. In 2008, lung cancer was the most common cause of death from cancer with an estimate of 342, 000 deaths in Europe [1]. Non-small cell lung cancer (NSCLC) accounts for at least 80 % of all lung cancer cases [2]. The majority of these NSCLC patients present advanced-stage disease (stage III and IV), which are considered inoperable [3].

For these patients, the combination of radiotherapy and chemotherapy shows improved treatment outcome [4, 5], however local tumor failure is still observed in approximately 70 % of patients [6]. Therefore early identification of patients with a high risk of local treatment failure is important, as these patients may potentially benefit from additional therapy. One method of investigating local treatment failure, is assessing metabolic response within the primary tumor after treatment with ^{18}Fluorodeoxyglucose (FDG) positron emission tomography (PET) imaging [7].

Several studies indicated that patients with metabolically active residual masses after treatment have a poorer prognosis compared to patients without residual metabolic activity [8, 9]. Although, other studies have shown that FDG uptake before treatment is prognostic for residual metabolic activity within the tumor [9-11], other pre-treatment clinical factors were not investigated for their prognostic capability.

Therefore, we hypothesize that also other pre-treatment factors, including demographic, tumor and treatment characteristics, can have prognostic value for predicting metabolic response after treatment. In the present study we examined the association between commonly used prognostic factors in NSCLC patients and metabolic response after treatment in a univariate and multivariate analysis.

MATERIAL AND METHODS

Patient characteristics

The electronic medical record system and the medical patient charts were retrospectively reviewed to obtain demographic, clinical, tumour and treatment data. One-hundred and one patients (40 women and 61 men) with inoperable non-small cell lung cancer (NSCLC), stage I-IV, were included in this study. Their age ranged from 43 to 86 years (mean: 65.6 years). All patients were treated with curative intent at MAASTRO Clinic with sequential chemo-radiotherapy (82 patients) or with radical radiotherapy alone (19 patients) between December 2004 and September 2007. All patients received a pre and post-treatment FDG-PET-CT scan. For patients receiving sequential chemo-radiotherapy the pre-treatment scan was performed after chemotherapy. The average time interval be-

tween the last radiotherapy and the second FDG-PET-CT scan was 99 days (range: 49-184 days). No treatment was given between the end of radiotherapy and the post-treatment scan.

FDG-PET-CT Imaging

Pre and post-treatment FDG-PET-CT scans were performed using a Siemens Biograph (Siemens, Knoxville, TN). All patients were instructed to fast at least six hours before the intravenous administration of FDG (Tyco Health Care, Amsterdam, The Netherlands), followed by physiologic saline (10 mL). The total injected activity of FDG was dependent on the patient weight: (weight*4) + 20 Mbq. After a period of 45 minutes, during which the patient was encouraged to rest, PET and CT imaging were performed [12].

Treatment characteristics

The radiotherapy treatment was delivered in fractions of 1.8 Gy, twice a day, with a mean lung dose (MLD) restricted to 19 Gy and a maximal allowed total tumor dose (TTD) of 79.2 Gy [12]. Patients with stage III disease, who where physically fit enough received sequential chemo-radiotherapy, consisting of three courses of gemcitabine in combination with cisplatin or carboplatin, followed by radiotherapy as described for stage I/II. No concurrent chemo-radiotherapy was given. The biologic equivalent dose was used as indication of the intensity of chest RT delivered to the tumor and was calculated using the quadratic model [13] and corrected for overall treatment time.

Metabolic response

Metabolic response was assessed for all patients with a FGD-PET-CT scan after treatment. Residual disease was defined as residual metabolic activity within the primary tumor, i.e., areas with FDG uptake higher than in the aortic arch ($SUV > SUV_{AORTA}$) [7, 8]. If there was no activity within the tumor, patients were defined as with a complete metabolic response [10]. Survival data were obtained by reviewing the Dutch Communal Data register. Survival time was defined as the date from the start of radiotherapy until the date of death or last follow-up. Survival status could not be retrieved for one patient.

Statistical Analysis

All data are expressed as means ± SD. Because the distribution of the continuous variables was rather skewed, the Mann-Whitney U test was used to determine statistical differences between the patients with and without residual disease. For categorical variables the Chi-square test was used. Differences were considered to be significant when the p-value was lower than 0.05. The area under the curve (AUC) of the receiver operating characteristic (ROC), a plot of the true positive rate (correctly classified positive samples) and false positive rate (incorrectly classified negative samples) was used to analyze the associ-

ation between the variables and residual disease in univariate analysis using a proximal-support vector machine (p-SVM) [14]. A p-SVM was also used to build a multivariate prediction model, using metabolic residual disease as outcome measure. Combinatorial feature selection was performed to obtain an optimal subset of features. The set of variables with the highest AUC of the ROC curve was included in the multivariate predictive model. The Kaplan-Meier method was used to estimate survival probabilities and statistical differences were assessed using the log-rank test. Data were considered right-censored if patients were alive at the time of last follow-up. All the analyses were performed in Matlab 2008b (The MathWorks Inc, Natick, MA, USA) and SPSS (Version 15.0 for Windows, Chicago, IL).

RESULTS

Patient characteristics

To assess the power of clinical parameters for the prediction of metabolic response, commonly known prognostic factors were collected before treatment and correlated with metabolic response after treatment. A total of 101 NSCLC patients were included in this analysis, of which 56 (55 %) patients showed persistent residual FDG uptake on the post-radiotherapy CT-PET scan and 45 (45 %) patients had a complete metabolic response (CMR) indicating no residual FDG uptake within the tumor post-radiotherapy. Patient, tumor and treatment characteristics for both groups are listed in Table 1.

The median follow-up duration was 23.9 months (range: 3.8 – 55.5 months). The patients with residual active areas post-treatment had a significantly worse survival (median survival: 13.4 months) compared to patients with a complete metabolic response (median survival not reached) (Figure 1; 95 % CI, 38.9 – 49.8 months, $p = 0.0001$). The hazard ratio for death for patients with residual areas compared to individuals without was 3.701 (95% confidence interval: 1.92 to 7.13; $p = 0.0001$ by the log-rank test, two-sided).

Univariate analysis

To assess the association between patient, tumour and treatment characteristics with post-radiotherapy outcome, a univariate analysis was performed. The area under the ROC curve of a univariate model for each parameter was estimated. These results are summarized in Table 1.

The volume of the primary tumor ($GTV_{primary}$), maximum FDG uptake and OTT had the highest predictive power, while other commonly used predictors such as FEV_1, WHO-performance status or clinical stage showed a low predictive ability. $GTV_{primary}$ was significantly higher for patients with residual areas than for patients with a complete metabolic response (103 cm^3 ± 126.13 cm^3 vs. 48.3 cm^3 ± 55.5 cm^3, $p = 0.008$).

Table 1: Patient characteristics and their association with post-RT outcome in univariate analysis. Comparison of groups with residual disease and with complete metabolic response.

Variable	Residual disease (n = 56)	Complete metabolic response (n = 45)	*p**	AUC
Age, years				
Mean	65	65	0.907	0.54
SD	10.6	7.5		
Gender				
Female	21 (38)	19 (42)	0.480	0.54
Male	35 (62)	26 (58)		
Stage				
I	8 (14)	7 (16)	0.882	0.54
II	1 (2)	1 (2)		
IIIA	13 (23)	9 (20)		
IIIB	33 (59)	28 (62)		
IV	1 (2)			
Histology				
SCC	17 (30)	7 (16)	0.327	0.54
Adenocarcinoma	9 (16)	11 (24)		0.56
Large cell	20 (36)	17 (38)		0.54
NSCLC, NOS	10 (18)	10 (22)		0.47
FEV 1				
Mean	75.3	72.2	0.454	0.52
SD	16.8	22.5		
WHO-PS				
0	15 (27)	14 (31)	0.468	0.54
1	32 (57)	24 (53)		
≥ 2	9 (16)	7 (16)		
$GTV_{primary}$ (cm^3)				
Mean	103.0	48.3	0.008	0.62
SD	126.13	55.5		
GTV_{nodal} (cm^3)				
Mean	24.9	34.4	0.368	0.54
SD	37.3	66.5		
Tumor load (cm^3)				
Mean	127.8	82.2	0.047	0.60
SD	124.6	97.5		

*Abbreviations: **TTD** = Total tumor dose; **OTT** = Overall Treatment Time; **SUV_{max}** = Standardized Uptake Value; **$EQD_{2,T}$** = Equivalent radiation dose at 2 Gy corrected for time; **FEV_1** = Forced expiratory volume in 1 s; **SCC** = Squamous cell carcinoma; **NOS** = Not specified otherwise; **WHO-PS** = World Health Organization-performance status; **PLNS** = Positive lymph node stations.*

** Comparison between residual disease group vs. complete metabolic response group for variables. The Mann-Whitney U test was used for continuous variables and the Chi-square test for categorical variables.*

Similarly, the maximum FDG uptake on the pre-RT scan was significantly higher for patients with residual disease compared to patients with a complete metabolic response (10.5 ± 5 vs. 7.7 ± 5.2, $p = 0.007$). The overall treatment time (OTT) was longer for patients with a complete metabolic response in comparison with patients with residual disease (27 ± 6 days vs. 24 ± 5 days, $p = 0.013$).

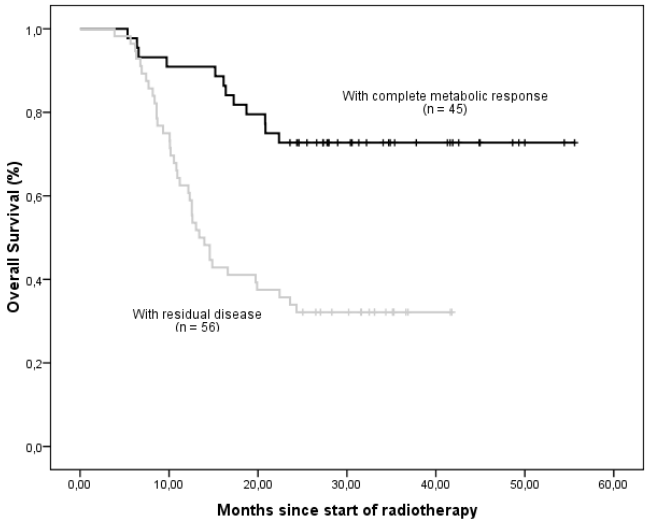


Figure 1

Kaplan-Meier estimates of overall survival of patients with residual metabolically active areas and with complete metabolic response on the post-radiotherapy PET-CT scan. Patients with residual metabolically active areas had significantly worse survival (p = 0.0001).

Kaplan-Meier survival curves for subgroups determined by the median for selected variables are shown in Figure 2.

Survival was significantly higher for patients with a tumor volume smaller than the median ($GTV_{primary}$= 46.6 cm^3) (p = 0.001). In patients with a SUV_{max} higher than the median (SUV_{max} = 8.4) in the pre-treatment scan, survival was significantly shorter, compared to patients with a SUV_{max} lower than the median (p = 0.040). Significant differences in survival were also observed for OTT, with a more prolonged survival for patients with a treatment time longer than the median of 25 days (p = 0.042). Survival differences in patients stratified according to TNM stage, were statistically not significant (p = 0.266). The same result was obtained for age. Older patients did not have different survival compared to younger patients (p = 0.998). Higher equivalent radiation dose was associated with better survival, however the difference was not statistically significant (p = 0.056).

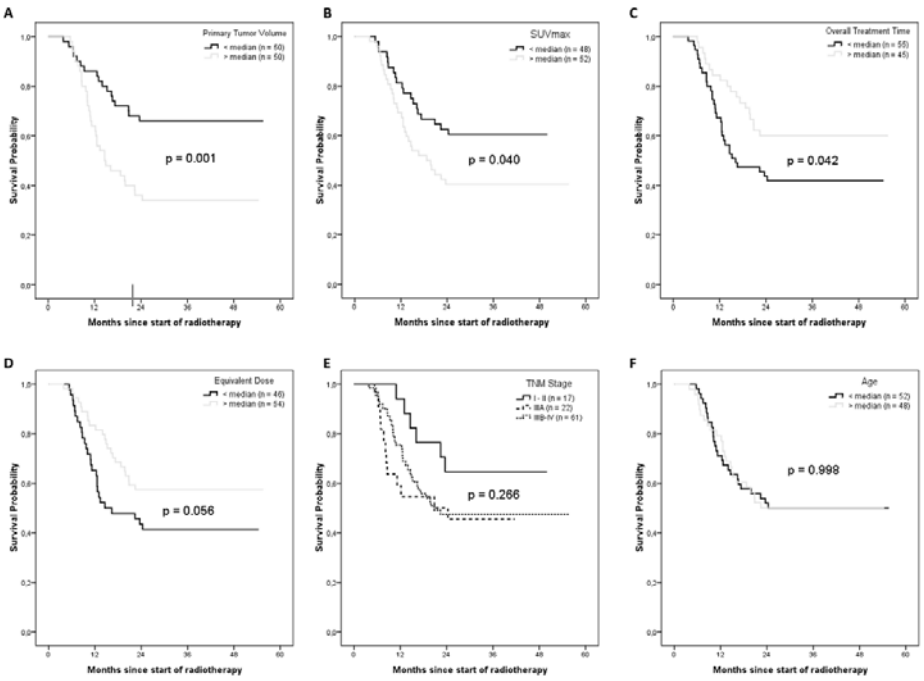


Figure 2

Survival among patients with advanced NSCLC for selected variables. For continuous variables, the cut-off value to stratify the patients was defined at the variable median. Shown are Kaplan-Meier curves for $GTV_{primary}$, SUV_{max}, OTT, $EQD_{2,T}$, TNM stage and age. In panel E, patients with stage I and II were grouped together due to the small number of cases. Stage IV (1 patient) was grouped with Stage IIIB.

Multivariate analysis

For the multivariate analysis, all the available variables were subjected to a combinatorial feature selection procedure. The combination with the highest AUC assessed by the leave-one-out cross validation approach was selected for the multivariate model. The variables included in the final multivariate p-SVM model were $GTV_{primary}$, maximum standardized FDG uptake, OTT and equivalent dose corrected for treatment time ($EQD_{2,T}$). Addition of other parameters to this model did not improve its performance. The area under the curve of the final predictive model was 0.71 (95% CI, 0.65-0.76; Fig 3). The variables included in the multivariate model showed also a significant association with the post-radiotherapy outcome in univariate analysis.

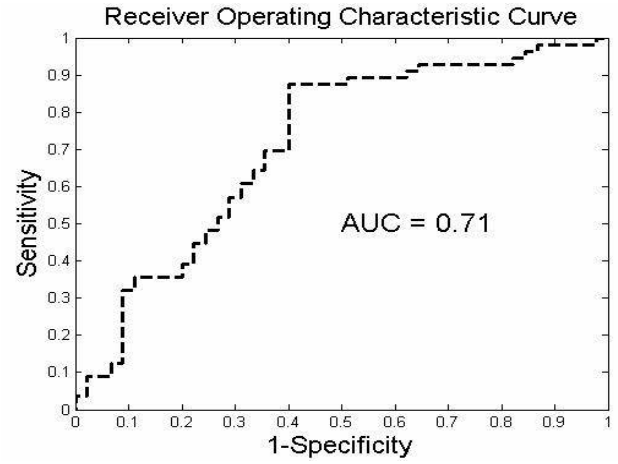


Figure 3

Area under the ROC curve assessed by the leave-one-out method for the multivariate model consisting on $GTV_{primary}$, SUV_{max}, OTT and $EQD_{2,T}$. A classifier with sensitivity of 1 and (1-specificity) of 0, point (0, 1) in graph, is ideal.

DISCUSSION

In this study we investigated the relationship of clinical parameters, including demographic, tumor and treatment characteristics, with metabolic response post-treatment. Our primary endpoint was defined as residual metabolic disease on a post-treatment PET-CT scan. Previous studies have shown that patients with residual metabolically active areas after treatment have a poorer prognosis compared with patients without [10, 15, 16]. In agreement with these studies, also our results showed that patients with residual disease had a significantly worse survival ($p = 0.0001$), compared to patients with a complete metabolic response, thus supporting the importance of our primary endpoint as surrogate for survival.

Previous studies examined the value of pre-treatment FDG-PET alone to determine treatment response after radiotherapy [17] and chemotherapy [18]. In our study, we explored not only the prognostic capability of FDG-PET but also the additional value of other clinico-pathological prognostic factors. Some of them, i.e., age, gender, tumor size, WHO performance status have been included in predictive models for survival in NSCLC patients [19-21]. In a retrospective study with a large patient population of NSCLC patients (stage I and II) which received resection with curative intent, Agarwal et al., reported that age and gender, tumor volume and type of surgery were important for the prediction of survival [22]. However, we did not find a significant association between age and metabolic response. Similarly, other studies have shown a relation between female gender and a fa-

vorable outcome [23]. We did not find a significant difference based on gender between responders and non-responders.

WHO performance status and FEV_1, have been cited as predictors of survival [19, 21], in which worse performance status and impaired lung function measurements are associated with shorter survival. We could not identify an association between these parameters and the post-treatment outcome. Although the tumor-node-metastasis (TNM) staging system is an important tool to estimate prognosis and choose the best treatment modality, several studies have reported that TNM has a poor predictive capability for survival in NSCLC patients [24]. In our cohort, the majority of patients were diagnosed with stage IIIA (22 %) and IIIB (61 %) disease. Therefore, stage was not a good predictor for residual disease, as differences in stage between the responding and the non-responding groups were not observed. Great interest has been given to the use of FDG-PET as a tool for tumor detection, staging and particularly for response assessment after radical radiotherapy or chemo-radiation [25, 26]. The maximum FDG uptake in the primary tumor measured on a pre-treatment scan has consistently been shown as an important prognostic factor for survival in NSCLC [15, 18, 25]. Our results showed that patients with residual metabolically active areas had a significantly higher FDG uptake on the pre-treatment scan, compared to patients with a complete metabolic response.

A high pre-treatment FDG uptake within the primary tumor was also significantly associated with worse survival (p = 0.040). Furthermore, the SUV_{max} showed a good predictive capability in univariate analysis.

Tumor volume also emerged as one of the most important predictors of residual disease. Our results are consistent with recently published studies, which have identified tumor size as an important prognostic factor of survival [27]. Here we confirmed the predictive capability of tumor size in assessment of metabolic response. This might indicate that specially for larger tumors, an effective dose could not be reached due to the dose constraints of the current protocol. The total tumor load ($GTV_{primary}$ + GTV_{nodal}) showed a strong association with the post-treatment outcome (Table 1). This association is due to the primary tumor volume, and perhaps enhanced by the addition of secondary volumes, however GTV_{nodal} alone did not show a predictive capability. A similar result was obtained for the number of positive lymph node stations on a pre-treatment PET-CT scan. Although the number of PLNSs is an important risk and staging factor for non-surgical patients [28], and has been included in multivariate models for survival in NSCLC, we did not find an added prognostic value for residual disease, perhaps because the outcome was defined in the primary tumor.

Despite an overall difference of two days, overall treatment time was significantly higher for patients with complete metabolic response in comparison with patients with residual disease. OTT was also significantly associated with the outcome in univariate analysis. It is generally accepted that a short treatment time should be chosen, to minimize the effect of accelerated repopulation [29]. The fact that we observe a longer treatment time in patients with a positive outcome is because those patients received a higher

dose. Higher total treatment dose has been associated with improved local tumor control and better survival [27, 30]. In the present study, the prescribed total dose was not different for patients with a complete response compared to patients with residual disease (p = 0.809).

Several predictive models of survival have been published for NSCLC patients, reporting different values of the area under the ROC as performance measurement, ranging from 0.65 to 0.86. These models were developed on populations that underwent different treatment modalities such as surgery [31], chemotherapy [28], radiotherapy or a combination [32] and consisted of patients with different tumor and patient characteristics. Thus, application of those models to different scenarios is still subject of research. Here we presented a multivariate model for prediction of residual disease. The final model consisted on tumor volume, overall treatment time, SUV_{max} and equivalent dose corrected for treatment time. This model yielded an AUC of 0.71 (95% CI, 0.65-0.76). This may have clinical relevance for patients identified at risk of treatment failure that may benefit from additional therapy. We were not able to analyze potential prognostic variables such as molecular markers or imaging surrogates [33-35] that may improve the ability of the presented model to predict the post- treatment failure. The lack of an external cohort to validate the presented model and confirm our results is an important limitation to our study. Our results may require validation according to the treatment modality to avoid possible confounding effects associated with multiple treatment modalities.

In conclusion, our results demonstrated that patients who do not respond to radiotherapy can be identified early in the course of their treatment. To our knowledge, this is the first study that examines different clinico-pathological predictors of residual disease. We identified important prognostic factors of residual disease and developed a multivariate model that identified patients at risk of treatment failure. Furthermore, we confirmed the validity of residual disease as a surrogate of survival. Our results could assist clinicians in the treatment decision-making process and in stratification of patients for clinical trials.

REFERENCES

1. Ferlay J, Parkin DM, Steliarova-Foucher E. Estimates of cancer incidence and mortality in Europe in 2008. Eur J Cancer; 46: 765-781.
2. Brawner EJ, Patrick Nana-Sinkam S, Jett JR. Lung cancer screening in 2008: A review and update. Respiratory Medicine CME 2008; 1: 2-9.
3. Scott WJ, Howington J, Feigenberg S, Movsas B, Pisters K. Treatment of non-small cell lung cancer stage I and stage II: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007; 132: 234S-242S.
4. Le Chevalier T, Arriagada R, Quoix E, Ruffie P, Martin M, Tarayre M, et al. Radiotherapy alone versus combined chemotherapy and radiotherapy in nonresectable non-small-cell lung cancer: first analysis of a randomized trial in 353 patients. J Natl Cancer Inst 1991; 83: 417-423.
5. Florin Sirzén EK, Sverre Sörenson, Eva Cavallin-ståhl. A Systematic Overview of Radiation Therapy Effects in Non-Small Cell Lung Cancer. Acta Oncologica 2003; 42: 493-515.

6. Zatloukal P, Petruzelka L, Zemanova M, Havel L, Janku F, Judas L, et al. Concurrent versus sequential chemoradiotherapy with cisplatin and vinorelbine in locally advanced non-small cell lung cancer: a randomized study. Lung Cancer 2004; 46: 87-98.
7. Aerts HJ, van Baardwijk AA, Petit SF, Offermann C, Loon J, Houben R, et al. Identification of residual metabolic-active areas within individual NSCLC tumours using a pre-radiotherapy (18)Fluorodeoxyglucose-PET-CT scan. Radiother Oncol 2009; 91: 386-392.
8. Aerts HJ, Bosmans G, van Baardwijk AA, Dekker AL, Oellers MC, Lambin P, et al. Stability of 18F-deoxyglucose uptake locations within tumor during radiotherapy for NSCLC: a prospective study. Int J Radiat Oncol Biol Phys 2008; 71: 1402-1407.
9. Weber WA, Figlin R. Monitoring cancer treatment with PET/CT: does it make a difference? J Nucl Med 2007; 48 Suppl 1: 36S-44S.
10. Mac Manus MP, Hicks RJ, Matthews JP, Wirth A, Rischin D, Ball DL. Metabolic (FDG-PET) response after radical radiotherapy/chemoradiotherapy for non-small cell lung cancer correlates with patterns of failure. Lung Cancer 2005; 49: 95-108.
11. Hoekstra CJ, Stroobants SG, Smit EF, Vansteenkiste J, van Tinteren H, Postmus PE, et al. Prognostic relevance of response evaluation using [18F]-2-fluoro-2-deoxy-D-glucose positron emission tomography in patients with locally advanced non-small-cell lung cancer. J Clin Oncol 2005; 23: 8362-8370.
12. van Baardwijk A, Wanders S, Boersma L, Borger J, Ollers M, Dingemans AM, et al. Mature results of an individualized radiation dose prescription study based on normal tissue constraints in stages I to III non-small-cell lung cancer. J Clin Oncol; 28: 1380-1386.
13. Steel G. Basic Clinical Radiobiology, 3rd ed. London: Hodder Arnold Publications; 2002.
14. Fung GM, Mangasarian OL. Multicategory Proximal Support Vector Machine Classifiers. Mach. Learn. 2005; 59: 77-97.
15. Decoster L, Schallier D, Everaert H, Nieboer K, Meysman M, Neyns B, et al. Complete metabolic tumour response, assessed by 18-fluorodeoxyglucose positron emission tomography (18FDG-PET), after induction chemotherapy predicts a favourable outcome in patients with locally advanced non-small cell lung cancer (NSCLC). Lung Cancer 2008; 62: 55-61.
16. Cerfolio RJ, Bryant AS, Winokur TS, Ohja B, Bartolucci AA. Repeat FDG-PET after neoadjuvant therapy is a predictor of pathologic response in patients with non-small cell lung cancer. Ann Thorac Surg 2004; 78: 1903-1909; discussion 1909.
17. Erdi YE, Macapinlac H, Rosenzweig KE, Humm JL, Larson SM, Erdi AK, et al. Use of PET to monitor the response of lung cancer to radiation treatment. Eur J Nucl Med 2000; 27: 861-866.
18. Vansteenkiste JF, Stroobants SG, De Leyn PR, Dupont PJ, Verbeken EK. Potential use of FDG-PET scan after induction chemotherapy in surgically staged IIIa-N2 non-small-cell lung cancer: a prospective pilot study. The Leuven Lung Cancer Group. Ann Oncol 1998; 9: 1193-1198.
19. Brundage MD, Davies D, Mackillop WJ. Prognostic factors in non-small cell lung cancer: a decade of progress. Chest 2002; 122: 1037-1057.
20. Dehing-Oberije C, De Ruysscher D, van der Weide H, Hochstenbag M, Bootsma G, Geraedts W, et al. Tumor volume combined with number of positive lymph node stations is a more important prognostic factor than TNM stage for survival of non-small-cell lung cancer patients treated with (chemo)radiotherapy. Int J Radiat Oncol Biol Phys 2008; 70: 1039-1044.
21. Solan MJ, Werner-Wasik M. Prognostic factors in non-small cell lung cancer. Semin Surg Oncol 2003; 21: 64-73.
22. Agarwal M, Brahmanday G, Chmielewski GW, Welsh RJ, Ravikrishnan KP. Age, tumor size, type of surgery, and gender predict survival in early stage (stage I and II) non-small cell lung cancer after surgical resection. Lung Cancer; 68: 398-402.
23. Fu JB, Kau TY, Severson RK, Kalemkerian GP. Lung cancer in women: analysis of the national Surveillance, Epidemiology, and End Results database. Chest 2005; 127: 768-777.
24. Ball D, Smith J, Wirth A, Mac Manus M. Failure of T stage to predict survival in patients with non-small-cell lung cancer treated by radiotherapy with or without concomitant chemotherapy. Int J Radiat Oncol Biol Phys 2002; 54: 1007-1013.

25. Hoang JK, Hoagland LF, Coleman RE, Coan AD, Herndon JE, 2nd, Patz EF, Jr. Prognostic value of fluorine-18 fluorodeoxyglucose positron emission tomography imaging in patients with advanced-stage non-small-cell lung carcinoma. J Clin Oncol 2008; 26: 1459-1464.
26. Forssell-Aronsson E, Kjellén E, Mattsson S, Hellström M, Swedish Cancer Society Investigation Group T. Medical Imaging for Improved Tumour Characterization, Delineation and Treatment Verification. Acta Oncologica 2002; 41: 604-614.
27. Kong FM, Ten Haken RK, Schipper MJ, Sullivan MA, Chen M, Lopez C, et al. High-dose radiation improved local tumor control and overall survival in patients with inoperable/unresectable non-small-cell lung cancer: long-term results of a radiation dose escalation study. Int J Radiat Oncol Biol Phys 2005; 63: 324-333.
28. 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 Biol Phys 2009; 74: 355-362.
29. Saunders M, Dische S, Barrett A, Harvey A, Griffiths G, Palmar M. Continuous, hyperfractionated, accelerated radiotherapy (CHART) versus conventional radiotherapy in non-small cell lung cancer: mature data from the randomised multicentre trial. CHART Steering committee. Radiother Oncol 1999; 52: 137-148.
30. Henning Willers FW, Henry Bünemann, Hans-Peter Heilmann. High-dose Radiation Therapy alone for Inoperable Non-small cell Lung Cancer: Experience with Prolonged Overall Treatment Times. Acta Oncologica 1998; 37: 101-105.
31. Mandrekar SJ, Schild SE, Hillman SL, Allen KL, Marks RS, Mailliard JA, et al. A prognostic model for advanced stage nonsmall cell lung cancer. Pooled analysis of North Central Cancer Treatment Group trials. Cancer 2006; 107: 781-792.
32. Blanchon F, Grivaux M, Asselain B, Lebas FX, Orlando JP, Piquet J, et al. 4-year mortality in patients with non-small-cell lung cancer: development and validation of a prognostic index. Lancet Oncol 2006; 7: 829-836.
33. Shepherd FA, Rosell R. Weighing tumor biology in treatment decisions for patients with non-small cell lung cancer. J Thorac Oncol 2007; 2 Suppl 2: S68-76.
34. Naqa IE, Deasy JO, Mu Y, Huang E, Hope AJ, Lindsay PE, et al. Datamining approaches for modeling tumor control probability. Acta Oncol.
35. Encan T, Hannisdal E. Blood Analyses as Prognostic Factors in Primary Lung Cancer. Acta Oncologica 1990; 29: 151-154.

CHAPTER 4

Development and validation of a nomogram for prediction of survival and local control in laryngeal carcinoma patients treated with radiotherapy alone: A cohort study based on 994 patients

Published in: Radiother Oncol. 2011; 100(1):108-15

Development and validation of a nomogram for prediction of survival and local control in laryngeal carcinoma patients treated with radiotherapy alone: a cohort study based on 994 patients.

Egelmeer AG, Rios Velazquez E*, de Jong JM, Oberije C, Geussens Y, Nuyts S, Kremer B, Rietveld D, Leemans CR, de Jong MC, Rasch C, Hoebers F, Homer J, Slevin N, West C, Lambin P.*

*These authors contributed equally to this work

ABSTRACT

Background

To advise laryngeal carcinoma patients on the most appropriate form of treatment, a tool to predict survival and local control is needed.

Materials and methods

We performed a population-based cohort study on 994 laryngeal carcinoma patients, treated with RT from 1977 until 2008. Two nomograms were developed and validated. Performance of the models is expressed as the Area Under the Curve (AUC).

Results

Unfavorable prognostic factors for overall survival were low hemoglobin level, male sex, high T-status, nodal involvement, older age, lower EQD (total radiation dose corrected for fraction dose and overall treatment time), and non-glottic tumor. All factors except tumor location were prognostic for local control. The AUCs were 0.73 for overall survival and 0.67 for local control. External validation of the survival model yielded AUCs of 0.68, 0.74, 0.76 and 0.71 for the Leuven (n = 109), the VU Amsterdam (n = 178), the Manchester (n = 403) and the NKI cohort (n = 205), respectively, while the validation procedure for the local control model resulted in AUCs of 0.70, 0.71, 0.72 and 0.62. The resulting nomograms were made available on the website www.predictcancer.org.

Conclusions

For patients with a laryngeal carcinoma treated with RT alone, we have developed visual, easy-to-use nomograms for the prediction of overall survival and primary local control. These models have been successfully validated in four external centers.

INTRODUCTION

In laryngeal carcinoma patients, treatment decisions are usually made by a multidisciplinary team based on guidelines. Patient- and tumor-related factors that are taken into consideration in this decision-making process are the TNM-stage, the functionality of the larynx, and the general condition of the patient (WHO performance status or Karnofski score) [1]. Though new developments are appearing in therapy, the primary treatment for early stage laryngeal carcinomas is radiotherapy (RT), laser surgery, or limited surgery. RT, open surgery (with or without postoperative radiotherapy), or a combination with systemic therapy are the current treatment options for more advanced cancers. Guidelines are used in the treatment decision process, and assessment of prognosis and preserved function are also taken into consideration.

Doctors often predict the prognosis fairly poorly [2–4], and so it is questionable whether this has an additional value. Besides the widely used predictors TNM-stage and general condition, other clinical factors are investigated for their prognostic and predictive value. An example of this is the pretreatment haemoglobin level. It is well established that patients with lower pretreatment hemoglobin levels have worse overall survival and local control than patients with normal hemoglobin levels [5–8].

Other prognostic factors that are investigated are sex and age [9– 12], with indications that women and younger patients have a better prognosis than men and elderly people.

Thus, to assist the doctor in deciding on the most appropriate treatment form, a tool to predict survival and local control is needed [13]. We, therefore, aimed to investigate which clinical and imaging factors are prognostic for the laryngeal carcinoma patients we have treated since 1977 with radiotherapy alone. We hypothesized that it is possible to develop nomograms for the prediction of survival and local control of laryngeal carcinoma patients treated with radiotherapy alone performing better than a nomogram based on TNM classification alone.

We tried to validate these models with four external datasets from the University Hospital of Leuven (Belgium), the VU University Medical Center of Amsterdam (The Netherlands), the Christie Hospital, Manchester (UK) and the Netherlands Cancer Institute-Antoni van Leeuwenhoek Hospital, Amsterdam (The Netherlands). These models will allow for improvement of the information given to patients about their prognosis. In the long term the models will allow for tailoring of the treatment to individual patients (e.g., for the choice surgeryradiotherapy), when combined with models predicting outcome after other therapies.

MATERIAL AND METHODS

Patient population

The patient and treatment characteristics of 1051 consecutively treated patients with a squamous cell laryngeal carcinoma were recorded in a database from January 1977 to December 2008. All patients were treated with radiotherapy alone at the MAASTRO

Clinic. Patients with a carcinoma in situ or distant metastasis at presentation (seven patients) were excluded from the study. Other exclusion criteria were treatment with Cobalt radiation (nine patients) and the use of chemotherapy (41 patients, three of whom had concurrent chemoradiation; 24 neoadjuvant chemotherapy; and 14 of whom were treated according to the ARCON-trial with carbogen and nicotinamide). A total of 994 patients were included in our cohort study. 528 (53.1%) of these patients had a T1 tumor, 264 (26.6%) a T2, 131 (13.2%) a T3, and 71 patients (7.1%) had a T4 tumor. Most of the patients (894, 89.9%) did not have positive lymph nodes. 45 patients (4.5%) had a N1 status, 42 (4.2%) a N2, and eleven (1.1%) patients had a N3 status.

The trial is registered on ClinicalTrials.gov with registration number 2263.

Diagnosis and staging were always undertaken according to the Dutch guidelines, including endoscopia under anesthesia and biopsy of the tumor. Also recommended in the latest Dutch guidelines are a computed tomography (CT) scan of the head and neck, a chest X-ray, ultrasonography of the neck (if necessary with punction), and blood tests.

Radiotherapy treatment

All patients were treated at the MAASTRO Clinic with a continuous course of radiotherapy delivered by a 4–6 MV linear accelerator after either a traditional simulation (patients before 1996) or a CT simulation (patients treated from 1996 onwards). During simulation and treatment, all patients were immobilized by a thermoplastic mask. Sixteen patients received a palliative radiation dose of less than 60 Gy. Patients were treated in line with the state-ofthe- art practices. T1–2 glottic tumors and T1 supraglottic tumors were treated with 60–66 Gy in fractions of 2–2.40 Gy, and other tumors were treated with 70 Gy over seven weeks in daily fractions of 2 Gy, and after 2000 with 68 Gy, the first 23 fractions 2 Gy daily, and the last 11 fractions twice daily in fractions of 2 Gy.

To correct for differences in radiation scheme, the biological equivalent dose in fractions of 2 Gy and corrected for overall treatment time was calculated, using the following formula:

$$EQD_{2T} = D \times ((d + \alpha/\beta) / (2 + \alpha/\beta)) - \gamma (T - Tk)$$

D is the total radiation dose, *d* is the fraction dose, α/β is 10 Gy, T is the overall treatment time, accelerated repopulation kick-off time (*Tk*) is 28 days, and loss in dose due to repopulation (*γ*) is 0.6 Gy/d [14]. As EQD_{2T} is not easily calculated in daily clinical practice, EQD_{2T}

values for the most common radiation schemes are given in Table 1. If the overall treatment time for a patient differs from the anticipated value, it is possible to recalculate EQD_{2T} after completing the treatment and thus to obtain an adjusted prediction.
The follow-up for all patients consisted of regular visits to the head and neck oncology department over five years after the curative radiotherapy treatment. These visits took place every second month for the first six months, then every third month for two years, every fourth month during the third year, then twice yearly until the end of follow-up (five years). At every visit, the medical history was taken and physical examination carried out. Thereafter, information was gathered from the general practitioner and the Dutch Registry of Births, Deaths, and Marriages (''Gemeentelijke Basis Administratie'', or ''GBA'').

Table 1: **EQD_{2T} of the most common radiotherapy schemes**

Radiotherapy scheme	EQD_{2T}
60 Gy in fractions of 2 Gy in 6 weeks	60 x ((2+10) / (2+10)) - 0.6 (40-28) = 52.8 Gy
66 Gy in fraction of 2 Gy in in 6.5 weeks	66 x ((2+10) / (2+10)) - 0.6 (45-28) = 55.8 Gy
68 Gy, 23 fractions of 2 Gy once daily, and 11 fractions of 2 Gy twice daily, in 6 weeks	68x ((2+10) / (2+10)) - 0.6 (39-28) = 61.4 Gy
70 Gy in fractions of 2 Gy in 7 weeks	70 x ((2+10) / (2+10)) - 0.6 (47-28) = 58.6 Gy
55 Gy in fractions of 2.2 Gy in 5 weeks	55 x ((2.2+10)/ (2+10)) - 0.6 (33-28) = 53.9 Gy

PET-CT analysis

Features of the PET–CT scans were analyzed for a subgroup of patients. ^{18F}FDG-PET-scans were available since 2004, and 115 of these scans were available and assessable. Contoured tumor volumes were available for 124 patients, mostly patients with a T3 or T4 tumor. The gross tumor volume (GTV) was measured as contoured in our radiotherapy treatment planning system by a radiation oncologist (Computerized Medical Systems, INC, St. Louis, MO). Several features were extracted from the pretreatment PETscans. A circle was drawn around the region of interest c.q. the larynx.
Within this region of interest the maximal Standard Uptake Value (SUV_{max}) is given. Dedicated software (TrueD; Siemens Medical, Erlangen, Germany) was used to calculate SUVmax per patient. Furthermore, the SUV of the deltoid muscle was calculated. The
SUVmax of the tumor and the SUV of the background was used to calculate a source-to-background ratio according to the following formula:

$$78.13 \times (SUV_{tumor} / SUV_{background})^{(-0.2988)}$$

The output of the formula is expressed as a percentage, which was used as the contouring percentage to determine the metabolic volume, i.e., the volume of the tumor that has higher FDG uptake than the contouring percentage of the SUV_{max}.

Statistical analysis

The prognostic factors tested were age at start of radiotherapy, sex, tumor location (glottic or non-glottic), pretreatment hemoglobin level, EQD_{2T}, T-stage, and N-stage (N0 or N+). Age, hemoglobin level, and EQD_{2T} were analyzed as continuous values. The endpoints of the study were overall survival and local control, both calculated from the start of radiotherapy. Patients were followed for at least 1 month up to a maximum of 72 months. Failure of local control was defined as persistent or recurrent local disease after the start of radiotherapy (i.e., the first relapse after therapy).

The Kaplan–Meier method was used for univariate survival analysis. For overall survival, data were considered right-censored if patients were still alive at the time of evaluation. For local control, data were considered right-censored if patients did not have recurrent local disease at the time of evaluation. Groups were compared using the log rank test. The Cox proportional hazards model was applied to perform a multivariate analysis. The proportional hazards assumption was tested by adding time-dependent covariates to the model. In addition, linearity of the variables was assessed. Missing values were imputed using predictive mean matching. A stepwise backward method was used to select a relevant set of variables ($p < 0.2$). Hazard ratios and 95% confidence intervals were reported. Performance of the models was expressed as the C-statistic (Harrell's C), which is comparable to the Area Under the Curve (AUC). The maximum value of the C-statistic is 1.0; indicating a perfect prediction model. A value of 0.5 indicates that 50% of the patients are correctly classified (i.e., as good as chance). Bootstrapping techniques were used to validate the models; that is, to adjust the estimated model performance for overoptimism or overfitting. The results of the multivariate analysis were used to develop a nomogram. These nomograms will be, after publication, publicly available on the website www.predictcancer.org.

The MAASTRO cohort was split into four subgroups according to quartiles of the risk score. To assess for differences in survival of the subgroups, Kaplan–Meier curves were made. In addition, the performance of the multivariate model was assessed using four external validation sets [15]. Analyses were performed using SPPS for Windows (version 17.0; SPSS Inc., Chicago) and Matlab 7.11.0 (The MathWorks Inc., Natick, MA, USA).

Validation datasets

Patient characteristics of the validation cohorts are shown in Table 2.

The validation cohort of Leuven consisted of 109 laryngeal carcinoma patients, treated with radiotherapy alone between March 2000 and January 2006. 45 of these patients (40.9%) had a T1 tumor, and 83 patients (75.5%) did not have nodal involvement.

None of the patients received chemotherapy. Two thirds of the patients received 2 Gy fraction until a total dose of 66–72 Gy. Most other patients were treated with 25 fraction of 2.2 Gy, to reach a total dose of 55 Gy.

Table 2: Patient characteristics

	MAASTRO Cohort	LEUVEN Cohort	VU AMSTERDAM Cohort	NKI /AVL AMSTERDAM Cohort	MANCHESTER Cohort
	Number (%)	Number (%)	Number (%)	Number (%)	Number (%)
Age					
18-60 years	360 (36.2)	40 (36.7)	62 (34.8)	75 (36.5)	154 (38.2)
>60 years	634 (63.8)	69 (63.3)	116(65.2)	130 (63.5)	249 (61.8)
Gender					
Male	883 (88.8)	99 (90.8)	154 (86.5)	162 (79.1)	357 (88.6)
Female	111 (11.2)	10 (9.2)	24 (13.5)	43 (20.9)	46 (11.4)
T-classification					
T1	528 (53.1)	45 (41.3)	67 (37.6)	86 (41.9)	252 (62.5)
T2	264 (26.6)	30 (27.5)	91 (51.1)	119 (58.1)	124 (30.8)
T3	131 (13.2)	24 (22.0)	16 (9.0)	0 (0.0)	27(6.7)
T4	71 (7.1)	10 (9.2)	4 (2.2)	0 (0.0)	0 (0.0)
N-classification					
N0	894 (89.9)	82 (75.2)	165 (92.7)	184 (89.8)	398 (98.8)
N1	45 (4.5)	6 (5.5)	5 (2.8)	6 (2.9)	1 (0.2)
N2	42 (4.2)	18 (16.5)	6 (3.4)	11 (5.4)	3 (0.7)
N3	11 (1.1)	3 (2.8)	0 (0)	4 (1.9)	1 (0.2)
Missing	2 (0.2)	0 (0)	2 (1.1)	0 (0.0)	0 (0.0)
Location tumor					
Glottic	729 (73.3)	64 (58.7)	127 (71.3)	149 (72.7)	403 (100.0)
Supraglottic	245 (24.6)	39 (35.8)	43 (24.2)	56 (27.3)	0 (0.0)
Subglottic	13 (1.3)		2 (1.1)		
Transglottic	7 (0.7)		6 (3.4)		
Other		6 (5.5)			0 (0.0)
Hemoglobin level					
Low [a]	168 (16.9)	20 (18.3)	44 (24.7)	35 (17.1)	90 (22.3)
Normal-High	667 (67.1)	46 (42.2)	123 (69.1)	145 (70.8)	255 (63.3)
Missing	159 (16.0)	43 (39.4)	11 (6.2)	25 (12.1)	58 (14.4)
Total radiation dose					
< 60 Gy	16 (1.6)	45 (41.3)	1 (0.6)	2 (0.9)	402 (99.8)
60-66 Gy	437 (44.0)	12 (11.0)	69 (38.8)	94 (45.9)	1 (0.2)
> 66 Gy	541 (54.4)	52 (47.7)	108 (60.7)	109 (53.2)	0 (0.0)
Fraction dose					
1.6 -2.0	677 (68.1)	65 (59.6)	116 (65.2)	0 (0.0)	1 (0.2)
> 2.0	317 (31.9)	44 (40.4)	62 (34.8)	205 (100)	402 (99.8)
Overall treatment time					
< 40 days	321 (32.3)	43 (39.4)	164 (92.1)	162 (79.1)	401 (99.5)
40 - 50 days	595 (59.9)	41 (37.6)	11 (6.2)	38 (18.5)	2 (0.5)
> 50 days	78 (7.8)	25 (22.9)	3 (1.7)	5 (2.4)	0 (0.0)

[a] *male <8.5 mmol/L, female <7.5 mmol/L*

The VU Amsterdam cohort consists of 178 patients, which were treated between December 2001 and January 2007. 67 (37.6%) had a T1 tumor and 165 patients (92.7%) were N0. 97 patients were treated with two lateral fields and 19 patients with an IMRT technique. All patients were treated with radiotherapy alone, with most patients treated with 2 Gy fraction, until 68–70 Gy or 60 Gy in 2.5 Gy fractions.

The NKI/AVL Amsterdam cohort consisted of 205 patients with early larynx cancer (T1 tumors in 42% of cases, T2 tumors in 58% of cases) treated with primary radiation treatment between March of 2000 and July 2008. 184 patients (89.8%) were N0 at presentation. Patients with T1N0 glottic cancer were treated with 2 lateral opposing beams to the larynx only. The standard fractionation scheme for these patients was 25 fractions of 2.4 Gy to a total dose of 60 Gy in 5 weeks. Patients with T2 glottic cancers or with supraglottic cancer received radiation treatment to the larynx including prophylactic neck irradiation. The fractionation schedule for this latter group was 35 fractions of 2 Gy to a total dose of 70 Gy in 6 weeks, according to the DAHANCA schedule [16]. None of the patients received chemotherapy.

The Manchester cohort consists of 403 patients, which were treated between January 1998 and January 2005. All these patients had a 68lottis tumor and most of these tumors were small (252 (62.05%) T1 tumors and 124 (30.8%) T2 tumors). All but four patients
(1.2%) were N0. 189 patients received radiation through two lateral opposing fields and 240 patients were treated with an anterior oblique technique. The majority of patients with treated with 50–55 Gy, in 16 fractions (3.1–3.4 Gy per fraction).

RESULTS

MAASTRO cohort

Analyses were carried out for 994 patients of the MAASTRO cohort. The majority of the patients were male (88.8%) and the median age at the start of radiotherapy was 65.0 years (range 31– 91 years). Pretreatment hemoglobin level was available for 835 patients; this value was missing in 16.0% of cases. For more patient information, see Table 2. At the time of analysis, 476 patients were alive (47.9%). Median follow-up for surviving patients was 72 months (range 2– 72 months). Two-year overall survival was 82.8%, and five-year overall survival was 67.7%. Primary local control was 71.0% at two years and 54.0% at five years after the start of treatment. A total of 220 local failures were observed. Most of these (179/220, 81.4%) occurred within the first two years after treatment, and no local failures were recorded after five years as they are assumed to be second primary tumors.

Prognostic factors

Univariate Cox regression was carried out on the following factors: age at the start of radiotherapy, sex, tumor location, pretreatment hemoglobin level, EQD_{2T}, T-stage, and N-stage (N0 or N+). All factors were statistically significant for overall survival ($p < 0.05$). In the multivariate analysis, independent unfavourable prognostic factors for overall survival were low hemoglobin level, male sex, high T-status, presence of nodal involvement, older age, lower EQD_{2T}, and non-glottic tumor. See Table 3 for the hazard ratios, confidence intervals and p-values.

Table 3: Multivariate Cox proportional hazards analysis of potential prognostic factors

	Overall survival			Local control		
	HR	95% CI	*p*	HR	95% CI	*p*
Age	1.04	1.03-1.05	<0.0001	1.02	1.01-1.03	0.0012
Gender			0.0002			<0.0001
Female	1.00			1.00		
Male	2.30	1.49-3.55		2.47	1.69-3.60	
T-stage			<0.0001			<0.0001
T1	1.00			1.00		
T2	1.22	0.91-1.63		1.52	1.20-1.92	
T3	2.22	1.56-3.14		2.48	1.87-3.28	
T4	4.29	2.85-6.47		4.28	3.05-6.02	
N-stage			0.034			0.0059
N0	1.00			1.00		
N-positive	1.46	1.03-2.06		1.51	1.13-2.03	
Location tumor			0.0725			
Glottic	1.00			1.00		
Non-glottic	1.31	0.98-1.75		-		0.93
Hemoglobin level	0.67		<0.0001	0.75	0.67-0.85	<0.0001
EQD_{2T}	0.97	0.94-0.99	0.0037	0.97	0.95-0.99	0.0011

[a] ***LLN*** *Lower limit of normal.* ***HR*** *Hazard ratio.* ***CI*** *Confidence intervals*

The year of therapy had no prognostic significance for either survival ($p = 0.28$) or local control ($p = 0.48$). In order to test the hypothesis that modern radiotherapy (3D, in vivo dosimetry) would perform better than 2D radiotherapy, the cohort was split before (n = 532) and after January 1996 (n = 462). In the univariate analysis there was no difference between the two groups ($p = 0.14$), but in the Cox regression analysis there was a trend ($p = 0.077$) with a hazard ratio of 1.3 in favor of modern radiotherapy.

The clinical factors investigated for local control were the same as for overall survival. Tumor location (i.e., glottic vs non-glottic) was significant in the univariate analysis ($p < 0.001$), but not in the multivariate analysis ($p = 0.93$). All other factors were significant both in the univariate and in the multivariate analysis. Unfavorable prognostic factors for local control were low haemoglobin level, male sex, high T-status, nodal involvement, old-

er age, and lower EQD_{2T}. See Table 3 for the hazard ratios, confidence intervals and p-values.

PET – CT scans

Tumor volume was measured in only 124 patients. The GTV ranged between 0.0 and 128.2 cc with a median of 4.7 cc. In a subgroup analysis with these patients, the volume was a statistically significant predictor for overall survival ($p < 0.001$) and local control ($p < 0.001$). In the Cox regression analysis, we tested the prognostic value of tumor volume, sex, and N-status. None of these factors was statistically significant for either overall survival, or for local control. This subgroup is possibly too small to detect influences on overall survival or local control.

One-hundred and fifteen PET-scans were assessable and evaluable. SUV_{max} ranged between 1.9 and 23.8, with a median of 6.1. The metabolic volume ranged between 1.1 and 73.3 cc, with a median of 7.9 cc. The SUV_{max} and metabolic volume were not statistically significant for survival ($p = 0.093$ and $p = 0.93$, respectively), but SUV_{max} was a statistically significant predictor for local control ($p = 0.019$, metabolic volume: $p = 0.70$). In the Cox Regression analyses, the SUVmax lost significance, when corrected for T-status, N-status and GTV. GTV, T-status and the SUV_{max} are highly correlated.

Nomograms

Table 4: Nomograms performance in external datasets.

	Survival		Local control	
	Model based on multiple variables	**Model based on TNM**	**Model based on multiple variables**	**Model based on TNM**
MAASTRO (n=994)	0.73 (0.70- 0.77)	0.62 (0.58-0.63)	0.67 (0.64-0.71)	0.62 (0.55-0.63)
LEUVEN (n=109)	0.68 (0.50-0.82)	0.70* (0.45- 0.81)	0.70 (0.50-0.78)	0.62 (0.49-0.72)
VU AMSTERDAM (n=178)	0.74 (0.69-0.87)	0.65 (0.57- 0.75)	0.71 (0.66-0.81)	0.64 (0.57-0.74)
NKI/AVL AMSTERDAM (n=205)	0.71 (0.60-0.82)	0.57 (0.52- 0.69)	0.62 (0.55-0.75)	0.56 (0.49-0.63)
MANCHESTER (n=403)	0.76 (0.72-0.81)	0.63 (0.58- 0.69)	0.72 (0.67-0.78)	0.63 (0.58-0.69)
Pooled External Datasets	0.71 (0.70-0.76)	0.60 (0.57-0.62)	0.65 (0.62-0.68)	0.60 (0.58-0.61)

** 95% Confidence intervals (shown between brackets) were obtained in a bootstrap procedure (n = 1000). The obtained AUCs were significantly different for the multivariate model compared with the TNM based model for every cohort (p=0.001), except for survival prediction in the Leuven cohort where no significant differences were found.*

For the purpose of comparison, we analyzed the predictive value of the TNM-stage alone. The AUC of the model for overall survival was 0.62, which means that the model predicted overall survival correctly in only 62% of patients. The AUC of the model for local control was 0.62 too.

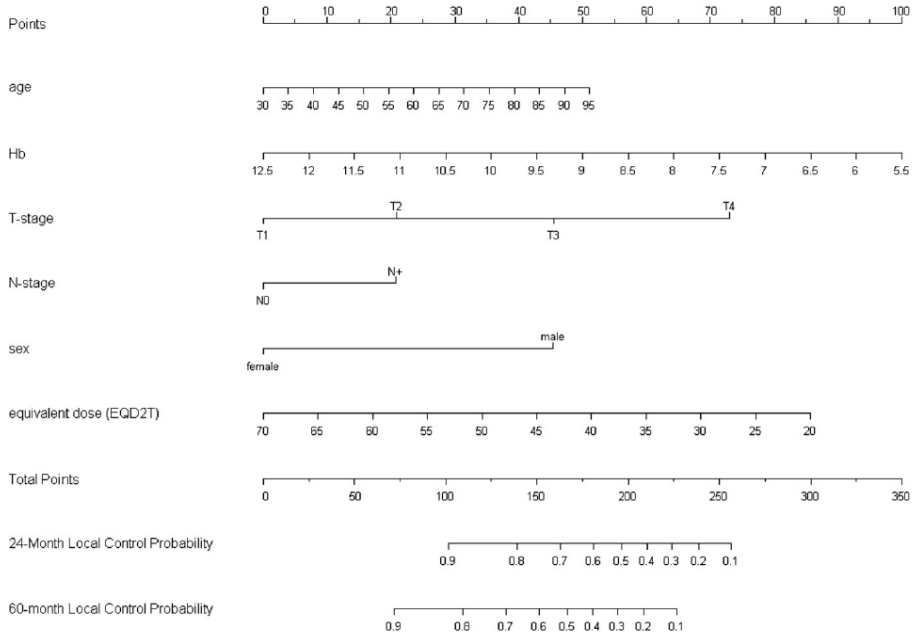


Figure 1

Nomogram for the prediction of 2-year and 5-year local control.

The MAASTRO cohort can be split into four subgroups, according to the quartiles of the risk score. The prognoses of the highand low-risk groups are distinctive, but the other two groups have overlapping 95% confidence intervals. They are, therefore, merged into one patient group with a medium risk score.

The two- and five-year survival rates were 82.1% (95% CI, 76.8–87.4%) and 76.3 (95% CI, 70.4–82.2%) for the low-risk group, 72.1% (95% CI, 67.8–76.4%) and 53.3% (95% CI, 48.6–58.4%) for the medium-risk group, and 47.3% (95% CI, 40.4–54.2%) and 28.3% (95% CI, 21.8–34.8%) for the high-risk group, respectively ($p < 0.001$). See Fig. 3 for the Kaplan–Meier curve. We validated these models with the datasets of Leuven, VU and NKI Amsterdam and Manchester. Validation of the survival model yielded AUCs of 0.68, 0.74, 0.71 and 0.76, while the validation procedure for the local control model resulted in AUCs of 0.70, 0.71, 0.62 and 0.72, respectively.

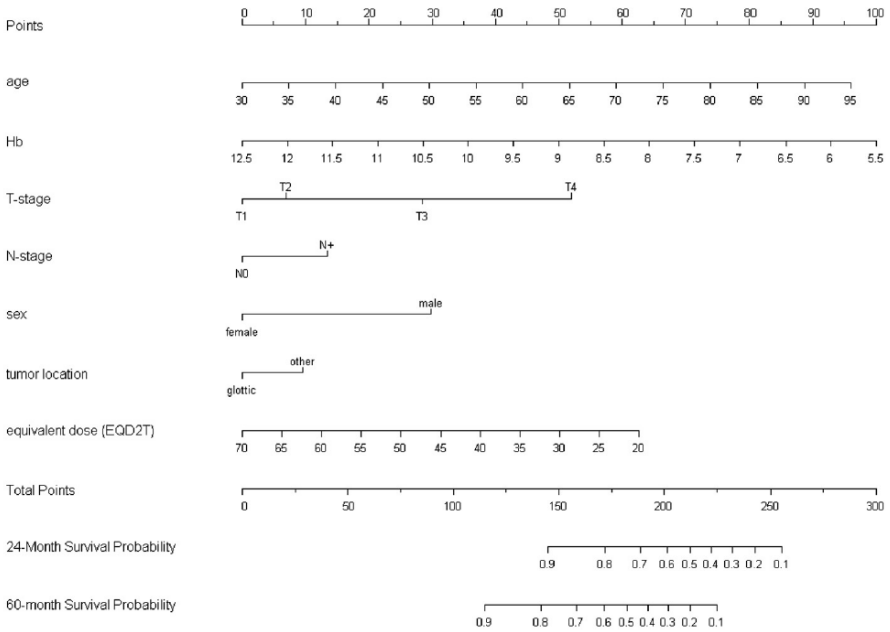


Figure 2

Nomogram for the prediction of 2-year and 5-year overall survival.

We validated the TNM model on the other datasets. The AUCs of the survival TNM model were 0.70 for the Leuven database, 0.65 for the VU Amsterdam database, 0.57 for the NKI database and 0.63 for the Manchester database.

The local control TNM model yielded AUCs of 0.62, 0.64, 0.56 and 0.63, respectively. Confidence intervals are shown in Table 4.

DISCUSSION

We have developed visual, ready-to-use nomograms for overall survival and primary local control in laryngeal carcinoma patients to predict outcome following radiotherapy alone. We did so after a multivariate analysis of several easily assessable clinical factors in a large group of unselected laryngeal cancer patients. The survival rates in this study are comparable with other studies and with those in the Dutch Cancer Registration (Nederlandse Kankerregistratie) [17], which observed a two- and five-year overall survival of 81% and 69%, respectively. The models we developed for both survival and local control yielded similar results in three other patient populations in hospitals in Leuven (Belgium), Amsterdam (The Netherlands), and Manchester (UK).

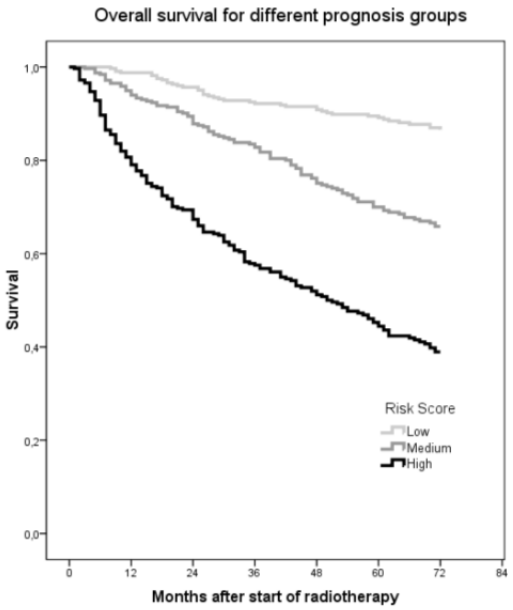


Figure 3

Kaplan Meier curve for overall survival according to the different prognosis groups.

The models perform better than models based on TNM staging alone. Johansen et al. [18] performed a multivariate analysis on 1127 patients with a laryngeal carcinoma treated with curative radiotherapy alone. This group of patients is comparable to ours with respect to age, sex, TNM-stage and tumor location. As in our patient population, they found sex, T-stage, N-stage, and hemoglobin level to be statistically significant for overall survival. They also found differentiation grade to be significant for survival, but this value was not available for our patients. And unlike our data, the haemoglobin level was not prognostic for local control.

In 2001 a study was published about a prediction model for head and neck cancer patients [19]. This model was based on 1662 head and neck cancer patients, using age, sex, cTNM-status, tumor (sub)site, and history of tumor as predictors for survival.

However, because the performance of the model is not indicated, its accuracy is unknown. Other studies with smaller groups of patients found correlations between outcome and T-site [20], low pretreatment hemoglobin level [5–7], and tumor volume [20 – 22]. Along with clinical factors, imaging features [23–25], serum markers [26,27] and other biomarkers [28–30] also have a predictive value for laryngeal carcinoma.

Several studies have been published with different findings about the predictive value of the PET–CT scan in head and neck cancer [31,32]. In contrast to earlier studies, our population consists exclusively of patients with a laryngeal carcinoma. Even though there are similarities between the different tumors in the head and neck region, there are

also a number of differences. Therefore, it is not possible to perform prognostic and predictive studies without adequate stratification for location of the tumor.
In the group of patients in this study, GTV was a significant prognostic factor for survival in the univariate analysis. The prognostic factors for local control in the univariate analysis were SUV_{max} and GTV.

This study is an observational – ''population-based'' cohort study, which included all laryngeal carcinoma patients treated with radiotherapy alone at our hospital. There is a potential selection bias, because treatment choice was made before inclusion in the study. During the inclusion period there were changes in diagnostics, staging methods, and treatment choice, which imply that there might be stage migration. Although treatment quality and control have improved in recent decades, we observed no significant improvement of overall survival over the course of this study.
A possible explanation for this is that co-morbidity in laryngeal carcinoma patients is high and influences overall survival more than the cancer itself. Data on WHO performance status or Karnofski score were missing in our population, thus we were not able to incorporate the effect of co-morbidity on the predictive nomogram which may limit its predictive performance. Although co-morbidity would certainly influence the estimation of overall survival, our estimation based on the predictive nomogram yielded similar AUCs on the external datasets and was significantly superior to TNM for overall survival prediction. Co-morbidity might be an important predictor and should be investigated in future studies. Patients treated with chemotherapy were excluded from this analysis, as they constituted only a small group of patients. Most of them received chemotherapy as a palliative treatment or within a study protocol. A recent meta-analysis demonstrates the benefit of concurrent chemotherapy [33–36], with an absolute benefit of 6.5% at five years for head and neck cancer patients. In subgroup analyses this effect seems stronger for younger patients (<60 years old) and patients with good performance status and locally advanced disease.

To allow for adequate decision making regarding treatment for laryngeal cancer patients, we are currently analyzing our data on patients with a laryngeal carcinoma treated with laser surgery or surgery, and patients treated with surgery and postoperative radiotherapy. By making nomograms for these patients too, we can create a useful tool for the treatment decision-making process.
A model consisting of solely clinical features is still too limited to allow clinical decision making. There is a need for adding biological and imaging data [29]. There is a need for a prospective multicentric randomized trial, preferably with banking of tissues, to validate and extend the results. A prognostic study of that kind would make it possible to collect data on biomarkers and imaging features, along with clinical data. Normal tissue reactions should then also be taken into account [37–39]. The availability of several validated nomograms for the different therapeutic options for laryngeal carcinoma consisting of clinical, biological, and imaging features will make a potent decision support model. The risk

groups illustrated in Fig. 3 can also be used for stratification in clinical trials or to customize more aggressive strategies to the risk of relapse.

CONCLUSIONS

We have built visual, ready-to-use nomograms for the prediction of survival and primary local control with several easy assessable clinical factors, for use on patients with laryngeal carcinoma treated with radiotherapy alone. The performance of these nomograms is significantly better than the predictive value of the TNMclassification alone, but still need additional data before being used in clinical practice.

REFERENCES

1. van der Schroeff MP, Baatenburg de Jong RJ. Staging and prognosis in head and neck cancer. Oral Oncol 2009;45:356–60.
2. Kellett J. Prognostication – the lost skill of medicine. Eur J Intern Med 2008;19: 155–64.
3. Chow E, Davis L, Panzarella T, et al. Accuracy of survival prediction by palliative radiation oncologists. Int J Radiat Oncol Biol Phys 2005;61:870–3.
4. Stockler MR, Tattersall MH, Boyer MJ, Clarke SJ, Beale PJ, Simes RJ. Disarming the guarded prognosis: predicting survival in newly referred patients with incurable cancer. Br J Cancer 2006;94:208–12.
5. Cho EI, Sasaki CT, Haffty BG. Prognostic significance of pretreatment hemoglobin for local control and overall survival in T1–T2N0 larynx cancer treated with external beam radiotherapy. Int J Radiat Oncol Biol Phys 2004;58:1135–40.
6. Lee WR, Berkey B, Marcial V, et al. Anemia is associated with decreased survival and increased locoregional failure in patients with locally advanced head and neck carcinoma: a secondary analysis of RTOG 85-27. Int J Radiat Oncol Biol Phys 1998;42:1069–75.
7. Prosnitz RG, Yao B, Farrell CL, Clough R, Brizel DM. Pretreatment anemia is correlated with the reduced effectiveness of radiation and concurrent chemotherapy in advanced head and neck cancer. Int J Radiat Oncol Biol Phys 2005;61:1087–95.
8. Hoff CM, Hansen HS, Overgaard M, et al. The importance of haemoglobin level and effect of transfusion in HNSCC patients treated with radiotherapy – results from the randomized DAHANCA 5 study. Radiother Oncol 2011;98:28–33.4.
9. Plataniotis GA, Theofanopoulou ME, Kalogera-Fountzila A, et al. Prognostic impact of tumor volumetry in patients with locally advanced head-and-neck carcinoma (non-nasopharyngeal) treated by radiotherapy alone or combined radiochemotherapy in a randomized trial. Int J Radiat Oncol Biol Phys 2004;59:1018–26.
10. Jeremic B, Milicic B. Pretreatment prognostic factors of survival in patients with locally advanced nonmetastatic squamous cell carcinoma of the head and neck treated with radiation therapy with or without concurrent chemotherapy. Am J Clin Oncol 2009;32:163–8.
11. Horiot JC, Bontemps P, Van den Bogaert W, et al. Accelerated fractionation (AF) compared to conventional fractionation (CF) improves loco-regional control in the radiotherapy of advanced head and neck cancers: results of the EORTC 22851 randomized trial. Radiother Oncol 1997;44:111–21.
12. Trinkaus M, Corry J, Rischin D. Comparison of self-reported smoking status and physician-recorded smoking status among patients with locally advanced squamous cell carcinoma of the head and neck (SCCHN). Radiother Oncol 2011;98:143–4.

13. Lambin P, Petit SF, Aerts HJ, et al. The ESTRO Breur Lecture 2009. From population to voxel-based radiotherapy: exploiting intra-tumour and intraorgan heterogeneity for advanced treatment of non-small cell lung cancer. Radiother Oncol 2010;96:145–52.
14. Fowler JF, Tome WA, Fenwick JD, Mehta MP. A challenge to traditional radiation oncology. Int J Radiat Oncol Biol Phys 2004;60:1241–56.
15. Dehing-Oberije C, Yu S, De Ruysscher D, 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 Biol Phys 2009;74:355–62.
16. Overgaard J, Hansen HS, Specht L, et al. Five compared with six fractions per week of conventional radiotherapy of squamous-cell carcinoma of head and neck: DAHANCA 6 and 7 randomised controlled trial. Lancet 2003;362: 933–40.
17. www.ikcnet.nl.
18. Johansen LV, Grau C, Overgaard J. Laryngeal carcinoma – multivariate analysis of prognostic factors in 1252 consecutive patients treated with primary radiotherapy. Acta Oncol 2003;42:771–8.
19. Baatenburg de Jong RJ, Hermans J, Molenaar J, Briaire JJ, le Cessie S. Prediction of survival in patients with head and neck cancer. Head Neck 2001;23:718–24.
20. Mendenhall WM, Morris CG, Amdur RJ, Hinerman RW, Mancuso AA. Parameters that predict local control after definitive radiotherapy for squamous cell carcinoma of the head and neck. Head Neck 2003;25:535–42.
21. van den Broek GB, Rasch CR, Pameijer FA, et al. Pretreatment probability model for predicting outcome after intraarterial chemoradiation for advanced head and neck carcinoma. Cancer 2004;101:1809–17.
22. Le Tourneau C, Velten M, Jung GM, Bronner G, Flesch H, Borel C. Prognostic indicators for survival in head and neck squamous cell carcinomas: analysis of a series of 621 cases. Head Neck 2005;27:801–8.
23. Machtay M, Natwa M, Andrel J, et al. Pretreatment FDG-PET standardized uptake value as a prognostic factor for outcome in head and neck cancer. Head Neck 2009;31:195–201.
24. Allal AS, Slosman DO, Kebdani T, Allaoua M, Lehmann W, Dulguerov P. Prediction of outcome in head-and-neck cancer patients using the standardized uptake value of 2-[18F]fluoro-2-deoxy-D-glucose. Int J Radiat Oncol Biol Phys 2004;59:1295–300.
25. Gupta T, Jain S, Agarwal JP, et al. Diagnostic performance of response assessment FDG-PET/CT in patients with head and neck squamous cell carcinoma treated with high-precision definitive (chemo)radiation. Radiother Oncol 2010;97:194–9.
26. Eleftheriadou A, Chalastras T, Ferekidou E, et al. Clinical effectiveness of tumor markers in squamous cell carcinoma of the larynx. Anticancer Res 2006;26:2493–7.
27. Kimura Y, Fujieda S, Takabayashi T, Tanaka T, Sugimoto C, Saito H. Conventional tumor markers are prognostic indicators in patients with head and neck squamous cell carcinoma. C cer Lett 2000;155:163–8.
28. Rewari A, Lu H, Parikh R, Yang Q, Shen Z, Haffty BG. BCCIP as a prognostic marker for rad therapy of laryngeal cancer. Radiother Oncol 2009;90:183–8.
29. de Jong MC, Pramana J, Knegjens JL, et al. HPV and high-risk gene expression profiles predict response to chemoradiotherapy in head and neck cancer, independent of clinical factors. Radiother Oncol 2010;95:365–70.
30. Lassen P, Eriksen JG, Krogdahl A, et al. The influence of HPV-associated p16- expression on accelerated fractionated radiotherapy in head and neck cancer: evaluation of the randomised DAHANCA 6&7 trial. Radiother Oncol 2011.
31. Troost EG, Schinagl DA, Bussink J, Oyen WJ, Kaanders JH. Clinical evidence on PET–CT for radiation therapy planning in head and neck tumours. Radiother Oncol 2010;96:328–34.
32. Moule RN, Kayani I, Moinuddin SA, et al. The potential advantages of 18FDG PET/CT-based target volume delineation in radiotherapy planning of head and neck cancer. Radiother Oncol 2010;97:189–93.
33. Pignon JP, Bourhis J, Domenge C, Designe L. Chemotherapy added to locoregional treatment for head and neck squamous-cell carcinoma: three meta-analyses of updated individual data. MACH-NC Collaborative Group. Meta-analysis of chemotherapy on head and neck cancer. Lancet 2000;355:949–55.

34. Pignon JP, le Maitre A, Bourhis J. Meta-analyses of chemotherapy in head and neck cancer (MACH-NC): an update. Int J Radiat Oncol Biol Phys 2007;69:S112–4.
35. Pignon JP, le Maitre A, Maillard E, Bourhis J. Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): an update on 93 randomised trials and 17, 346 patients. Radiother Oncol 2009;92:4–14.
36. Overgaard J. Chemoradiotherapy of head and neck cancer – can the bumble bee fly? Radiother Oncol 2009;92:1–3.
37. Langendijk JA, Doornaert P, Rietveld DHF, Verdonck-de Leeuw IM, René Leemans C, Slotman BJ. A predictive model for swallowing dysfunction after curative radiotherapy in head and neck cancer. Radiother Oncol 2009;90:189–95.
38. Langius JAE, Doornaert P, Spreeuwenberg MD, Langendijk JA, Leemans CR. Schueren, MAEvB-dvd. Radiotherapy on the neck nodes predicts severe weight loss in patients with early stage laryngeal cancer. Radiother Oncol 2010;97:80–5.
39. Rennemo E, Zätterström U, Evensen J, Boysen M. Reduced risk of head and neck secon primary tumors after radiotherapy. Radiother Oncol 2009;93:559–62.8.

CHAPTER 5

Externally validated HPV-based prognostic nomogram for oropharyngeal carcinoma patients yields more accurate predictions than TNM staging

Accepted for publuication in Radiotherapy and Oncology

Externally validated HPV-based prognostic nomogram for oropharyngeal carcinoma patients yields more accurate predictions than TNM staging

Emmanuel Rios Velazquez, Frank Hoebers, Hugo J.W.L. Aerts, Michelle M. Rietbergen, Ruud H. Brakenhoff, C. René Leemans, Ernst-Jan Speel, Jos Straetmans, Bernd Kremer and Philippe Lambin

ABSTRACT

Background

Due to the established role of the human papillomavirus (HPV), the optimal treatment approach for oropharyngeal carcinoma is currently under debate. The purpose of this study was to evaluate the most important determinants of treatment outcome and to develop a multifactorial predictive model that could provide individualized predictions of treatment outcome in oropharyngeal carcinoma patients.

Materials and methods

We analyzed the association between clinico-pathological factors and overall and progression-free survival in 168 OPSCC patients treated with curative radiotherapy or concurrent chemo-radiation. A multivariate model was validated in an external dataset of 189 patients and compared to the TNM staging system. This nomogram will be made publicly available at www.predictcancer.org.

Results

Predictors of unfavorable outcomes were negative HPV-status, moderate to severe comorbidity, T3-T4 classification, N2b – N3 stage, male gender, lower hemoglobin levels and smoking history of more than 30 pack years. Prediction of overall survival using the multi-parameter model yielded a C-index of 0.82 (95% CI, 0.76 – 0.88). Validation in an independent dataset yielded a C-index of 0.73 (95% CI, 0.66 – 0.79. For progression-free survival, the model's C-index was 0.80 (95% CI, 0.76 – 0.88), with a validation C-index of 0.67, (95% CI, 0.59 – 0.74). Stratification of model estimated probabilities showed statistically different prognosis groups in both datasets ($p < 0.001$).

Conclusion

This nomogram was superior to TNM classification or HPV status alone in an independent validation dataset for prediction of overall and progression-free survival in OPSCC patients, assigning patients to distinct prognosis groups. These individualized predictions could be used to stratify patients for treatment de-escalation trials.

INTRODUCTION

An increasing body of evidence has shown the relationship between the human papilloma virus (HPV) and squamous cell carcinoma of the oropharynx (OPSCC)[1-3].

Several studies have demonstrated that HPV- associated oropharyngeal cancer has a distinctly better survival after treatment, compared to HPV-negative tumors. However, the prognosis of HPV-positive oropharyngeal cancer seems to be significantly worse if there is a history of smoking[4,5]. This accumulated evidence suggests that tailored OPSCC cancer therapies, in which specific information about HPV status, and other patient characteristics are taken into account, need to be designed as a step forward from current population based therapies.

A tool that combines these factors to accurately anticipate patient´s outcome is needed. An analysis of the RTOG 0129 study proposed a stratification algorithm, combining HPV, T-stage, N-stage and smoking history, to assign patients into different prognostic groups[2]. This single-cohort based algorithm, although able to discriminate patients according to their risk of failure, was based on patients treated within a randomized trial with strict inclusion criteria, including mainly patients with T3-T4 tumors and with limited comorbidity.

A recent approach called rapid learning, which aims to drive the process of knowledge discovery by routinely and iteratively learning from data generated through patient care, proposes an alternative for knowledge extraction to evidence based clinical trials[6,7]. Clinical and outcomes information of unselected patients treated with different treatment modalities and with a larger heterogeneity in terms of stages, demographics and comorbidities can be analyzed to generate evidence representative of the consecutive patient in daily clinical practice, particularly for the advance elderly or with high comorbidities, frequently excluded from clinical trials.

This approach has been explored recently, successfully stratifying consecutive patients into distinct risk groups[8]. Unfortunately, an external validation of this model is not yet available.

In this study we evaluated the most important prognostic factors in OPSCC patients, treated with (chemo) radiation, such as HPV and smoking history, in combination with other patient and tumor characteristics to develop a robust nomogram that could provide individualized predictions of treatment outcome. The proposed predictive nomogram was externally validated in an independent cohort of consecutive OPSCC patients. As this knowledge was extracted from patients in routine clinical care, it can subsequently be implemented in clinical practice: it will improve the information given to patients regarding their prognosis, and could allow eligibility for treatment de-escalation trials.

MATERIAL AND METHODS

Patient population

All consecutive patients with OPSCC, stages (I-IVb) treated at Maastro Clinic between January 2000 and October 2011. 168 patients were included, treated with curative intent (including definitive radiotherapy or concurrent chemo-radiation). This analysis was approved by the Institutional Review Board (No. 11-29-14/09-intern-6430; NCT01985984).

Treatment details

Treatment options were either definitive radiotherapy alone or concurrent chemo-radiation with high dose cisplatin every 3 weeks. Patients treated with definitive radiotherapy received a continuous course of radiotherapy delivered by 4–6 MV linear accelerator. Patients were treated with fractionation schedules: patients with early oropharyngeal cancers (stage I-II) were treated with Accelerated Fractionated RadioTherapy (AFRT) to 68 Gy in 34 fractions over 37-38 days, the first 23 fractions 2 Gy daily, and the last 11 fractions twice daily in fractions of 2 Gy. Patients in moderate general condition, who were deemed unfit for AFRT received standard fractionated radiotherapy to 70 Gy in 35 fractions over 7 weeks.

HPV testing

To determine HPV status formalin-fixed, paraffin-embedded (FFPE) biopsy material of histopathologically confirmed OPSCC were retrieved from the archives of the Department of Pathology, University Hospital Maastricht, The Netherlands. FFPE material had been classified by histopathology and analyzed by means of $p16^{INK4A}$ immunostaining and for the presence of oncogenic HPV16 DNA by PCR in 168 available specimens[9]. A tumor was considered HPV positive if the HPV16 DNA by PCR results were positive.

Statistical analysis

The factors evaluated for their prognostic potential were HPV status, smoking and alcohol history, patient comorbidity, pre-treatment hemoglobin levels, gender, age, tumor location and TNM classification. All patient and treatment characteristics were collected from medical records. Patient comorbidity was scored using the Adult Comorbidity Evaluation 27[10].

N – stage was subdivided into two categories comparing N0 – N2a stage against N2b – N3 stage since patients in these categories have different clinical implications[8,11]. Missing values were imputed using the predictive mean matching algorithm.

Study endpoints were progression-free survival and overall survival, calculated from the start of radiotherapy. An event for progression-free survival was defined as

death or the first documented recurrence either recurrent local-regional disease or distant metastases after treatment. For overall survival, data were considered right-censored if patients were still alive at the time of last follow-up. For progression-free survival analysis, data were considered right-censored if patients did not develop a local-regional recurrence or distant metastases and were alive at the time of last follow-up.

The X^2-test was used for comparisons of categorical variables. For univariate survival analysis, the Kaplan Meier method was used. Groups were compared using the log rank test.

A multivariate Cox Proportional Hazard Regression analysis was performed to establish factors independently contributing to treatment-outcome. Two-sided p-values of <0.05 were considered statistically significant. A multivariate model combining the most important predictors was converted into a visual nomogram[12], and validated in an external cohort of patients from the VU University Medical Center, Amsterdam, The Netherlands. Model performance was evaluated using the C-index. The maximum value of the C-index is 1.0; indicating a perfect prediction model. A value of 0.5 indicates that 50% of the patients are correctly classified. Bootstrapping was used to obtain model prediction confidence intervals. The Maastro and external validation cohorts were split, using this model, into three subgroups according to the 33 and 66 percentiles of the risk score. The nomogram will be publicly available on the website www.predictcancer.org, after publication. Raw data of the training dataset will be available on www.cancerdata.org. Analyses were performed using SPPS 19.0 (SPSS Inc., Chicago) and Matlab 7.11.0 (The MathWorks Inc., Natick, MA).

Validation cohort

Patient characteristics of the validation cohort are shown in Supplementary Table 1. It consisted of a consecutive series of 189 OPSCC patients curatively treated at the VU University Medical Center, Amsterdam, The Netherlands, between January 2000 and December 2006. Treatment options included definitive radiotherapy alone and chemo-radiation. The definitive radiotherapy regime consisted of standard fractionated radiotherapy to 70 Gy in fractions of 2 Gy over 7 weeks. The concomitant chemo-radiation scheme included daily fractionation of 2 Gy up to 70 Gy with a concomitant intra-venous administration of cisplatin with a dose of 100 mg/m^2 at three weeks intervals.

RESULTS

The patient, tumor and treatment characteristics are shown in Table 1. The majority of the patients were male (74.4 %) and the median age at the start of therapy was 59 years (range: 43 – 83 years). The median follow-up of all patients was 26 months (range 2.5 – 127.2) and it was 37.5 months (range 6.4 – 127.2) for patients alive at last follow-up.

Table 1: Patient and tumor characteristics and univariate analysis results; Maastro cohort (n = 168).

	Frequency (%)	Log Rank test *p* Overall survival	Log Rank test *p* PF survival
Age (years)	59,5 (43 – 83)	0,498	0,282
Gender		0,004	0,006
Male	74,4		
Female	25,6		
Primary tumor sub-location		0,052	0,140
Tonsillar fossa	36,9		
Base of tongue	29,8		
Oropharynx overlap	25,6		
Soft palate	7,7		
Differentiation grade		0,486	0,379
Good	6,5		
Moderate	51,8		
Poor	27,4		
SCC nos	10,7		
Undifferentiated	3,6		
Smoking pack years	30 (0 – 100)		
Split by median (>30)		0,025	0,026
Split by percentiles		0,078	0,083
Alcohol unit years	134 (0 – 660)		
Split by median (>134)		0,042	0,004
Split by percentiles		0,047	0,005
Comorbidity score (ACE-27)		0,000	0,000
None	33,3		
Mild	41,1		
Moderate	19,0		
Severe	6,5		
T-stage		0,001	0,004
T1	14,9		
T2	27,4		
T3	22,6		
T4	35,1		
N-stage		0,048	0,065
N0	34,5		
N1	17,3		
N2	44,1		
N3	3,6		
Nx	0,6		
N0 – N2a vs N2b – N3		0,021	0,053
P16 immunostaining			
Positive	34,5	0,000	0,000
Negative	64,3		
Unknown	1,2		
HPV status			
Positive	30,4	0,000	0,000
Negative	69,6		
Treatment		0,065	0,060
Radiation only	67,9		
Chemo-radiation	32,1		
RT Dose (Gy)	68 (60 – 70)	0,888	0,865
Pre-RT Haemoglobin levels (mmol/L)	8,5 (5,1 – 11,3)	0,006	0,004

At the time of last follow-up 60.1% of patients were alive and 39.9 % had deceased. Progression free survival was 47% at 5 years. A total of 29 (17.3%) local-regional recurrences were observed.

Patient characteristics and HPV status

Immunostaining for p16 was positive in 58 cases (34.5 %) and missing in 1.2% of the cases. After HPV DNA testing, a total of 51 (30.4 %) was considered as HPV positive. Due to its importance in OPSCC patients, we evaluated the association between HPV status and other patient and tumor characteristics. Overall survival was significantly better for patients with an HPV-positive OPSCC (CI, 83.66 – 120.21 months), compared to patients with an HPV-negative OPSCC (CI, 48.6 – 68.2 months; $p<0.0001$). The 5-year overall survival rates were 82% in the HPV-positive group and 39% in the HPV-negative group. For progression-free survival, the surviving rates were 83% and 35% for the HPV-positive and HPV-negative groups respectively ($p < 0.001$).

HPV status was positive in 32.5% and 29.6% of female and male patients respectively. Patients with HPV-positive status were more likely to have none to moderate comorbidity (72.5% of HPV positive cases, p = NS; ACE-27 score 0 – 1); these patients also showed a clear tendency towards moderate smoking and alcohol consumption compared to HPV-negative patients ($p < 0.001$). No significant differences were observed when comparing HPV status and nodal status, tumor stage and age. There was a higher incidence of HPV-positive tumors in the tonsils and base of tongue, compared to the other oropharyngeal sub-locations ($p = 0.001$). Poorly differentiated tumors had significant higher incidence of HPV-positivity compared to well differentiated tumors ($p < 0.006$).

Prognostic factors for overall survival and progression-free survival

Univariate analysis was performed to evaluate the prognostic significance of the tumor and patient characteristics shown in Table 1. The variables that were associated with shorter overall survival were male gender ($p = 0.004$), pack years of smoking higher than the median value (median = 30 pack years; $p = 0.025$), unit years of alcohol consumption higher than the median value (median = 134 unit years; $p = 0.042$), higher ACE-27 comorbidity index ($p < 0.0001$), higher T-stage ($p < 0.0001$), N2b-N3 stage ($p = 0.021$), negative HPV status ($p < 0.0001$) and lower pre-radiotherapy hemoglobin levels than the median value (median = 8.5 mmol/L; $p < 0.006$). Differentiation grade did not show significant differences in overall survival ($p = 0.654$). Tumors located in the posterior oropharynx wall had a trend towards worse survival, compared to other tumor sub-locations ($p = 0.052$).

Treatment parameters such as radiotherapy delivered dose and overall treatment time did not show a correlation with overall survival ($p > 0.05$). Likewise, no significant differences in overall survival were observed based on treatment type (radiation only vs. chemoradiation, $p = 0.065$).

Gender, pack years of smoking, unit years of alcohol consumption, comorbidity, T-stage, HPV status and pre-radiotherapy hemoglobin levels were individually associated with progression-free survival (Table1).

Table 2: Multivariate Cox proportional hazards analysis of potential prognostic factors

	Overall survival			Local control		
	HR	95% CI	*p*	HR	95% CI	*p*
Age	1,003	973 - 1,034	,856	1,015	,986 -1,044	,320
Gender						
Female						
Male	2,511	1,199 - 5,258	,015	2,101	1,094 -4,037	,026
Pre-RT Hemoglobin levels	,693	,531 - ,903	,007	,802	,633 – 1,016	,067
Pack years of smoking	1,006	,992 - 1,021	,792	1,006	,993 -1,020	,368
Unit years of alcohol comsumption	1,001	,999 - 1,002	,554	1,002	1,000 - 1,003	,057
T-Stage						
T1						
T2	,945	,337 - 2,759	,945	1,157	,457 - 2,931	,759
T3	3,284	1,287 - 8,437	,014	2,941	1,225 - 7,063	,016
T4	2,281	,875 - 5,942	,092	2,216	,914 - 5,374	,078
N-Stage						
N0 - N2a						
N2b - N3	2,588	1,413 - 4,738	,002	2,332	1,362 - 3,995	,002
Comorbidity						
Score (0 -1)						
Score (2 -3)	2,347	1,307 - 4,216	,004	1,728	1,004 - 2,972	,048
HPV status						
Positive						
Negative	6,027	2,487 - 14,607	,001	4,746	2,183 - 10,322	,001

Some prognostic factors in the univariate analysis (Table 1) were no longer significant in the multivariate cox-regression analysis. For overall survival, the factors that remained as independent contributors of unfavorable treatment outcome were male gender, low pre-treatment hemoglobin levels (<median), higher T-stage, N2b – N3 stage, negative HPV status and high comorbidity (moderate to severe). Multivariate hazard ratios, confidence intervals and significance levels are shown in Table 2.

For progression-free survival male gender, high comorbidity, higher T-stage, N2b – N3 stage and negative HPV status remained as significant independent prognostic factors. All other parameters did not show a significant correlation with progression-free survival (Table 2). Prediction of overall survival yielded a C-index of 0.82 (95% CI, 0.76 – 0.88) based on the Maastro Clinic dataset. In the independent external validation dataset, the C-index was 0.73 (95% CI, 0.66 – 0.79).

The resulting nomogram, shown in Figure 1, estimates outcome probabilities by assigning a score to each predictor value. The sum of these scores corresponds to an outcome event probability.

Table 3: Multivariate model performance (c-index) and comparison with TNM and HPV.

	Maastro cohort		VUMC cohort	
	OS	**PFS**	**OS**	**PFS**
Multivariate model	0.82 (CI, 0.76 – 0.88)	0.80 (CI, 0.75 – 0.87)	0.73 (CI, 0.66 – 0.79)	0.67 (CI, 0.59 – 0.74)
TNM	0.66* (CI, 0.61 – 0.75)	0.65* (CI, 0.60 – 0.72)	0.64* (CI, 0.59 – 0.73)	0.60* (CI, 0.53 – 0.68)
HPV	0.68* (CI, 0.61 – 0.72)	0.68* (CI, 0.60 – 0.74)	0.68* (CI, 0.63 – 0.73)	0.54* (CI, 0.49 – 0.59)
Ang's Model[2]	0.76* (CI, 0.65 – 0.80)	0.74 (CI, 0.70 – 0.82)*	0.72 (CI, 0.64 – 0.78)	0.66 (CI, 0.60 – 0.73)

*C-index confidence intervals were obtained in a bootstrap procedure (n = 100). *Indicates whether the multivariate model performance was significantly higher than TNM or HPV (p <0.0001).*

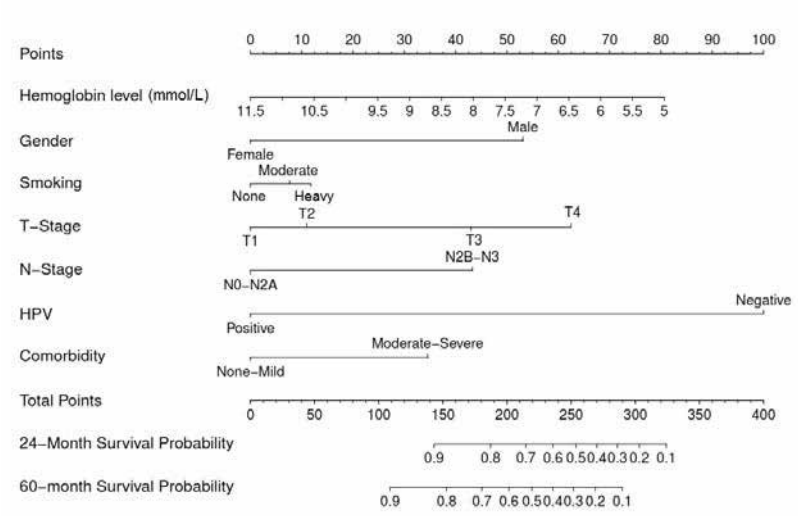


Figure 1

Multivariate model converted to a graphic nomogram for prediction of overall survival. Each variable in the model, corresponding to the characteristics of an individual patient, is assigned to an individual score. A probability for overall survival can be calculated by drawing a vertical line from each predictor value to the score scale at the top—'points'. After manually summing up the scores, the 'total points' correspond to the probability of overall survival (or progression free survival respectively, Supplementary Figure 1), which are estimated by drawing a vertical line from this value to the bottom scales to estimate overall survival. Smoking was categorized as none, moderate (1-30 pack years of smoking) and heavy (> 30 pack years of smoking).

The most important factor in the nomogram to estimate overall survival is HPV status. Kaplan-Meier curves of the model estimates for the development and validation cohorts are shown in Figure 2a. This stratification showed significant differences in outcomes for the three proposed risk groups, in both datasets ($p < 0.001$). For progression-free survival, the model's C-index was 0.80 (95% CI, 0.76 – 0.88), with a validation C-index in the external dataset of 0.67 (95% CI, 0.59 – 0.74).

Table 4: Median survival times for the stratification risk groups as estimated with the multivariate model.

	95% Confidence Interval	
	Lower Bound	Upper bound
Overall survival (Maastro)		
High (n = 55)	21,63	35,25
Intermediate (n = 56)	57,04	86,70
Low (n = 57)	100,09	123,24
Overall survival (VUMC)		
High (n = 62)	20,81	32,14
Intermediate (n = 63)	35,81	46,76
Low (n = 64)	45,00	54,36
Progression-free survival (Maastro)		
High (n = 55)	16,80	31,03
Intermediate (n= 56)	43,39	68,07
Low (n = 57)	94,74	120,13
Progression-free survival (VUMC)		
High (n = 62)	20,47	32,88
Intermediate (n = 63)	34,39	45,98
Low (n = 64)	39,38	50,85

Differences in survival among the risk groups were statistically significant in all cases (log rank test, p < 0,0001).

Again, the predictive nomogram was able to estimate individual progression-free survival rates and assign patients to clearly distinct risk groups in the validation cohort ($p < 0.001$; Figure 2b). A comparison of the multivariate model performance with TNM staging, HPV alone and Ang's model[2] is shown in Table 3. Median survival rates for the distinct risk groups are summarized in Table 4.

DISCUSSION

We evaluated the prognostic significance of HPV and other factors of clinical interest, in a large cohort of consecutive OPSCC patients, to develop a multifactorial predictive model that can provide individual estimations of treatment outcome in this patient population.

Combining the most important prognostic factors in a multivariate model, including HPV status, comorbidity score, T-stage, N-stage, pack years of smoking, gender and pre-treatment hemoglobin levels yielded high predictive performances, as shown by the C-index for overall survival of 0.82 (95% CI, 0.76 – 0.88) and of 0.80 (95% CI, 0.76 – 0.88) for

progression-free survival. This model was validated in an external unselected cohort of OPSCC patients (n = 189), showing reliable validation model performances for overall survival (0.73; 95% CI, 0.66 – 0.79) and progression-free survival (0.67; 95% CI, 0.59 – 0.74). Model predictions were significantly better than using TNM or HPV alone.

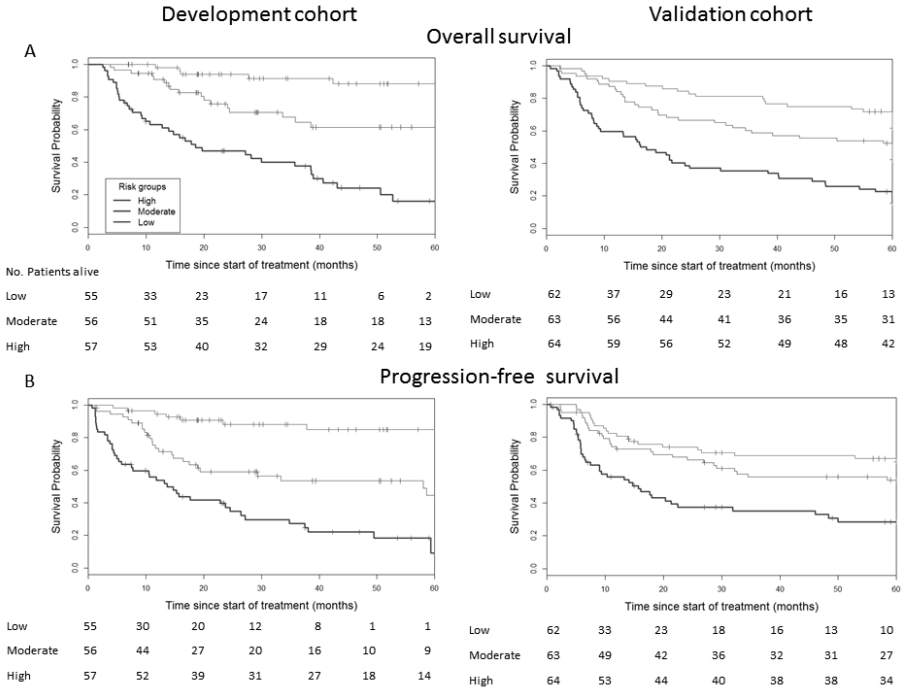


Figure 2

Kaplan-Meier curves of risk group stratification for (a) overall survival and (b) progression-free survival. Nomogram risk group stratifications are shown for the development cohort (left), and for the validation cohort (right). All survival curves are statistically different (log rank test, p < 0.0001)

Also, this multivariate model showed higher C-indexes when compared with the published model based on the RTOG 0129 study including HPV, T-classification, N-classification and smoking history[2].

This prognostic model for OPSCC patients has been validated in an independent dataset by directly applying the model weights to the validation raw data. Previously published models have been evaluated in a single development cohort[2,8], although Ang's model has been recently evaluated by two groups[8,13]. Our model was able to stratify patients according to their estimated risk of failure into distinct risk groups, in both cohorts, for overall and progression-free survival. However, the performance was lower for prediction progression-free survival in the validation cohort.

We followed the so-called rapid-learning approach in which knowledge is derived from unselected patient databases, as compared to medical evidence derived from clinical

trials[6,14]. This approach has an obvious advantage of including a more heterogeneous group of patients, in terms of clinical stage, comorbidity and treatments. In this way, the knowledge derived can be used for decisions concerning new patients, including the elderly patient or the patient with severe comorbidity, which would not be included in a clinical trial. The clinical and patient characteristics included in this study were selected based on medical expertise, known prognostic importance from literature and availability[1,2,15].

In our study, overall survival rates and progression-free survival rates were comparable with other studies[8,16]. The frequency of HPV-associated OPSCC in our cohort is comparable to other recent European series[8,9,16]. Similarly, we found HPV-positive status to be associated with low smoking and alcohol consumption, and less likely to have severe comorbidity ($p<0.0001$). Furthermore, the presence of HPV correlated positively with poor differentiation grade ($p < 0.006$) and was more often present in tumors of the tonsils and the base of the tongue ($p = 0.001$). These findings are in line with previously observed correlations between HPV incidence and patient demographics and tumor characteristics[9,15,17]. HPV-positive cancers have been associated with smaller primary tumors and with greater regional disease[1], in our study, no significant differences in HPV-prevalence were observed among different T-stages or N-stages.

Tobacco smoking has been established as a major independent-prognostic factor for patients with OPSCC[4,9,17], these studies showed that cancer progression and risk of death increases with tobacco exposure, independently of tumor HPV status and treatment. In our study, pack years of smoking was a significant prognostic factor for overall and progression-free survival, however, in the multivariate analysis, it did not remain as independent prognostic factor.

A limitation in our study, inherent to its retrospective nature is the lack of standardization in which data has been collected over the years. Furthermore, smoking behavior during therapy, which has been recently reported as important prognostic factor[4,5,18], was not available in our study.

This further highlights the increasing need for systematic routine patient care data collection, warehouse and semantic inter-operable data retrieval systems, to assure improved and standardized data retrieval and allow external applicability[14,19,20].

Moderate to severe comorbidity, higher T-stage and advanced N-stage were independent unfavorable prognostic factors for overall survival and progression-free survival. We used the ACE-27 comorbidity score, a validated comorbidity scoring system, which has been previously associated with patient prognosis in head and neck cancers[8,21,22]. Advanced clinical T-classification has been reported as a significant risk factor for progressive disease and death in oropharyngeal carcinoma patients[17]. Indeed, T3-T4 tumors showed poorer survival, compared to T1-T2 tumors. Similarly, we observed that higher N-stage was associated with worse survival; however this association was less significant with progression-free survival. Comparing N0-N2a nodal stages against N2b-N3 stages showed marked differences in survival, with the latter being an unfavorable prognostic factor. This re-grouping of N-stage has shown prognostic value previously[8,11]. Male gender was a

strong negative prognostic factor for overall survival and progression-free survival, however this effect remained significant in the multivariate setting only for overall survival. Other studies have shown male gender to be an unfavorable prognostic factor, as well as in other head and neck cancer sites, however in OPSCC this association can be confounded by the fact that men have a higher incidence of HPV-positive OPSCC than women[4,17,18,23].

We showed that combining tumor HPV status with other important prognostic factors, increased the accuracy in the predictions, compared to the traditional TNM staging system or individually. 95% CI of the model predictions were significantly better than those obtained with TNM alone or HPV status alone, which underlines the importance of multifactorial prediction models.

This model performance is acceptable for clinical support, particularly due to the clear distinction in risk groups, in both cohorts; however it is still far from optimal. Combining clinical parameters with HPV, is a first step into developing validated decision support systems in head and neck cancer; however we anticipate that adding other features, such as diagnostic and molecular imaging, and other important biomarkers such as EGFR or CA-IX will increase model accuracy[14,19,24,25]. Standardization and systematic collection of routine patient care data will likewise increase model reliability and allow further validation.

In conclusion, we showed that combining HPV status with a set of important clinical parameters allows the development of multifactorial models to predict overall and progression-free survival. Applying this model to individual patients can support their stratification according to their estimated risk and their eligibility for different treatment approaches[11,26], for instance, ongoing trials are evaluating treatment de-intensification for OPSCC with estimated good prognosis (NCT01663259). Thus, population-based learning can improve the information given to patients regarding their prognosis as well as in the long term allow stratification in prospective clinical trials and treatment individualization.

ACKNOWLEDGMENTS

Authors acknowledge financial support from the CTMM framework (AIRFORCE project, grant 030-103), EU 6th and 7th framework program (METOXIA, EURECA, ARTFORCE), euroCAT (IVA Interreg - www.eurocat.info) and the Dutch Cancer Society (KWF UM 2011-5020, KWF UM 2009-4454).

REFERENCES

1. Adelstein DJ, Rodriguez CP: Human papillomavirus: changing paradigms in oropharyngeal cancer. Curr Oncol Rep 12:115-20, 2010

2. Ang KK, Harris J, Wheeler R, et al: Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med 363:24-35, 2010
3. Marur S, D'Souza G, Westra WH, et al: HPV-associated head and neck cancer: a virus-related cancer epidemic. Lancet Oncol 11:781-9, 2010
4. Ang KK, Sturgis EM: Human papillomavirus as a marker of the natural history and response to therapy of head and neck squamous cell carcinoma. Semin Radiat Oncol 22:128-42, 2012
5. Gillison ML, Zhang Q, Jordan R, et al: Tobacco smoking and increased risk of death and progression for patients with p16-positive and p16-negative oropharyngeal cancer. J Clin Oncol 30:2102-11, 2012
6. Abernethy AP, Etheredge LM, Ganz PA, et al: Rapid-learning system for cancer care. J Clin Oncol 28:4268-74, 2010
7. Lambin P, Roelofs E, Reymen B, et al: 'Rapid Learning health care in oncology' - An approach towards decision support systems enabling customised radiotherapy'. Radiother Oncol 109:159-64, 2013
8. Rietbergen MM, Brakenhoff RH, Bloemena E, et al: Human papillomavirus detection and comorbidity: critical issues in selection of patients with oropharyngeal cancer for treatment De-escalation trials. Ann Oncol 24:2740-5, 2013
9. Hafkamp HC, Manni JJ, Haesevoets A, et al: Marked differences in survival rate between smokers and nonsmokers with HPV 16-associated tonsillar carcinomas. Int J Cancer 122:2656-64, 2008
10. Kallogjeri D, Piccirillo JF, Spitznagel EL, Jr., et al: Comparison of Scoring Methods for ACE-27: Simpler Is Better. J Geriatr Oncol 3:238-245, 2012
11. O'Sullivan B, Huang SH, Siu LL, et al: Deintensification candidate subgroups in human papillomavirus-related oropharyngeal cancer according to minimal risk of distant metastasis. J Clin Oncol 31:543-50, 2013
12. Iasonos A, Schrag D, Raj GV, et al: How to build and interpret a nomogram for cancer prognosis. J Clin Oncol 26:1364-70, 2008
13. Granata R, Miceli R, Orlandi E, et al: Tumor stage, human papillomavirus and smoking status affect the survival of patients with oropharyngeal cancer: an Italian validation study. Ann Oncol 23:1832-7, 2012
14. Lambin P, Rios-Velazquez E, Leijenaar R, et al: Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441-6, 2012
15. Rischin D, Young RJ, Fisher R, et al: Prognostic significance of p16INK4A and human papillomavirus in patients with oropharyngeal cancer treated on TROG 02.02 phase III trial. J Clin Oncol 28:4142-8, 2010
16. Straetmans JM, Olthof N, Mooren JJ, et al: Human papillomavirus reduces the prognostic value of nodal involvement in tonsillar squamous cell carcinomas. Laryngoscope 119:1951-7, 2009
17. Naeini KM, Pope WB, Cloughesy TF, et al: Identifying the mesenchymal molecular subtype of glioblastoma using quantitative volumetric analysis of anatomic magnetic resonance images. Neuro Oncol 15:626-34, 2013
18. Garden AS, Kies MS, Morrison WH, et al: Outcomes and patterns of care of patients with locally advanced oropharyngeal carcinoma treated in the early 21st century. Radiat Oncol 8:21, 2013
19. Lambin P, van Stiphout RG, Starmans MH, et al: Predicting outcomes in radiation oncology--multifactorial decision support systems. Nat Rev Clin Oncol 10:27-40, 2013
20. Roelofs E, Persoon L, Nijsten S, et al: Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. Radiother Oncol 108:174-9, 2013
21. Habbous S, Harland LT, La Delfa A, et al: Comorbidity and prognosis in head and neck cancers: Differences by subsite, stage, and human papillomavirus status. Head Neck, 2013
22. Ankola AA, Smith RV, Burk RD, et al: Comorbidity, human papillomavirus infection and head and neck cancer survival in an ethnically diverse population. Oral Oncol 49:911-7, 2013
23. Egelmeer AG, Velazquez ER, de Jong JM, et al: Development and validation of a nomogram for prediction of survival and local control in laryngeal carcinoma patients treated with radiotherapy alone: a cohort study based on 994 patients. Radiother Oncol 100:108-15, 2011
24. Lassen P, Overgaard J, Eriksen JG: Expression of EGFR and HPV-associated p16 in oropharyngeal carcinoma: Correlation and influence on prognosis after radiotherapy in the randomized DAHANCA 5 and 7 trials. Radiother Oncol, 2013

25. Reimers N, Kasper HU, Weissenborn SJ, et al: Combined analysis of HPV-DNA, p16 and EGFR expression to predict prognosis in oropharyngeal cancer. Int J Cancer 120:1731-8, 2007
26. Bossi P, Orlandi E, Miceli R, et al: Treatment-related outcome of oropharyngeal cancer patients differentiated by HPV dictated risk profile: a tertiary cancer centre series analysis. Ann Oncol 25:694-9, 2014

PART 2

Radiomics

CHAPTER 6

Radiomics: Extracting more information from medical images using advanced feature analysis

Published in: European Journal of Cancer (2012) 48, 441– 446

Radiomics: Extracting more information from medical images using advanced feature analysis *Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, Ruud G.P.M. van Stiphout, Patrick Granton, Catharina M.L. Zegers, Robert Gillies, Ronald Boellard, Andre Dekker, Hugo J.W.L. Aerts*

ABSTRACT

Solid cancers are spatially and temporally heterogeneous. This limits the use of invasive biopsy based molecular assays but gives huge potential for medical imaging, which has the ability to capture intra-tumoural heterogeneity in a non-invasive way. During the past decades, medical imaging innovations with new hardware, new imaging agents and standardized protocols, allows the field to move towards quantitative imaging. Therefore, also the development of automated and reproducible analysis methodologies to extract more information from image-based features is a requirement. Radiomics – the high-throughput extraction of large amounts of image features from radiographic images – addresses this problem and is one of the approaches that hold great promises but need further validation in multi-centric settings and in the laboratory.

INTRODUCTION

The use and role of medical imaging technologies in clinical oncology has greatly expanded from primarily a diagnostic tool to include a more central role in the context of individualised medicine over the past decade (Fig. 1). It is expected that imaging contains complementary and interchangeable information compared to other sources, e.g. demographics, pathology, blood biomarkers, genomics and that combining these sources of information will improve individualised treatment selection and monitoring.[1]

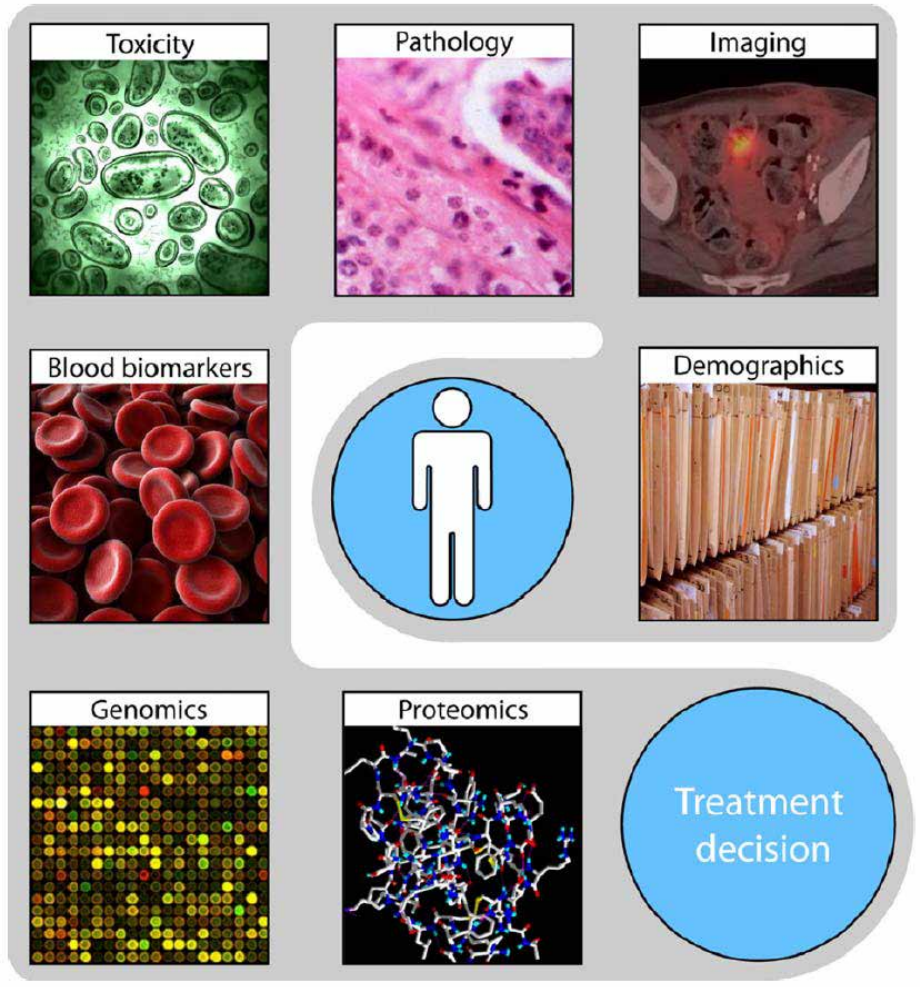


Figure 1
Different sources of information, e.g. demographics, imaging, pathology, toxicity, biomarkers, genomics and proteomics, can be used for selecting the optimal treatment.

Cancer can be probed in many ways depending on the non-invasive imaging device used or the mode by which it operates (Fig. 2). Classically, anatomical computed tomography (CT) imaging is a often used modality, acquiring images of the 'anatome' in high resolution (e.g. 1 mm^3). CT imaging is now routinely used and is playing an essential role in all phases of cancer management, including prediction, screening, biopsy guidance for detection, treatment planning, treatment guidance and treatment response evaluation.[2,3]

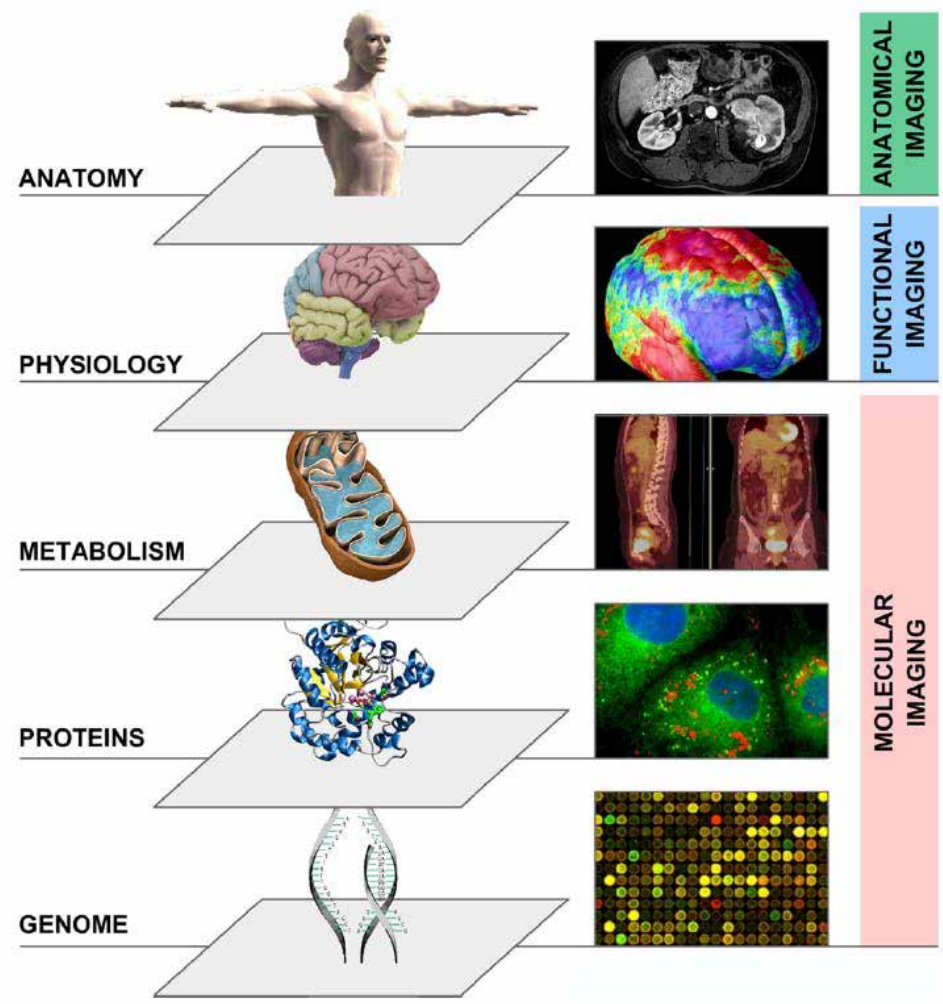


Figure 2
Multilevel imaging: anatomical, functional, and molecular imaging.

CT is used in the assessment of structural features of cancer but it is not perceived to portray functional or molecular details of solid tumours. Functional imaging concerns physiological processes and functions such as diffusion, perfusion and glucose uptake. Here, commonly used methodologies are dynamic contract enhanced-magnetic resonance imaging (DCE-MRI), assessing tumour perfusion and fluoro-2-deoxy-D-glucose (FDG) positron emission tomography (PET) imaging, assessing tumour metabolism, which both often are found to have prognostic value.[4–6] Finally, another modality is molecular imaging, visualising at the level of specific pathways or macro-molecule in vivo. For example, there are molecular markers assessing tumour hypoxia or labelled antibodies, assessing receptor expression levels of a tumour.[1,7]
Over the past decades, medical imaging has progressed in four distinct ways:

- Innovations in medical devices (hardware): This concerns improvements in imaging hardware and the development of combined modality machines. For example, in the last decade we moved from single slice CT to multiple slices CT and CT/PET. More recent developments are dual-source and dualenergy CT. These techniques significantly increase the temporal resolution for 4-D CT reconstructions allowing visualisation of fine structures in tissues, also in several stages in the cardiac or respiration phase. Moreover, dual-energy CT can be used to improve identification of tissue composition and density.

- Innovations in imaging agents: Innovations in imaging agents (or imaging biomarker, imaging probe, radiotracer), i.e. molecular substances injected in the body and used as an indicator of a specific biological process occurring in the body. This is achieved by contrast agents, i.e. an imaging agent using positive emission tomography (radiotracer). A common use is to find indications of pathological processes, e.g. hypoxia markers using PET imaging.

- Standardised protocol allowing quantitative imaging: Historically radiology has been a qualitative science, perhaps with the exception of the quantitative use of CT based electron densities in radiotherapy treatment planning. The use of standardised protocols like common MRI spin-echo sequences helps to allow multicentric use of imaging as well as transforming radiology to a more quantitative, highly reproducible science.

- Innovations in imaging analysis: The analysis of medical images has a large impact on the conclusions of the derived images. More and more software is becoming available, allowing for more quantification and standardisation. This has been illustrated by the development of the computer-assisted detection (CAD systems) that improves the performance of detecting cancer in mammography or in lung diseases.[8]

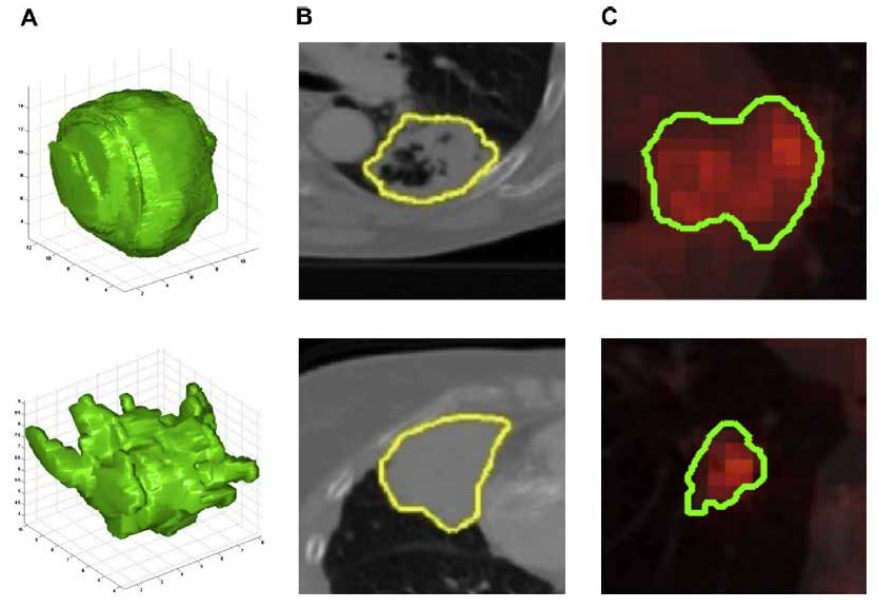


Figure 3

(A) Two representative 3-D representations of a round tumour (top) and spiky tumour (bottom) measured by computed tomography (CT) imaging. (B) Texture differences between non-small cell lung cancer (NSCLC) tumours measured using CT imaging, more heterogeneous (top) and more homogeneous (bottom). (C) Differences of FDG-PET uptake, showing heterogeneous uptake.

Radiomics focuses on improvements of image analysis, using an automated high-throughput extraction of large amounts (200+) of quantitative features of medical images and belongs to the last category of innovations in medical imaging analysis.

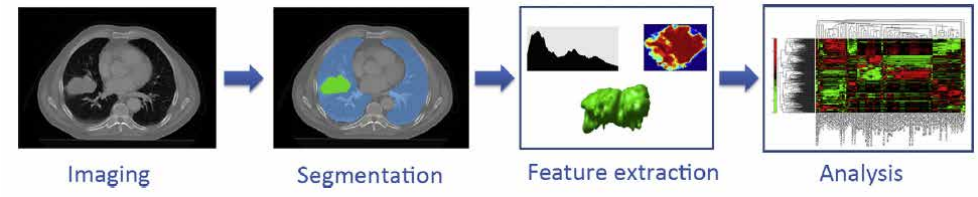


Figure 4

The Radiomics workflow. On the medical images, segmentation is performed to define the tumour region. From this region the features are extracted, e.g. features based on tumour intensity, texture and shape. Finally, these features are used for analysis, e.g. the features are assessed for their prognostic power, or linked with stage, or gene expression.

The hypothesis is that quantitative analysis of medical image data through automatic or semi-automatic software of a given imaging modality can provide more and better infor-

mation than that of a physician. This is supported by the fact that patients exhibit differences in tumour shape and texture measurable by different imaging modalities (Fig. 3).

THE WORKFLOW OF RADIOMICS: A (SEMI) HIGH-THROUGHPUT APPROACH

Fig. 4 depicts the processes involved in the Radiomics workflow. The first step involves the acquisition of high quality and standardised imaging, for diagnostic or planning purposes. From this image, the macroscopic tumour is defined, either with an automated segmentation method or alternatively by an experienced radiologist or radiation oncologist. Quantitative imaging features are subsequently extracted from the previously defined tumour region. These features involve descriptors of intensity distribution, spatial relationships between the various intensity levels, texture heterogeneity patterns, descriptors of shape and of the relations ofthe tumour with the surrounding tissues (i.e. attachment to the pleural wall in lung, differentiation).

The extracted image traits are then subjected to a feature selection procedure. The most informative features are identified based on their independence from other traits, reproducibility and prominence on the data. The selected features are then analysed for their relationship with treatment outcomes or gene expression. The ultimate goal is to provide accurate risk stratification by incorporating the imaging traits into predictive models for treatment outcome and to evaluate their added value to commonly used predictors.

THE RADIOMICS HYPOTHESIS: INFERRING PROTEO-GENOMIC AND PHENOTYPIC INFORMATION FROM RADIOLOGICAL IMAGES

The underlying hypothesis of Radiomics is that advanced image analysis on conventional and novel medical imaging could capture additional information not currently used, and more specifically, that genomic and proteomics patterns can be expressed in terms of macroscopic image-based features. If proven, we can infer phenotypes or gene–protein signatures, possibly containing prognostic information, from the quantitative analysis of medical image data.

This hypothesis is supported by image-guided biopsies, which demonstrated that tumours show spatial differences in protein expressions.[9] More specifically, it has been demonstrated that major differences in protein expression patterns within a tumour can be correlated to radiographic findings (or radiophenotypes) such as contrast-enhanced and non-enhanced regions based on CT data.[10] The authors suggest that image-guided proteomics holds promise for characterising tissues prior to treatment decisions and with-

out imaging there is indeed a risk that the optimum treatment decision could be neglected (i.e. the use or not of a targeted agent).
Also, Kuo et al. reported the association of CT-derived imaging traits with histo-pathologic markers, and several pre-defined gene expression modules on liver cancer. [11,12] In ovarian carcinoma, an imaging feature describing the enhancement fraction as proportion of enhancing tumour tissue on a pre-treatment CT scan, was found predictive for outcome after first line chemotherapy.[13] In lung cancer, CT derived information has been limited to pre-treatment assessment of tumour volume and as response evaluation defined as tumour size reduction.[14]

For PET imaging, the maximum and median FDG uptake has often been investigated, indicating strong prognostic power.[6,19] However, more complex descriptions of FDG uptake are only investigated on a limited scale. There was a study of El Naqa et al.[15], investigating the predictive power of intensity–volume histogram (IVH) metrics, shape and texture features to assess response to treatment of a limited set of patients with head and neck and cervix cancers. Tixier et al. also explored the potential of SUV based, shape and texture features extracted from baseline FDG-PET, images, to assess response to therapy and prognosis in order to predict response to combined chemo-radiation treatment in oesophageal cancer.[16] Also, textural features in FDG PET images exhibited small variations due to different acquisition modes and reconstruction parameters.[17]
These examples open the question of whether quantitative extraction of additional imaging features on conventional imaging improves the ability of currently used parameters to predict or monitor response to treatment. Furthermore, Radiomics can be linked with the concept of radio-genomics, which assumes that imaging features are related to gene signatures. An interesting finding in recent literature is that tumours with more genomic heterogeneity are more likely to develop a resistance to treatment and to metastasise.[18] This links to the concept that more heterogeneous tumours have a worse prognosis.

According to the Radiomics hypothesis, the genomic heterogeneity could translate to an expression in an intra-tumoural heterogeneity that could be assessed through imaging and that would ultimately exhibit worse prognosis. This hypothesis has been sustained by Jackson et al.[19] and as well as by Diehn et al.[20] who quite convincingly showed that proliferation and hypoxia gene expression patterns can be predicted by mass effect and tumour contrast enhancement, respectively. They also showed that a specific imaging pattern could predict overexpression of epidermal growth factor receptor (EGFR), a known therapeutic target. Moreover, in their analysis the presence of certain image features was highly predictive of outcome. The authors concluded that imaging in this case MR provided an '*in vivo portrait*' of genome-wide gene expression in glioblastoma multiform. Similar findings have been found in hepatocellular carcinomas by Segal et al.[21], showing that the combination of only 28 imaging traits was sufficient to reconstruct the variation of 116 gene expression modules.

These types of studies will need to be extended, by including more patients with external validation datasets, more tumour types that exhibit phenotypes such as invasive-

ness. This will be the focus of the QuIC-ConCePT consortium, to confirm experimentally the Radiomics hypothesis, namely to establish a causal relationship between gene expression patterns and image features.

CONCLUSIONS

Solid cancers have extraordinarily spatial and temporal heterogeneity at different levels: genes, proteins, cells, microenvironment, tissues and organs. This limits the use of biopsy based molecular assays but in contrast gives a huge potential for non-invasive imaging, which has the ability to capture intra-tumoural heterogeneity in a non-invasive way. Medical imaging innovations with new hardware, new imaging agents and standardized protocol now allow for quantitative imaging but require the development of 'smart' automated software to extract more information from image-based features.
Radiomics – the high-throughput extraction of image features from radiographic images – is one approach that holds great promises but needs further validation in a multi-centric setting and in the laboratory.

ACKNOWLEDGEMENTS

The authors are members of the QuIC-ConCePT project partly funded by EFPIA companies and the Innovative Medicine Initiative Joint Undertaking (IMI JU) under Grant Agreement No. 115151. We also acknowledge financial support from the CTMM framework (AIRFORCE project), EU 6th and 7th framework program (Euroxy and Metoxia program), Interreg (www.eurocat.info), and the Dutch Cancer Society (KWF UM 2011-5020, KWF UM 2009-4454).

REFERENCES

1. Lambin P et al. The ESTRO Breur Lecture 2009. From population to voxel-based radiotherapy: exploiting intra-tumour and intraorgan heterogeneity for advanced treatment of non-small cell lung cancer. Radiother Oncol 2010;96:145–52.
2. Fass L. Imaging and cancer: a review. Mol Oncol 2008;2:115–52.
3. Eisenhauer EA et al.. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 2009;45:228–47.
4. van Baardwijk A et al. The maximum uptake of (18)F-deoxyglucose on positron emission tomography scan correlates with survival, hypoxia inducible factor-1alpha and GLUT-1 in nonsmall cell lung cancer. Eur J Cancer 2007;43:1392–8.

5. Berghmans T et al. Primary tumor standardized uptake value (SUVmax) measured on fluorodeoxyglucose positron emission tomography (FDG-PET) is of prognostic value for survival in non-small cell lung cancer (NSCLC): a systematic review and meta-analysis (MA) by the European Lung Cancer Working Party for the IASLC Lung Cancer Staging Project. J Thorac Oncol 2008;3:6–12.
6. Paesmans M et al. Primary tumor standardized uptake value measured on fluorodeoxyglucose positron emission tomography is of prognostic value for survival in non-small cell lung cancer: update of a systematic review and meta-analysis by the European Lung Cancer Working Party for the International Association for the Study of Lung Cancer Staging Project. J Thorac Oncol 2010;5:612–9.
7. Aerts HJWL et al. Disparity between in vivo EGFR expression and 89Zr-labeled cetuximab uptake assessed with PET. J Nucl Med 2009;50:123–31.
8. Li H et al.. Evaluation of computer-aided diagnosis on a large clinical full-field digital mammographic dataset. Acad Radiol 2008;15:1437–45.
9. Van Meter T et al. Microarray analysis of MRI-defined tissue samples in glioblastoma reveals differences in regional expression of therapeutic targets. Diagn Mol Pathol 2006;15:195–205.
10. Hobbs SK et al.. Magnetic resonance image-guided proteomics of human glioblastoma multiforme. J Magn Reson Imaging 2003;18:530–6.
11. Kuo MD, Gollub J, Sirlin CB, Ooi C, Chen X. Radiogenomic analysis to identify imaging phenotypes associated with drug response gene expression programs in hepatocellular carcinoma. J Vasc Interv Radiol 2007;18:821–31.
12. Rutman AM, Kuo MD. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. Eur J Radiol 2009;70:232–41.
13. O'Connor JPB et al. Enhancing fraction predicts clinical outcome following first-line chemotherapy in patients with epithelial ovarian carcinoma. Clin Cancer Res 2007;13:6130–5.
14. Dehing-Oberije C et al. Development and validation of a prognostic model using blood biomarker information for prediction of survival of non-small-cell lung cancer patients treated with combined chemotherapy and radiation or radiotherapy alone (NCT00181519, NCT00573040, and NCT00572325). Int J Radiat Oncol Biol Phys 2010. doi:10.1016/j.ijrobp.2010.06.011.
15. El Naqa I et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognit 2009;42:1162–71.
16. Tixier F et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. J Nucl Med 2011;52:369–78.
17. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol 2010;49:1012–6.
18. Campbell PJ et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature 2010;467:1109–13.
19. Jackson A, O'Connor JPB, Parker GJM, Jayson GC. Imaging tumor vascular heterogeneity and angiogenesis using dynamic contrast-enhanced magnetic resonance imaging. Clin Cancer Res 2007;13:3449–59.
20. Diehn M et al. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. Proc Natl Acad Sci U S A 2008;105:5213–8.
21. Segal E et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nat Biotechnol 2007;25:675–80.10.

CHAPTER 7

A semiautomatic CT-based ensemble segmentation of lung tumors: Comparison with oncologists' delineations and with the surgical specimen

Published in: Radiotherapy and Oncology 105 (2012) 167–173

A semiautomatic CT-based ensemble segmentation of lung tumors: Comparison with oncologists' delineations and with the surgical specimen

Emmanuel Rios Velazquez, Hugo J.W.L. Aerts, Yuhua Gu, Dmitry B. Goldgof, Dirk De Ruysscher, Andre Dekker, René Korn, Robert J. Gillies, Philippe Lambin

ABSTRACT

Purpose

To assess the clinical relevance of a semiautomatic CT-based ensemble segmentation method, by comparing it to pathology and to CT/PET manual delineations by five independent radiation oncologists in non-small cell lung cancer (NSCLC).

Methods

For 20 NSCLC patients (stages Ib–IIIb) the primary tumor was delineated manually on CT/PET scans by five independent radiation oncologists and segmented using a CT based semi-automatic tool. Tumor volume and overlap fractions between manual and semiautomatic-segmented volumes were compared. All measurements were correlated with the maximal diameter on macroscopic examination of the surgical specimen. Imaging data are available on www.cancerdata.org.

Results

High overlap fractions were observed between the semi-automatically segmented volumes and the intersection (92.5 ± 9.0, mean ± SD) and union (94.2 ± 6.8) of the manual delineations. No statistically significant differences in tumor volume were observed between the semiautomatic segmentation (71.4 ± 83.2 cm^3, mean ± SD) and manual delineations (81.9 ± 94.1 cm^3; p = 0.57). The maximal tumor diameter of the semiautomatic-segmented tumor correlated strongly with the macroscopic diameter of the primary tumor (r = 0.96).

Conclusions

Semiautomatic segmentation of the primary tumor on CT demonstrated high agreement with CT/PET manual delineations and strongly correlated with the macroscopic diameter considered as the ‘‘gold standard’’. This method may be used routinely in clinical practice and could be employed as a starting point for treatment planning, target definition in multi-center clinical trials or for high throughput data mining research. This method is particularly suitable for peripherally located tumors.

INTRODUCTION

Lung cancer is the deadliest type of cancer worldwide [1]. About 80% of the lung cancer patients present advanced-stage disease (stages III and IV) and are considered inoperable due to loco-regional tumor extension, extra thoracic spread or poor physical condition at the time of diagnosis [2]. For these patients, external beam radiotherapy (RT) often combined with chemotherapy is the primary treatment modality [3].

The success of radiotherapy depends upon a good target definition and dose coverage of the target volume while limiting the radiation dose to highly radiosensitive surrounding organs. A consistent and accurate target definition is of utmost importance for accurate radiotherapy treatment planning and for treatment response evaluation. Multiple studies have reported the uncertainties and high – intra and inter – observer variability associated with target delineation in lung cancers [4–10]. Efforts have been made to reduce the observer variation for target definition, including standardized delineation protocols on CT scans and the addition of fused FDG-PET-CT information on the delineation process [11–14]. The latter has diminished the inter-observer variability [15], however differences among observers are still observed for visual delineations [16]. Various PET-based methods have been developed for semiautomatic tumor delineations, ranging from simple fixed (absolute and relative) threshold based segmentations, to the more complex signal-to-background ratio and watershed clustering methods [17,18]. A few studies have compared FDG PET based automatic segmentation tools with pathological examinations [19,20] and demonstrated their utility in reducing inter-observer variability [16,21]. To our knowledge, no CT-based semiautomatic delineation method has been compared with both oncologists' manual delineations and with pathological examination. In practice, target volume and organs at risk are generally defined on a planning CT scan [22], which remains as the reference imaging modality in the treatment planning of non-small cell lung cancer (NSCLC).

Given the observed variability, complexity and time required for target definition, a semi-automated method to accurately segment lung tumors on a CT scan would be of clinical value by providing a consistent initial target definition and would optimize the daily workflow. In this study we present a CT-based region growing method to semi-automatically segment lung tumors, that incorporates expert knowledge and is based on the cognition network technology [23].

Our aim is to evaluate the potential clinical usefulness of a CT-based semiautomatic-segmentation method, by comparing it with CT-PET manual delineations of five independent radiation oncologists and with the pathological examination of the surgical specimen.

MATERIAL AND METHODS

CT-PET scans

This study was approved by the local Medical Ethics Committee according to the Dutch law, and all patients provided written informed consent. Twenty consecutive patients with histologically proven non-small cell lung cancer, stages Ib–IIIb, were included in this retrospective study. All patients had undergone a diagnostic whole body PET-CT scan (Biograph, SOMATOM Sensation 16 with an ECAT ACCEL PET scanner; Siemens, Erlangen, Germany). Patients were instructed to fast at least 6 h before the intravenous administration of ^{18}F-fluoro-2-deoxy-glucose (FDG) (MDS Nordion, Liège, Belgium), followed by physiologic saline (10 mL). The total injected activity of FDG was dependent on the patient weight expressed in kg: (weight ⁄ 4) + 20 Mbq. After a period of 45 min, during which the patient was encouraged to rest, free-breathing PET and CT images were acquired. The CT scan was a spiral CT scan of the whole thorax with intravenous contrast. The PET images were acquired in 5-min bed positions. The CT data set was used for attenuation correction of PET images. The complete data set was then reconstructed iteratively with a reconstruction increment of 5 mm.

Click and grow auto-segmentation algorithm

Pre-processing

The algorithm was developed using the Cognition Network Language (CNL) running on Definiens Developer XD with the LuTA extension [24]. CNL is, an object-based procedural computer language designed to allow automated implementation of complex, context-dependant image analysis tasks. This algorithm was designed to enable an accurate and efficient analysis of lung lesions guided by an operator. The implementation here used the original algorithm previously described [23], and was extended with a multi- seed point segmentation routine. In a first step the LuTA preprocessing was used to classify context objects like body, background, lungs and bones. They were segmented based on intensity and object size without user interaction. In a second optional step, LuTA allows an optional three-dimensional semi-automated correction of the lung boundary. The algorithm workflow is summarized in the Fig. 1S.

Click and grow

The third step involved the identification and segmentation of the lung lesion. The user identified the tumor, within the segmented lung and placed a seed-point at the perceived center of the lesion. From this starting seed-point an initial seed object was automatically segmented using the LuTA region growing [23].

To improve the lesion segmentation and to reduce sensitivity towards the location of the initial seed-point, the original click-and-grow algorithm was further extended with a single click ensemble segmentation algorithm (SCES) [25]. SCES used the previously defined region, within which multiple seed points were automatically generated. Briefly, the initially segmented tumor was divided into eight regions using three perpendicular planes (xy, yz and zx), within each sub-region a seed-point was placed.

Two additional seed-points were placed, one at the center of mass of the segmented tumor and one more randomly. Each seed-point was grown into a new candidate tumor region using the same LuTA region growing algorithm (multiple runs using same segmentation technique but different initial seed-point). Finally the candidate regions were merged into one consensus tumor region using a voting strategy: a voxel is classified as tumor voxel if more than half of the voxels in its 3 x 3 x 3 neighborhood window were labeled as tumor voxels in at least half of the segmented candidate tumor regions. This approach reduced inter-observer variability and operator interactions compared to the original algorithm [25].

GTV manual delineations

For comparison with the CT-based semiautomatic-segmentation, five observers independently carried out manual GTV delineations based on fused PET-CT images using a standard clinical delineation protocol. Briefly, the protocol included fixed window level settings of both CT (lung W 1,700; L –300, mediastinum W 600; L 40) and PET scan (W 30,000; L 15,000) to be used for delineation [16,26,27]. All observers were blinded to each others delineations.

The primary gross tumor volume (GTV) was defined for each patient based on combined CT and PET information. Observers were given transversal, coronal, sagittal and 3D views simultaneously. Delineations were performed on a treatment planning system (XiO; Computer Medical System, Inc., St. Louis, MO).

Pathology

The surgical specimen was examined according to national guidelines [28]. All patients underwent a surgical resection of their lung tumor and a standardized routine pathology examination was performed directly on the fresh specimen maintained on ice within 30 min after resection. Before slicing, the maximal diameter of the primary tumor was measured by macroscopic examination. The interval time between the CT scan and the surgery or biopsy was in average 39 days (range: 7–112).

Statistical analysis

In a similarity analysis, an intersection volume (agreement between all observers) and a union volume (merging of all regions delineated by all observers) were defined and used

for comparison with the semi automatically-segmented volume (Fig. 1). The overlap fraction was used to estimate the agreement between the various volumes and was d fined as the volume of overlap divided by the smallest volume [29,30]:

$$Overlap\ fraction = \left(\frac{A \cap B}{Smallest\ volume}\right) * 100$$

where A is the semi automatically-segmented tumor and B is either the observers intersection volume or the observers union volume depending on the comparison. An overlap fraction equal to 100 indicates two perfectly matched volumes while an overlap fraction equal to zero indicating two disjoined volumes. The first overlap fraction value indicates whether the semiautomatic-segmentation method covers the common agreement (intersection volume) of the manual delineations while the second overlap fraction value indicates whether the algorithm falls within the inter-observer variability (union volume).

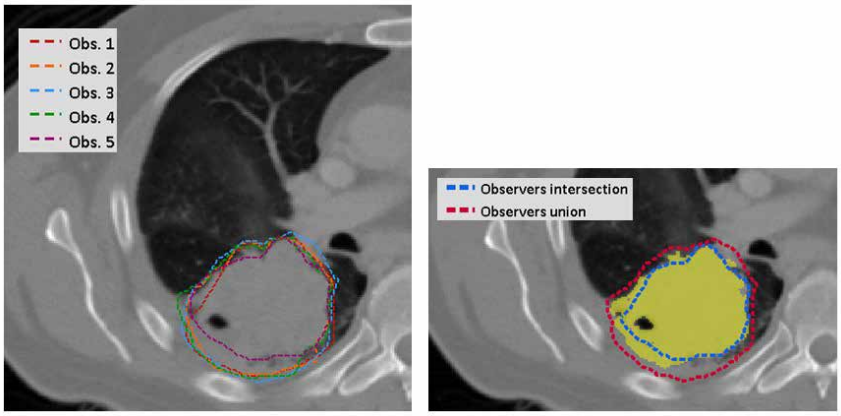


Figure 1

The image in the left side shows the variability observed for CT/PET manual delineations. To summarize inter-observer variability the observers intersection (common agreement) and observers union (sum of all delineated areas) were defined and compared with the CT semi auto segmentation method (yellow color wash) in the right panel. These images correspond to patient 12.

A volume comparison was conducted with the raw volumes expressed in cm^3. Results are summarized as the mean and standard deviations. Groups were compared with a paired Student *t*-test. Differences were considered to be significant when the p-value was lower than 0.05. Pearson's correlation coefficient was used to compare the maximal diameter estimates from pathology with the maximal diameter of the semi automatically-segmented volumes. Additionally, we used the Bland–Altman analysis to evaluate the agreement between the various measurements. The Bland–Altman plot is a scatter plot that shows in the vertical axis the difference between two measures (Y–X), against their

average on the horizontal axis $\left(\frac{Y+X}{2}\right)$. Horizontal lines are superimposed in the scatter plot indicating the mean difference between the measurements and the 95% limits of agreement. If the tumor size measurements are comparable with the ''gold standard'', the differences in the Bland–Altman plot should be small and close to zero. A negative value means that the semi-automatic segmentation overestimates the macroscopic tumor diameter, while a positive value indicates that the semiautomatic segmentation underestimates the gold standard. All data are expressed as mean ± SD. All the analyses were performed in Matlab 2010b (The MathWorks Inc., Natick, MA, USA).

RESULTS

To evaluate clinical validity, semiautomatic-segmentation of the primary tumor on CT was compared with CT/PET manual delineations of five independent observers. A similarity analysis was performed and the overlap fraction was calculated to estimate the agreement between the manually contoured and SCES volumes.

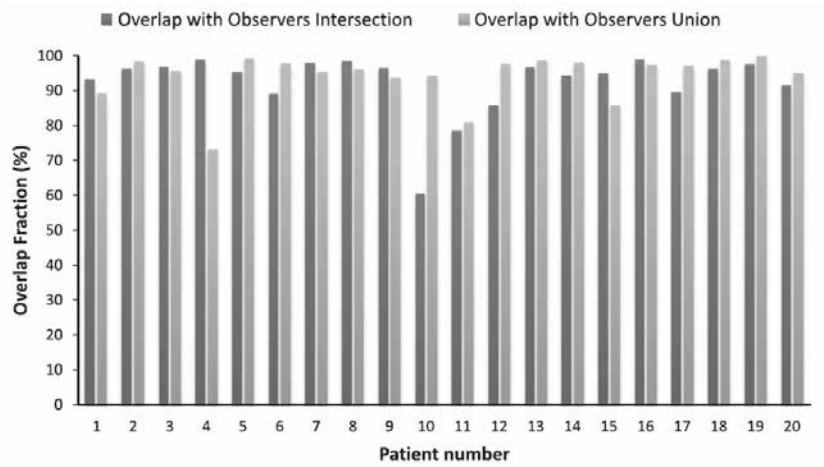


Figure 2

Overlap fractions between the semi auto-segmented volumes and observers' intersection (agreement between all observers) and union (merging of all regions delineated by all observers) volumes. An overlap fraction equal to 100 indicates two perfectly matched volumes while an overlap fraction equal to zero indicates two disjoined volumes.

The overlap fractions between the semi automatically-segmented volumes and the observers union and intersection are shown in Fig. 2.

High overlap fractions were obtained with the observers' intersection (92.5 ± 9.0) and the observers union (94.2 ± 6.8).

The raw semi automatically-segmented and manual volumes are reported in Table 1, mean and standard deviation of the manual volumes are shown as well. No statistical differences were observed in tumor volume between the SCES volumes (71.4 ± 83.2 cm^3) and manual delineations (81.9 ± 94.1 cm^3; p = 0.57). In the majority of the cases the semi automatically-segmented volumes fell within the observers' variability, i.e. 75% of the cases were included in the mean ± 1SD range of the manually delineated volumes.

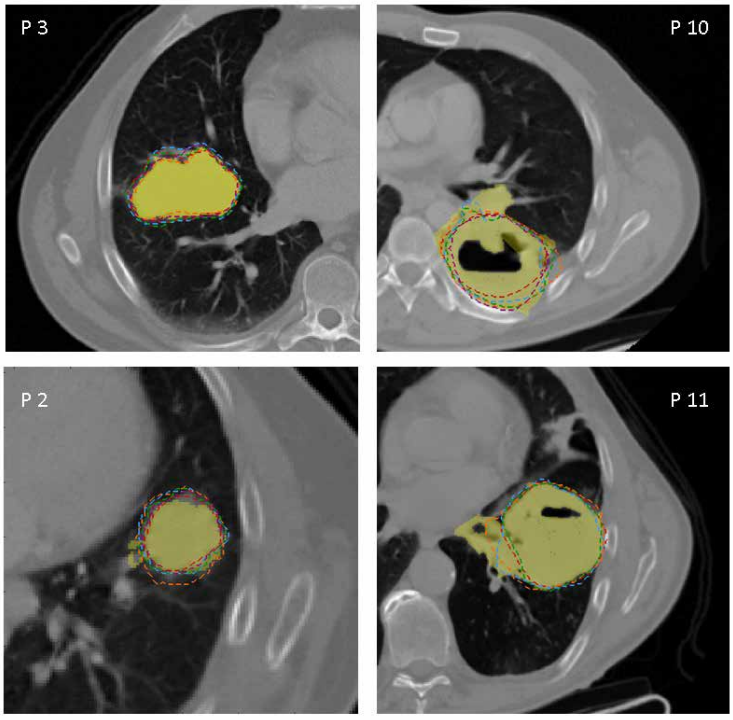


Figure 3

Representative CT images of NSCLC patients. Lung tumors were segmented using a click and grow ensemble segmentation algorithm (yellow solid color wash) and manually delineated by five independent observers (color dotted lines).

Three cases fell off the observers' variability, two of which were centrally located tumors and one peripherally located. For visual comparison, representative examples of both the semi-automatically segmented volumes and manual delineations are shown in Fig. 3. To further evaluate its usability, the click and grow semiautomatic- segmentation algorithm was compared with macroscopic examination of the surgical specimen. A strong correlation was found between the maximal diameter of the SCES volumes and the macroscopic diameter of the primary tumors (Pearson correlation coefficient, 0.96) (Fig 4). The correlation of maximal diameters on manual CT/PET delineations with the pathology examination ranged from 0.88 to 0.96 (0.93 ± 0.02) for different observers.

Table 1: Raw volumes in cm^3 as determined by CT/PET manual delineations and compared with the CT semi auto-segmentation method.

Patient No.	Tumor stage	Location	Observer 1	Observer 2	Observer 3	Observer 4	Observer 5	Mean observers	SD observers	Semi auto-segmentation
1	IIIa	Central	69.8	79.4	103.9	80.3	69.5	80.6	14.0	83.9
2	IIIa	Peripheral	8.0	17.6	13.1	11.9	8.5	11.8	3.9	9.0
3	IIIa	Central	355.0	320.4	380.4	353.1	309.5	343.7	28.6	334.6
4	IIIa	Peripheral	3.1	4.9	4.8	4.2	3.2	4.1	0.9	5.8
5	IIIa	Peripheral	10.6	16.1	21.2	15.9	10.2	14.8	4.5	10.4
6	IIIa	Central	188.1	219.0	275.2	206.8	150.7	208.0	45.6	163.9
7	Ib	Peripheral	46.1	59.2	57.6	57.1	48.4	53.7	6.0	50.1
8	IIIa	Peripheral	18.0	35.5	43.2	25.5	16.8	27.8	11.4	26.2
9	IIIa	Central	29.8	31.8	37.9	34.5	28.5	32.5	3.8	31.6
10	IIIa	Central	196.4	266.8	215.4	259.6	175.1	222.7	39.7	146.1
11	IIIa	Central	151.2	184.2	184.6	164.0	144.3	165.7	18.5	150.7
12	IIIa	Peripheral	199.1	193.5	193.4	194.9	205.1	197.2	5.0	158.2
13	IIIa	Peripheral	55.9	77.9	65.6	68.5	56.8	65.0	9.1	58.0
14	IIIa	Peripheral	6.7	9.4	12.2	9.3	9.1	9.3	1.9	7.9
15	IIIa	Central	18.2	21.9	12.5	14.5	12.5	15.9	4.1	19.6
16	IIIa	Peripheral	38.5	40.9	24.2	36.3	34.3	34.8	6.5	36.7
17	IIIa	Central	31.0	37.1	39.9	36.6	31.7	35.3	3.8	29.5
18	Ib	Peripheral	82.2	101.9	98.9	97.9	92.4	94.7	7.8	88.2
19	IIb	Peripheral	5.7	5.1	1.7	4.8	4.0	4.3	1.6	2.6
20	IIIa	Peripheral	12.2	14.0	30.3	19.1	9.8	17.1	8.1	14.5

Tumor diameters ranged from 1.8 to 9.0 cm on pathological examination (4.2 ± 1.9), from 1.7 to 9.8 (5.1 ± 1.7) on manual CT/ PET delineations and from 2.0 to 9.1 (4.9 ± 2.2) on CT semi automatically-segmented tumors.

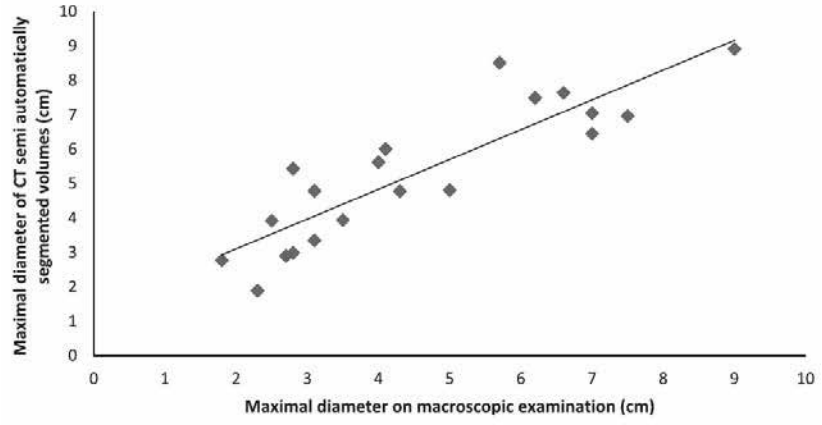


Figure 4

Maximal diameter of the primary tumor determined by the CT semi auto-segmentation algorithm compared with the tumor maximal diameter on macroscopic examination of the surgical specimen. Pearson correlation coefficient was 0.96.

Comparative Bland–Altman plots are shown in Fig. 5, the CT semiautomatic-segmentation algorithm slightly overestimated the maximal tumor diameter found at macroscopic examination, indicated by a mean difference of -0.46 cm, (95% CI, -1.64 to 0.70 cm). However, for manual CT/PET delineations, the differences were slightly larger (mean difference -0.80 cm; 95% CI, -2.42 to 0.82 cm) compared to semi automatically-segmented volumes. For the majority of the patients, the difference between the macroscopic diameter and the diameters on the semi automatically-segmented volumes was within 1 cm.

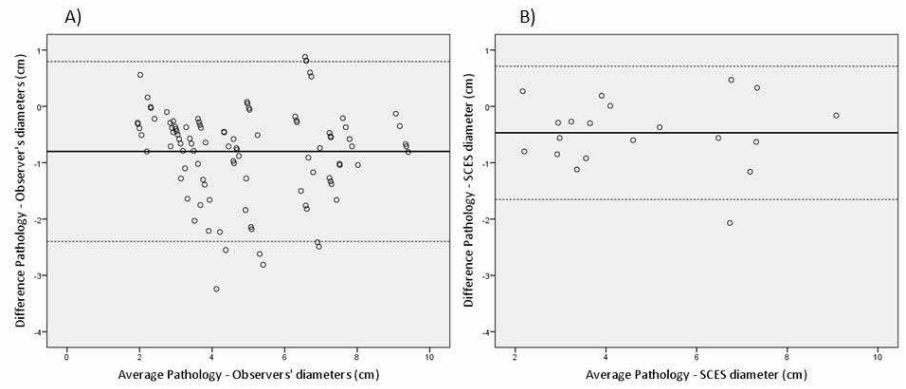


Figure 5

Bland Altman plots showing the discrepancies of CT/PET manual delineations (A) and CT semi auto-segmentations (B) with the 'golden truth' maximal diameter on pathology. The dotted lines represent the confidence intervals around the mean difference (solid line).

DISCUSSION

Target definition remains highly dependent on human interpretation of visual imaging information, making it error prone and subjective. Uncertainties associated with target definition have been largely reported, and especially for lung cancer, high intra and inter observer variability have been observed [4–10]. In practice, human interaction, regarding CT/PET visual interpretation of imaging information and tuning of parameters in complex auto segmentation algorithms still remains as the largest source of uncertainty for target definition [5,7].

To our knowledge, this is the first study that compares a semi-automated method to segment lung tumors on CT images, with multiple manual delineations used routinely in the clinic and furthermore compared to pathology. CT-Based semiautomatic ensemble segmentation of lung tumors showed high agreement with radiation oncologist's manual CT/PET delineations and correlated with pathological tumor measurements (Pearson's correlation, 0.96).

Multiple methods have been developed aiming to improve target definition [17–20]. For example, in the case of PET imaging, the simplest method uses an absolute threshold of the standardized uptake value. Single threshold methods are subject to considerable variation due to heterogeneities in tumor size and FDG-uptake and lack of standardization in PET acquisition settings [31]. More sophisticated segmentation algorithms such as the individualized threshold based on the source-to-background ratio [16,32], the gradient based delineation or the watershed clustering have been proposed, and are generally preferred over single threshold methods [18,20,33]. These studies showed the large differences on the resulting volumes depending on the delineation method. Recently, Wanet et al. compared different automatic delineation methods, including the gradient based method, the source-to-background based method and fixed thresholds at 40% and 50% of the SUV_{max}, with manually delineated contours on the macroscopic specimen and on CT images [20]. Although this method was compared with a three-dimensional reconstruction of the macroscopic specimen, only 10 patients were evaluated in the study. Despite these efforts, a consensus over the most reliable method for automatic delineation of the GTV based on PET is lacking.

CT remains as the reference imaging modality for the treatment planning in NSCLC [22], it is widely available and the image acquisition protocols are better standardized across institutions compared to PET. Thus, we believe a CT based semiautomatic method to segment lung tumors that limits user interaction has clinical value.

In this regard, automatic methods to interpret chest CT images have largely focused on the early detection of lung nodules [34,35], on the differential diagnosis of malignant versus benign nodules [36] and in the measurement of nodule or tumor size as treatment response criteria [37,38]. This study however, concentrates on the definition of locally advanced stage primary tumors, with the exception of two stage Ib patients, which tend to be large, irregular masses often adjacent to other anatomical structures, with the minimal user guidance.

Other studies have used supervised approaches using different machine learning techniques to classify lung tissues [39,40], however these techniques are trained with manually delineated regions of interest, rather than performing automatic detection and were not adequately validated. A general disadvantage of the supervised approach is that it requires beforehand labelled samples used for training of the algorithm.

Kakar et al. proposed an automatic method to classify lung lesions and healthy tissue on CT images, being able to classify the tissue in twelve different categories, differing mainly based on location (i.e. left lung upper, left lung lower) rather than on tissue type [41]. Although fairly accurate classification values were reported they failed to validate their approach since the segmentations were not adequately compared with a ground truth estimation. Furthermore, they artificially generated 500 samples by under or over sampling the original data to develop and evaluate their method, which questions the validity of the testing data.

Our study allowed the possibility of comparing our results with examination of the surgical specimen, by including surgical patients. However, this can also bias the algorithm performance towards peripherally located tumors. In fact, in 75% of the cases, of which 40% were centrally located, semiautomatic-segmentation of tumors on CT images proved to be adequate and could probably be further extended to late stage, inoperable patients treated with radiotherapy. However, in 25% of the cases, the algorithm showed reduced overlap with the observer's delineation. Two of the cases were centrally located tumors, these cases had large central cavities that were covered by the observer's delineation (shown in Fig. 4), but that the algorithm judges as not being part of the tumor.

This largely explains the apparent mismatch between the observer's delineations and the semi auto-segmented volumes. Furthermore, the centrally located tumors also displayed larger observer's variability. In other tumors with central location, the algorithm showed overlap fractions of approx. 95%. An additional case with lower overlap fractions was a small isolated tumor in the upper lobe of the left lung. In this case the algorithm overestimated the tumor extension, presumably by extending the tumor voxels to an adjacent bronchiole. In these situations, supervision is warranted.

Our results showed a high correlation with the maximal diameter of the surgical specimen measured on macroscopic examination. However, there are limitations intrinsic to the method employed to determine the pathological tumor diameter that should be addressed. To date only a few studies have validated auto-delineation methods with the gold standard of pathology [16,20,42]. Daisne et al. and Stroom et al. have proposed techniques to obtain a three-dimensional digital reconstruction of the pathological specimen, by fixating, slicing and photographing the surgical specimen under controlled conditions [42,43]. In our study, the maximal diameter on pathology was determined with a ruler in one dimension, a fairly simple measurement prone to tumor shrinkage and deformation. Measurements were performed before fixation which reduces the influence of tumor shrinkage.

In future work, three-dimensional assessment of the pathological specimen is warranted. Of importance is the time interval between the imaging study and the date of surgery or biopsy, which was in average 39 days (range: 7–112 days). This time interval could impact the results in the case of tumor growth. This relatively large time span could not be avoided as those were the treatment schedules.

However, if large changes in tumor growth occurred between the day of the imaging study and the day of the surgical intervention, these would have been observed in the Bland–Altman analysis.

This analysis showed that the difference between the macroscopic diameter and the diameters on the semi automatically-segmented volumes was within 1 cm in 85% of the cases. We believe this has clinical value in terms that the semi-auto segmented volumes fairly compare to those delineated by the clinical experts and the proposed method agrees with the gold standard at least as good as the medical experts as demonstrated by the Bland– Altman plot.

Because of the retrospective nature of this study, we lacked thin-slice reconstruction CT images; slice thickness in our data was 5 mm. Thin-slice CT images could enhance the ability to accurately determine the tumor volume.

Besides reducing inter-observer variability, semi-automated segmentation of tumors has proven useful to reduce the target delineation time in other tumor sites [44–46]. The time needed to manually delineate the tumors was not recorded, however, in a multicentric study Steenbakkers reported a mean delineation time of 16 min (SD 10 min) on CT scans and of 12 min (SD 8 min) on CT/PET scans [15]. The time needed to segment the lung tumor with the semiautomatic software, including importing and loading the CT data, pre-processing, initial seed-point definition and ensemble segmentation was in average 12.02 min (SD 0.4 min).

Finally, this method could be relevant for adaptive radiotherapy, where treatment plans are modified under restricted time slots, and in the first instance could be used as a pre-delineated structure that the clinical expert could modify. Since this algorithm fairly compared with the standardized delineations that are the current clinical practice, we view this algorithm as an approximation of the target volume that could facilitate the target definition step in radiotherapy treatment planning, but not as a replacement of the medical experts.

To conclude, this method could be employed for target definition routinely on clinical practice, as an approximation that can be refined by the medical experts, and in high-throughput data mining research, based on the overlap with manual delineations and its correlation with the pathological examination, in particular for peripherally located tumors. Further validation will require a multi-center investigation.

ACKNOWLEDGEMENTS

This study was supported by the Dutch Cancer Society (KWF UM-2010-4776) and the National Institutes of Health (NIH-USA U01 CA 143062-01, Radiomics of NSCLC). The authors wish to acknowledge Definiens AG, for the algorithm development and support. We also acknowledge financial support from the QuICConCePT project, partly funded by EFPIA companies and the Innovative Medicine Initiative Joint Undertaking (IMI JU) under Grant Agreement No. 115151.

REFERENCES

1. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics. CA Cancer J Clin 2010;60:277–300.
2. Auperin A, Le Pechoux C, Rolland E, et al. Meta-analysis of concomitant versus sequential radiochemotherapy in locally advanced non-small-cell lung cancer. J Clin Oncol 2010;28:2181–2190.

3. Jett JR, Schild SE, Keith RL, Kesler KA. Treatment of non-small cell lung cancer stage IIIB: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 2007;132:266S–76S.
4. Bradley J, Bae K, Choi N, et al. A phase II comparative study of gross tumor volume definition with or without PET/CT fusion in dosimetric planning for non-small-cell lung cancer (NSCLC): primary analysis of Radiation Therapy Oncology Group (RTOG) 0515. Int J Radiat Oncol Biol Phys 2012;82:e431.
5. Giraud P, Elles S, Helfre S, et al. Conformal radiotherapy for lung cancer: different delineation of the gross tumor volume (GTV) by radiologists and radiation oncologists. Radiother Oncol 2002;62:27–36.
6. Grabarz D, Panzarella T, Bezjak A, McLean M, Elder C, Wong RK. Quantifying interobserver variation in target definition in palliative radiotherapy. Int J Radiat Oncol Biol Phys 2011;80:1498–504.
7. Greco C, Rosenzweig K, Cascini GL, Tamburrini O. Current status of PET/CT for tumour volume definition in radiotherapy treatment planning for non-small cell lung cancer (NSCLC). Lung Cancer 2007;57:125–34.
8. Steenbakkers RJ, Duppen JC, Fitton I, et al. Observer variation in target volume delineation of lung cancer related to radiation oncologist-computer interaction: a 'Big Brother' evaluation. Radiother Oncol 2005;77:182–90.
9. Van de Steene J, Linthout N, de Mey J, et al. Definition of gross tumor volume in lung cancer: inter-observer variability. Radiother Oncol 2002;62:37–49.
10. Vorwerk H, Beckmann G, Bremer M, et al. The delineation of target volumes for radiotherapy of lung cancer patients. Radiother Oncol 2009;91:455–60.
11. Senan S, Van Sornsen de Koste J, Samson M, et al. Evaluation of a target contouring protocol for 3D conformal radiotherapy in non-small cell lung cancer. Radiother Oncol 1999;53:247–55.
12. Bowden P, Fisher R, Mac Manus M, et al. Measurement of lung tumor volumes using three-dimensional computer planning software. Int J Radiat Oncol Biol Phys 2002;53:566–73.
13. Caldwell CB, Mah K, Ung YC, et al. Observer variation in contouring gross tumor volume in patients with poorly defined non-small-cell lung tumors on CT: the impact of 18FDG-hybrid PET fusion. Int J Radiat Oncol Biol Phys 2001;51:923–31.
14. Buijsen J, van den Bogaard J, van der Weide H, et al. FDG-PET-CT reduces the interobserver variability in rectal tumor delineation. Radiother Oncol 2012;102:371–6.
15. Steenbakkers RJ, Duppen JC, Fitton I, et al. Reduction of observer variation using matched CT-PET for lung cancer delineation: a three-dimensional analysis. Int J Radiat Oncol Biol Phys 2006;64:435–48.
16. van Baardwijk A, Bosmans G, Boersma L, 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 Biol Phys 2007;68:771–8.
17. Lee JA. Segmentation of positron emission tomography images: some recommendations for target delineation in radiation oncology. Radiother Oncol 2010;96:302–7.
18. Nestle U, Kremp S, Schaefer-Schuler A, et al. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-Small cell lung cancer. J Nucl Med 2005;46:1342–8.
19. Wu K, Ung YC, Hornby J, et al. PET CT thresholds for radiotherapy target definition in non-small-cell lung cancer: how close are we to the pathologic findings? Int J Radiat Oncol Biol Phys 2009;77:699–706.
20. Wanet M, Lee JA, Weynand B, et al. Gradient-based delineation of the primary GTV on FDG-PET in non-small cell lung cancer: a comparison with thresholdbased approaches, CT and surgical specimens. Radiother Oncol 2010;98:117–25.
21. van Loon J, van Baardwijk A, Boersma L, Ollers M, Lambin P, De Ruysscher D. Therapeutic implications of molecular imaging with PET in the combined modality treatment of lung cancer. Cancer Treat Rev 2011;37:331–43.
22. Sonke JJ, Belderbos J. Adaptive radiotherapy for lung cancer. Semin Radiat Oncol 2010;20:94–106.
23. Bendtsen C, Kiotzmann M, Korn R, Mozley PD, Schmidt G, Binnig G. X-ray computed tomography: semiautomated volumetric analysis of late-stage lung tumors as a basis for response assessments. Int J Biomed Imaging 2011;2011:11.

24. Athelogou M, Schmidt G, Schaepe A, Baatz M, Binnig G. Definiens cognition network technology – a novel multimodal image analysis technique for automatic identification and quantification of biological image contents. In: Imaging cellular and molecular biological functions. New York, NY, USA: Springer; 2007.
25. Gu, Y, Kumar, V, Hall, LO, et al. Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach. Pattern Recognition 2012;In press.
26. van Baardwijk A, Bosmans G, Boersma L, et al. Individualized radical radiotherapy of non-small-cell lung cancer based on normal tissue dose constraints: a feasibility study. Int J Radiat Oncol Biol Phys 2008;71:1394–401.
27. van Baardwijk A, Wanders S, Boersma L, et al. Mature results of an individualized radiation dose prescription study based on normal tissue constraints in stages I to III non-small-cell lung cancer. J Clin Oncol 2010;28:1380–6.
28. Richtlijn Niet-kleincellig longcarcinoom: stadiering en behandeling. Vereniging van Integrale Kankercentra 2004. Available from: http:// www.cbo.nl/thema/Richtlijnen/Overzicht-richtlijnen/Oncologie/.
29. Aerts HJ, van Baardwijk AA, Petit SF, et al. Identification of residual metabolicactive areas within individual NSCLC tumours using a pre-radiotherapy (18)fluorodeoxyglucose-PET-CT scan. Radiother Oncol 2009;91:386–92.
30. Hanna GG, Hounsell AR, O'Sullivan JM. Geometrical analysis of radiotherapy target volume delineation: a systematic review of reported comparison methods. Clin Oncol (R Coll Radiol) 2010;22:515–25.
31. Nestle U, Kremp S, Grosu AL. Practical integration of [18F]-FDG-PET and PETCT in the planning of radiotherapy for non-small cell lung cancer (NSCLC): the technical basis, ICRU-target volumes, problems, perspectives. Radiother Oncol 2006;81:209–25.
32. Daisne JF, Sibomana M, Bol A, Doumont T, Lonneux M, Gregoire V. Tridimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. Radiother Oncol 2003;69:247–50.
33. Hatt M, Cheze le Rest C, Descourt P, et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. Int J Radiat Oncol Biol Phys 2010;77:301–8.
34. Dehmeshki J, Amin H, Valdivieso M, Ye X. Segmentation of pulmonary nodules in thoracic CT scans: a region growing approach. IEEE Trans Med Imaging 2008;27:467–80.
35. Armato 3rd SG, Giger ML, MacMahon H. Automated detection of lung nodules in CT scans: preliminary results. Med Phys 2001;28:1552–61.
36. McNitt-Gray MF, Hart EM, Wyckoff N, Sayre JW, Goldin JG, Aberle DR. A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: preliminary results. Med Phys 1999;26:880–8.
37. Nishino M, Guo M, Jackman DM, et al. CT tumor volume measurement in advanced non-small-cell lung cancer: performance characteristics of an emerging clinical tool. Acad Radiol 2011;18:54–62.
38. Yankelevitz DF, Reeves AP, Kostis WJ, Zhao B, Henschke CI. Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation. Radiology 2000;217:251–6.
39. Uchiyama Y, Katsuragawa S, Abe H, et al. Quantitative computerized analysis of diffuse lung disease in high-resolution computed tomography. Med Phys 2003;30:2440–54.
40. Ye X, Beddoe G, Slabaugh G. Automatic graph cut segmentation of lesions in CT using mean shift superpixels. Int J Biomed Imaging 2010:983963.
41. Kakar M, Olsen DR. Automatic segmentation and recognition of lungs and lesion from CT scans of thorax. Comput Med Imaging Graph 2009;33:72–82.
42. Stroom J, Blaauwgeers H, van Baardwijk A, et al. Feasibility of pathologycorrelated lung imaging for accurate target definition of lung tumors. Int J Radiat Oncol Biol Phys 2007;69:267–75.
43. Daisne JF, Duprez T, Weynand B, et al. Tumor volume in pharyngolaryngeal squamous cell carcinoma: comparison at CT, MR imaging, and FDG PET and validation with surgical specimen. Radiology 2004;233:93–100.
44. Strassmann G, Abdellaoui S, Richter D, et al. Atlas-based semiautomatic target volume definition (CTV) for head-and-neck tumors. Int J Radiat Oncol Biol Phys 2010;78:1270–6.

45. Anders LC, Stieler F, Siebenlist K, Schäfer J, Lohr F, Wenz F. Performance of an atlas-based autosegmentation software for delineation of target volumes for radiotherapy of breast and anorectal cancer. Radiother Oncol 2012;102:68–73.
46. Isambert Al, Dhermain Fdr, Bidault Fo, et al. Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. Radiother Oncol 2008;87:93–9.6.

CHAPTER 8

Volumetric CT-based segmentation of NSCLC using 3D-Slicer

Published in: Nature Scientific Reports. 3, 3529; 2013

Volumetric CT-based segmentation of NSCLC using 3D-Slicer

Emmanuel Rios Velazquez, Chintan Parmar*, Mohammed Jermoumi, Raymond H. Mak, Angela van Baardwijk, Fiona M. Fennessy, John H. Lewis, Dirk De Ruysscher, Ron Kikinis, Philippe Lambin & Hugo J. W. L. Aerts*

*These authors contributed equally to this work

ABSTRACT

Accurate volumetric assessment in non-small cell lung cancer (NSCLC) is critical for adequately informing treatments. In this study we assessed the clinical relevance of a semiautomatic computed tomography (CT)-based segmentation method using the competitive region-growing based algorithm, implemented in the free and public available 3D-Slicer software platform. We compared the 3D-Slicer segmented volumes by three independent observers, who segmented the primary tumour of 20 NSCLC patients twice, to manual slice-by-slice delineations of five physicians. Furthermore, we compared all tumour contours to the macroscopic diameter of the tumour in pathology, considered as the ''gold standard''. The 3D-Slicer segmented volumes demonstrated high agreement (overlap fractions>0.90), lower volume variability (p = 0.0003) and smaller uncertainty areas (p = 0.0002), compared to manual slice-by-slice delineations.

Furthermore, 3D-Slicer segmentations showed a strong correlation to pathology (r = 0.89, 95%CI, 0.81– 0.94). Our results show that semiautomatic 3D-Slicer segmentations can be used for accurate contouring and are more stable than manual delineations. Therefore, 3D-Slicer can be employed as a starting point for treatment decisions or for high-throughput data mining research, such as Radiomics, where manual delineating often represent a time-consuming bottleneck.

INTRODUCTION

Lung cancer is a disease that affects about 1.6 million individuals worldwide every year[1]. Non-small cell lung cancer (NSCLC) accounts for 85% of all lung cancer cases and it is characterized by poor prognosis and low survival rates, due to high incidence of loco-regional and distant recurrences[2]. In lung cancer, tumour delineation is critical for accurate volumetric assessment to evaluate response to therapy, which can inform treatment decisions. However, tumour delineation can be a source of uncertainty, since typically, the tumour delineation process involves an experienced physician, interpreting and manually contouring computed tomography (CT) alone or combined with Fluorodeoxyglucose (FDG) - positron emission tomography (PET) imaging, on a slice-by-slice basis[3–6]. Despite efforts in standardization of CT or FDG-PET-CT image acquisition and standardized guidelines for tumour delineation, definition of lung tumours remains prone to inter-observer variability and is time consuming[6–9].

To reduce these problems, a number of CT or FDG-PET based semi-automatic methods have been investigated, that aim to provide equivalent segmentations to those delineated manually by physicians, or to provide a starting point for the manual delineation process, thereby reducing the overall required time. The various segmentation methods, that range from simple threshold based methods to complex level set, watershed, or region growing-context based methods, have been compared to manual delineations provided by physicians and compared to the pathological measurements of tumour size, with varying success rates[10–16]. However, the application of these methods is limited, often due to accessibility of the method within the clinical delineation process.

In this study we evaluated the utility of the GrowCut algorithm to segment lung tumours, implemented in 3DSlicer – a free open source software platform for biomedical research[17]. This cellular automaton-based algorithm performs automatic tumour segmentation after drawing boundaries within the image volume. It provides an alternative to the manual slice-by-slice segmentation process and is found to be significantly faster and less user intensive[17].

Our hypothesis is that 3D-Slicer contours are more stable for inter-observer variation compared to manual contouring. To evaluate the accuracy of the 3D-Slicer segmentations, three independent observers segmented 20 NSCLC patients twice using 3D-Slicer.We compared these six 3D-Slicer segmentations to manual delineations provided by five physicians. Furthermore, the segmented volumes were compared with the maximum diameter measured from the tumour after resection, considered as the gold standard. Because 3D-Slicer is publicly available and easily accessible by download, its application in NSCLC could be useful for the clinical investigations where tumour contours are necessary for assessing therapy response, therapy planning, or in high-throughput data mining research of medical imaging in clinical oncology (Radiomics)[18–21].

RESULTS

Clinical reliability of the 3D slicer's semi-automatic segmentations was measured in terms of its agreement with the CT/PET manual tumour delineations of five independent observers and with pathological measurements after surgery. To quantify the agreement between the manual and 3D-Slicer segmentations, we performed an uncertainty analysis. The uncertainty region was defined as the region that varied between the segmentations of the different observers. In figure 1, the uncertainty region of five manual and six 3D-Slicer segmentations (three observers segmented twice with different seed point initialization) is illustrated. This example shows that the uncertainty region is larger for manual delineations compared to 3D-Slicer.

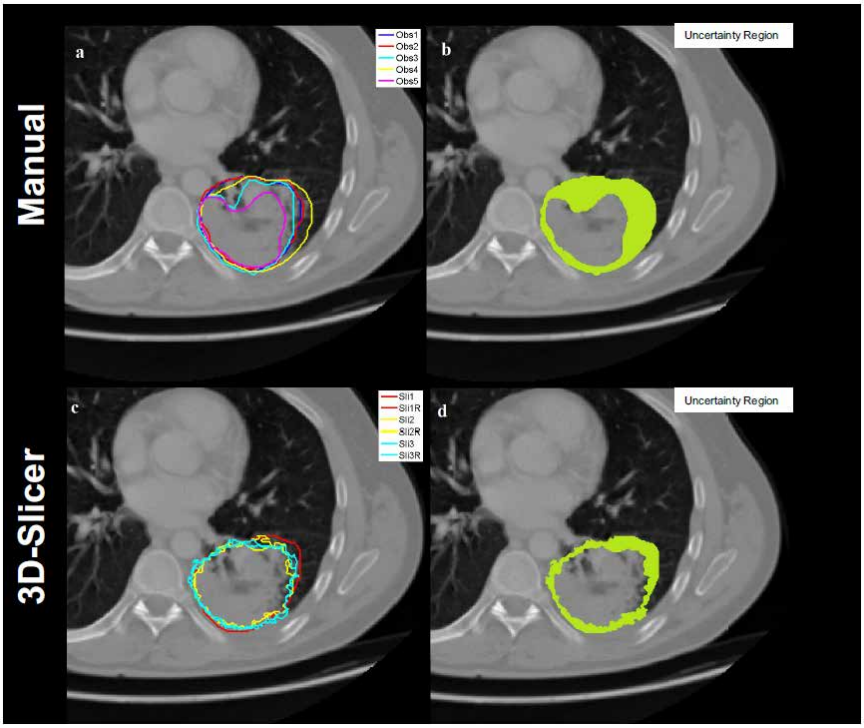


Figure 1. Segmentation uncertainty

Left: representative example showing differences in CT/PET manual delineations (top) and 3D-Slicer segmentations (bottom). Right: This variability is quantified with the uncertainty region, defined as the difference between the observers' agreement and observers' union (highlighted in green). The smaller the uncertainty region is, the lower the variability among multiple contours.

In the Supplementary Figure S1, a heat map depicting the overlap fractions for each patient between the GrowCut segmentations and manual delineations' union and intersection are shown.

The results demonstrate a high spatial agreement of the manual and 3D-Slicer segmentations.

Overlap fractions

To examine the spatial agreement of the manual and 3D-Slicer contours, Overlap Fractions (OF) were calculated. OFs were computed between each of the six 3D-Slicer segmentations with the uncertainty region of the manual delineations. The intersection is defined as the inner boundary of the uncertainty region (i.e. the region that all manual observers delineated), and the union as the outer boundary of the uncertainty region (i.e. the region at least one of the manual observers delineated). High OFs were observed with the observers' intersection (mean±SD: 94.3±4.4%, range: 76.8– 99.8) and union (mean±SD: 97.2±5.1%; range: 72.6–100) [See figure 2].

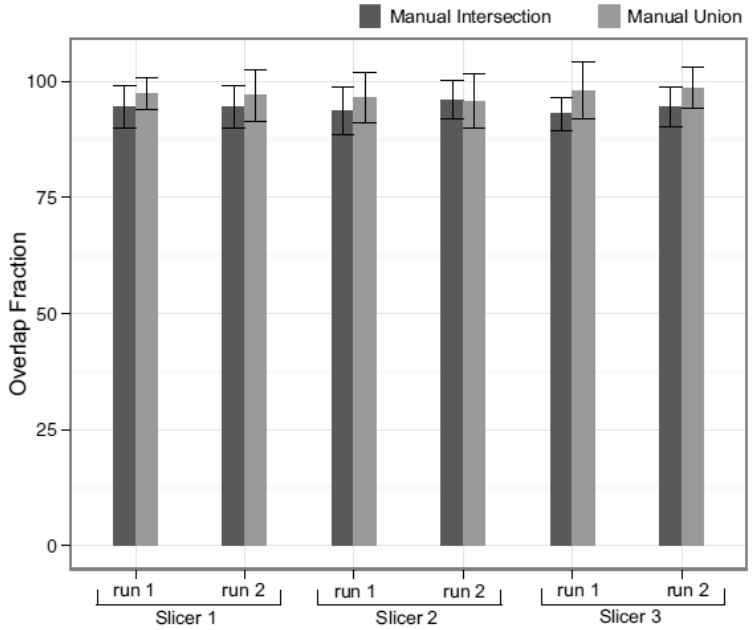


Figure 2. Overlap fractions between the 3D-Slicer segmented volumes and the observers' intersection and union volumes.

High overlap fraction indicates high agreement (spatial overlap) between volumes.

Uncertainty regions

To investigate the robustness of 3D-Slicer segmentations we compared its uncertainty region against the manual uncertainty region [Figure 1]. The analysis showed that the uncertainty region, defined as the difference between uncertainty region inner and outer boundaries, was smaller for the 3D-Slicer segmentations [See Figure 3A]. Manual delinea-

tions had significantly larger uncertainty areas compared to 3D-Slicer segmentations (Wilcoxon test $p = 0.0002$).

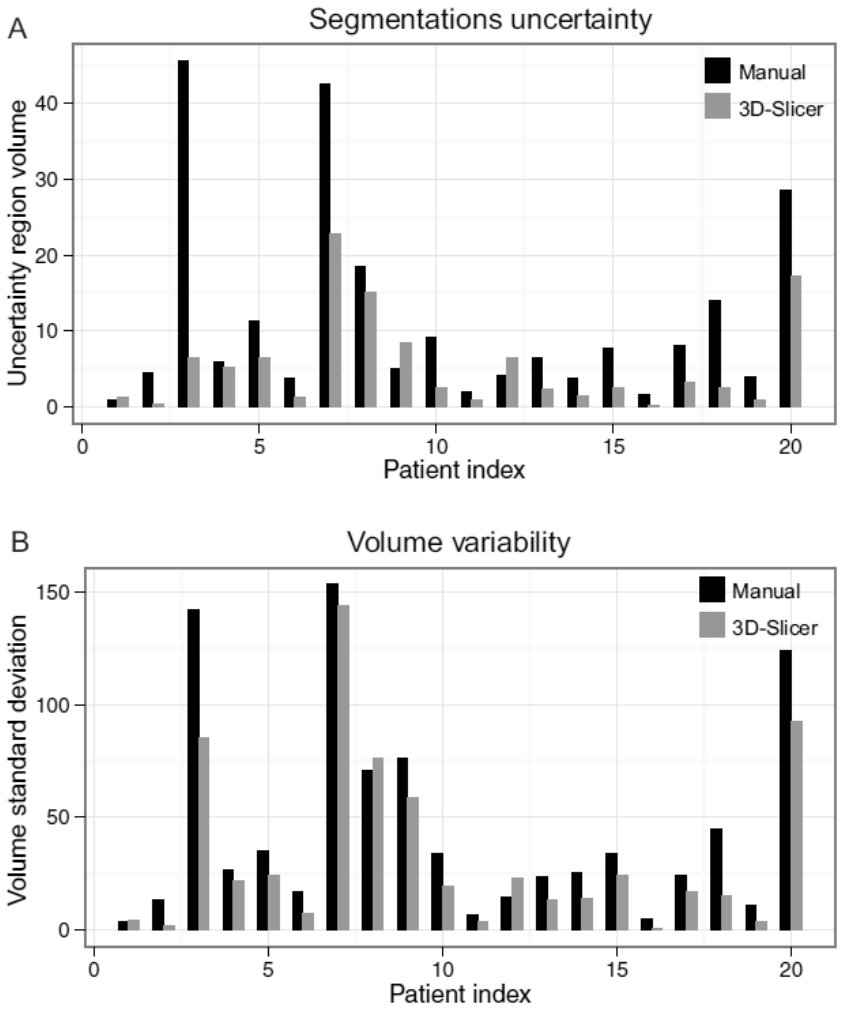


Figure 3.

(A): Comparison of volume uncertainty (as defined as the region that varied between the contours of multiple observers) of manual delineations and 3D-Slicer segmentations. See figure 2 for an illustrative example of the uncertainty region. (B): Comparison of volume variability (cm^3) of observers' manual delineations and 3D-Slicer segmentations.

Segmented volumes

We then investigated the volumes of the segmentations. There was a high agreement between the volumes of the manual and 3D-Slicer contours, as we found no statistically sig-

nificant difference between the volumes of the five manual delineations (82.03 ± 94.31 cm^3) and six 3D-Slicer (72.27 ± 86.62 cm^3, mean ± SD) segmentations, using Kruskal–Wallis one way analysis of variance (p = 0.98). Figure 3B, displays the tumour volume variability, for both manual and 3D-Slicer for all patients. In 17 cases (85%), the volume variability was significantly lower for 3DSlicer segmentations (p = 0.0003).

3D-Slicer segmentation process

To investigate the stability of 3DSlicer algorithm against user seed-points initialization, we compared the intra-observer variability for each of the 3D-Slicer users. High overlap fractions were observed for the 3D-Slicer users: 95.01% ± 5.33%, 94.11% ± 3.95 and 97.08% ± 2.54% [mean ± SD], respectively.

To assess the duration of the 3D-Slicer segmentation process, we recorded the duration of all segmentation phases. The total segmentation times were in average 10.6 min (range: 4.85–18.25 min), 9.97 (range 6.39–13.83 min) and 9.94 min (range: 4.38–20.25 min), for the three 3D-Slicer users respectively. In average, the times measured for each 3D-Slicer segmentation phase were: loading (28 seconds), algorithm initialization (2.79 min), running the 3D-Slicer algorithm (32 seconds) and editing final phase (6.52 min).

Pathology

Further validation was provided by comparing the maximum diameter of the 3D slicer segmentations with that of the surgical specimen. Strong correlations were observed between the maximum diameter of 3D-Slicer volumes and the macroscopic diameter of the surgical tumours (spearman r, mean ± SD 50.89 ± 0.05, range: 0.81–0.94).
Similarly, the maximum diameters of the manual CT/PET delineations were highly correlated with the macroscopic diameter (spearman r, mean ± SD = 0.92 ± 0.02, range: 0.91–0.95).

Figure 4 displays the scatter plot between macroscopic diameter and the diameters of CT segmentations (manual and 3D slicer). The diameters of surgery had a range of 1.8–9 and average of 4.5 ± 2.03 (mean ± SD). The manual delineations had a range of 1.42–12.53 and average of 6.09 ± 2.71 (mean ± SD). The semi-automatic delineations were: range 1.41– 12.20 and average of 6.17 6 2.89. These twelve different diameter vectors were also compared using the Kruskal-Wallis test and no statistically significant difference was observed (p = 0.97).

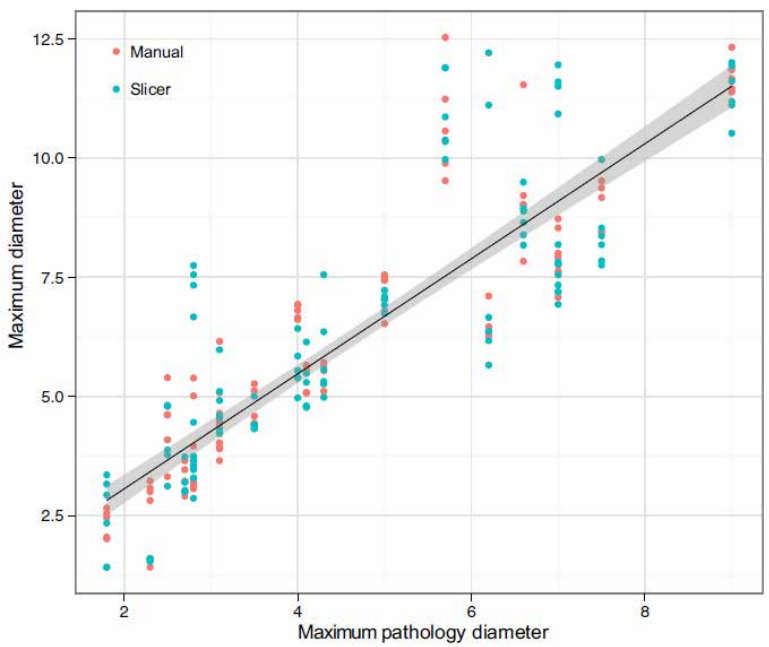


Figure 4. Scatter plot between maximal diameter of surgical specimen and the maximal diameter of computed tomography (CT) segmented volumes for both manual and semiautomatic 3D-Slicer diameters.

Spearman's correlation coefficient was 0.89 (95%CI, 0.81–0.94).

DISCUSSION

Despite the efforts in CT-PET imaging standardization and tumour delineation protocols, target definition remains subjected to observer variation. With respect to manual delineations, the addition of PET information to CT imaging in standardized delineation protocols has reduced the observer variability, however, human interaction and interpretation of medical images is still a considerable source of variation[3,22,23]. Furthermore, slice-by-slice manual contouring of two-dimensional images is a time consuming process. Here, we evaluated the utility of a freely accessible 3D-Slicer algorithm, a cellular automaton-based algorithm, by performing a volumetric comparison with tumour delineations made by five independent oncologists following standardized protocols[24], as well as by comparing it with the maximal diameter obtained from pathological measurements.

The volumetric comparison showed that the 3D-Slicer algorithm provides tumour segmentations, statistically equivalent to physicians CT/PET manual contours. To evaluate the accuracy of the 3D-Slicer segmentations, the overlap fraction (%) was calculated and resulted in high values between the semi-automatically segmented volumes and the intersection (mean±SD: 94.3±4.4%, range: 76.8–99.8) and union (mean ± SD: 97.2 ± 5.1%;

range: 72.6–100) of the manual delineations. Importantly, semi-automatic segmentations showed overall lower volume variability (p=0.0003) and smaller uncertainty areas (p = 0.0002) compared to manual delineations. 3D-Slicer segmentations showed robustness towards user initialization, the OF's between the first Slicer segmentation and the second slicer segmentation were for each user in average: 95.01% ± 5.33%, 94.11% ± 3.95 and 97.08% ± 2.54%, respectively.

Additionally, we observed a strong correlation between the 3DSlicer segmentations and the maximal diameter as measured on pathological examination (r=0.89; 95% CI, 0.81–0.94).

The average time to perform a complete segmentation was 9.8 minutes using Slicer. Loading the images and running the algorithm takes in average half a minute respectively. Due to the retrospective nature of our analysis we were not able to compare the 3D-Slicer segmentation times with the manual delineation times, since those were not available. However 3D slicer's volume segmentation has been shown to be substantially faster and less user intensive compared to manual delineation in other tumour sites[17]. Furthermore, manual delineation is well known to be a very time consuming task.

To minimize observer variability and reduce user interactions, several CT and PET semi-automatic segmentation methods have been introduced. Simple methods such as threshold-based segmentations are widely available but often fail to accurately define the tumour borders[10,11,16]. Various more complex methods have been investigated, including signal-to-background ratio individualized thresholding, watershed-based methods or complex fuzzy locally adaptive thresholding methods[11,14,15,25–27]. These methods have showed generally better correlations with pathology and manual delineations than the simple fixed threshold methods; however they often require significant tuning of algorithm parameters and are not widely available. PET-based methods are intrinsically better choices to segment the highly active metabolic areas of the tumour. In contrast,
CT-based methods provide an anatomical segmentation with higher spatial resolution.

In radiation therapy, CT is the reference imaging modality for treatment planning, and an accurate gross tumour volume definition is fundamental to assure adequate target coverage. Therefore, we believe that CT-based semi-automatic segmentations have clinical utility, if they provide segmentations as accurate as those generated manually by the medical experts, despite the intrinsic CT limitations to distinguish areas of the tumour that are metabolically more active.

Cheebsumon et al, compared several commonly used PET-based segmentation methods with pathology and with a CT manually delineated volume[11]. They reported PET-based methods to have a better agreement with pathology compared to CT delineation. In their study, CT manual delineation significantly overestimated the tumour size compared to pathology. CT manual delineation is known to be prone to inter-observer variation and usually overestimates tumour dimensions. In their exhaustive methods comparison, they lacked a comparison with semi-automatic CT-based segmentation methods, which have shown better correlations with pathology than manual delineations[28].

We previously evaluated a CT-based click-and-grow ensemble segmentation (SCES) algorithm, which showed good overlap with medical expert's tumour delineations and with pathological measurements[28]. The SCES also showed robustness towards user initialization, as it involved an iterative segmentation process, with a bootstrapping routine with multiple initializations, which resulted in highly reproducible final segmentations[29]. Unfortunately, this algorithm is only available in commercial packages and therefore not available for the broader community.

A comparison of CT-based and PET-based methods with pathological measurements and manual delineations is still lacking though. We anticipate that methods combining CT and PET information will be the winner in the lung tumour segmentation race, though not all centers are equipped with integrated PET-CT scanners. However, intrinsic differences between CT and PET information should be taken into account. The present 3D-Slicer algorithm, provided accurate tumour segmentations for 85% of the cases. In three cases the 3D-Slicer failed to define accurately the border, these cases showed larger volume variability with 3D-Slicer compared to manual delineations; two of these cases were large masses with pleural attachment, however only one had a central location. The third case was a very small isolated tumour, adjacent to a main blood vessel, in this case due to the volume size, small variations in border definition due to the adjacent vessel, resulted in significant volume variations. Nevertheless, a medical expert should supervise auto-segmentation algorithms in all cases.

The current correlation between the 3D-Slicer delineation and pathology could possibly be improved if the CT and PET-CT would have been performed in 4D-mode. It is well recognized that a free breathing CT and even more PET scan will result in blurred edges of the tumour and erroneous CT densities or SUV values. In further research, 4D scans should be used. A general drawback when comparing segmentation algorithms with pathological dimensions is that often only tumour sizes in one dimension are available (maximal diameter). Furthermore, pathological measurements can be affected by tumour shrinkage and deformation after surgery. In this study only the maximal diameter on pathology was compared, which is less prone to error than volumetric comparisons with pathology. The timing-span between the image acquisition and surgery may impact the comparison of the segmentation methods with pathology due to tumour growth. Given the correlation observed with pathological tumour diameter, this time difference may not have a strong impact in the evaluated cases.

In conclusion, the open source 3D-Slicer algorithm, provided tumour segmentations comparable to those manually delineated by physicians and with lower variability. Since the semi-automatic segmentations are statistically comparable to manual delineations and correlated well with pathology, they could be used as a starting point for treatment planning delineations and in high-throughput data mining research, such as Radiomics[18–21], where manual tumour delineations are often not available, or represent a considerable time consuming bottleneck.

METHODS

CT-PET scans

CT-PET scans. The imaging data was acquired at MAASTRO Clinic in The Netherlands, as reported previously by Baardwijk et al[7]. In short, twenty consecutive patients with histologically verified non-small cell lung cancer, stage IB-IIIB, were included in this study. All patients received a diagnostic whole body positron emission tomography (PET)-computed tomography (CT) scanning (Biograph, SOMATOM Sensation 16 with an ECAT ACCEL PET scanner; Siemens, Erlangen, Germany). Patients were instructed to fast at least six hours before the intravenous administration of ^{18}F-fluoro-2-deoxy-glucose (FDG) (MDS Nordion, Liege, Belgium), followed by physiologic saline (10 mL). The total injected activity of FDG was dependent on the patient weight expressed in kg: (*weight* * 4) + 20 Mbq. Free breathing PET and CT images were acquired after a period of 45 minutes, during which the patient was encouraged to rest. The whole thorax spiral CT scan was acquired with intravenous contrast. The PET images were obtained in 5-min bed positions. The CT data set was used for attenuation correction of PET images. The complete data set was then reconstructed iteratively with a reconstruction increment of 5 mm. Imaging data are available on www.cancerdata.org. This study was conducted according to national laws and guidelines and approved by the appropriate local trial committee at Maastricht University Medical Center (MUMC1), Maastricht, The Netherlands. For more details see Baardwijk et al[7].

GrowCut semi-automatic segmentation method in 3D-Slicer

GrowCut semi-automatic segmentation method in 3D-Slicer. GrowCut is an interactive region growing segmentation method. Given an initial small set of label points the algorithm automatically segments the remaining image by using cellular automation. The algorithm uses a competitive region growing approach and is considered as having good accuracy and speed for the 2D and 3D image segmentation. For N-class segmentation the algorithm needs N initial sets of pixels (one set corresponding to each class) from user. Using these pixel sets, the algorithm automatically generates the region of interest (ROI), which is the convex hull of the user-labelled pixels with an additional margin. In the next step, it iteratively labels all the pixels in the ROI using the user-given pixel labels. The algorithm converges when all the pixels in the ROI have unchanged labels across several iterations. Pixel labelling is done using a weighted similarity score, which is a function of the neighbouring pixel weights. An unlabelled pixel is labelled corresponding to the neighbouring pixels that have the highest weights.

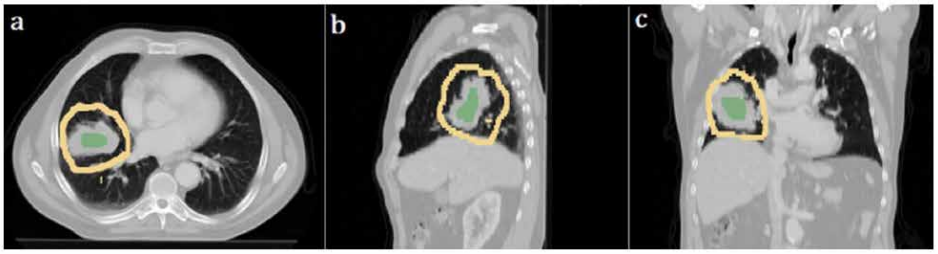


Figure 5. Initialization step of 3D-Slicer segmentation
Marked foreground (green) and background (yellow) are shown. Axial (a), sagittal (b) and coronal (c) views are shown.

NSCLC tumor GrowCut segmentation in 3D-Slicer

3D-Slicer gives a user friendly GUI as the frontend and an efficient algorithm as the back end for the GrowCut segmentation. After loading the patient data, the process began with the initialization of the foreground and background by marking the area inside and outside the tumour region with few initial seed pixels [Figure 5]. The next step was automatic competing region-growing, which segmented the region of interest into foreground and background. Background and surrounding isolated foreground pixels were removed after visual inspection.

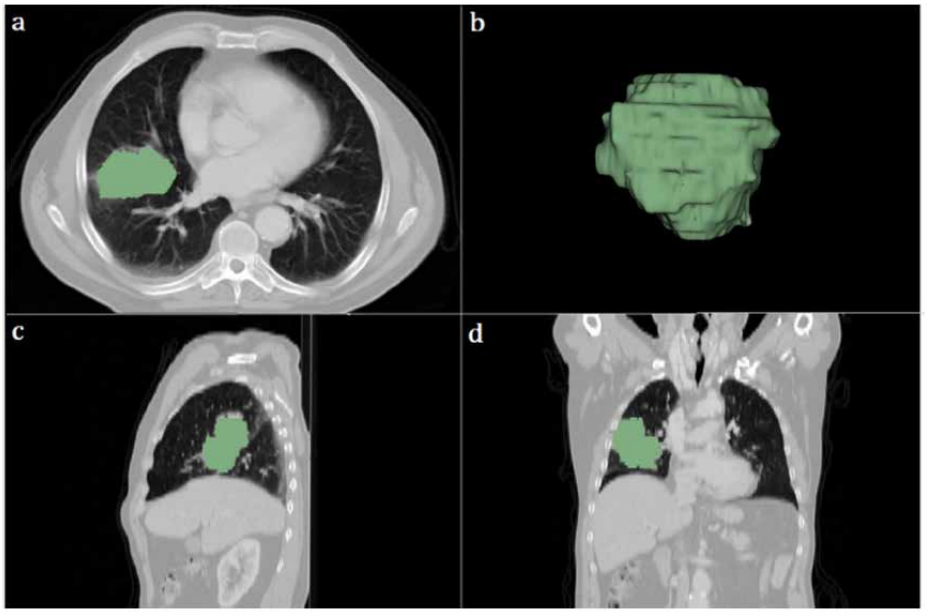


Figure 6. Semi-automatically segmented tumour (green) using 3D-Slicer.
Axial (a), three dimensional (b), sagittal (c) and coronal (d) views are shown.

Figure 6 displays the final segmented tumour region. In Supplementary Figure S2 four representative tumour segmentations generated using the 3D-Slicer algorithm are compared with the manual delineations of five independent observers. Visual comparison shows a high agreement of the manual delineations with the semiautomatic one. We performed Slicer GrowCut segmentations by three independent users, which repeated the process two times, with a three day interval between each time. Segmentation times using GrowCut were recorded for every step of the analysis.

Manual tumor delineations

To validate the semiautomatic segmentation method, five radiation oncologist have manually delineated the gross tumour volume (GTV) of the primary tumour, based on fused PET-CT images using standard delineation protocol, which includes fixed window-level settings of both CT (lung W 1,700; L 2300, mediastinum W600; L 40) and PET scan (W 30,000; L 15,000)[2,7,24]. Radiation oncologists were mutually blind of each other's delineations. The primary GTV was defined for each patient based on combined CT and PET information in the axial plane. The radiation oncologists were given transversal, coronal, sagittal and 3D views simultaneously. A treatment planning system (XiO; Computer Medical System, Inc., St. Louis, MO), was used for performing delineations.

Pathology

The examination of surgical specimen was carried out according to national guidelines[7]. Surgical resections were performed on all the patients. Before slicing, the maximal diameter of the primary tumour was measured by macroscopic examination. The interval time between the CT scan and the surgery or biopsy was in average 39 days (range: 7–112).

Statistical analysis

Overlap Fraction (OF) was used to evaluate the 3D slicer's segmentations in terms of its spatial overlap with manual delineations. Intersection and union volumes were defined for manual delineations (Figure 1). OFs were calculated between the semiautomatic segmentations and these intersection and union delineations. OF was defined as the as the volume of overlap divided by the smallest volume[30]:

$$OFInter = \frac{SV \cap OBi}{\min\{SV, OBi\}} * 100 \text{ and } OFunion = \frac{SV \cap OBu}{\min\{SV, OBu\}} * 100$$

SV, *OBi* and *OBu* are the semiautomatic, observers' intersection and union volumes respectively. OF value of 100 suggests a perfect match while OF value 0 points to two disjoint volumes and thus no match. OF_{inter} indicates whether the semiautomatic-

segmentation method covers the common agreement (intersection volume) of the manual delineations while OF_{union} indicates whether the algorithm falls within the inter-observer variability (union volume). Furthermore, using the above described concept of union and intersection volumes, we calculated and compared the uncertainty of the GrowCut segmentations and the manual delineations. The uncertainty was defined as the difference between the union and intersection volumes, which is the area that belongs to the union but not to the intersection volumes. This region can be seen in Figure 1, highlighted in green. The lower the difference between union and intersection volumes the lower the uncertainty. If all contours were equal, with no variation, the union and intersection volumes would be identical with no uncertainty areas. Overlap fractions were used to compare the first 3D-Slicer segmentation against the second 3D-Slicer segmentation for the same observer.

A volume (cm^3) comparison was also carried out. Volumes calculated from different segmentation methods were compared using the Kruskal-Wallis test. Two methods were considered to be significantly different when the p-value was lower than 0.05.
We compared the volume variability of the 3D-Slicer segmentations against manual delineations using the standard deviation of the 3D-Slicer and manual volumes. The Wilcoxon test was used to compare the volume variability and uncertainty differences between the two types of segmentations.

Spearman correlation coefficient was used to compare the maximal diameter of pathology with the maximal diameter of 3D-Slicer and the manual segmentations.
Further we also compared all these twelve maximal diameter groups: 3D-Slicer (three observers twice), pathology, and five manual using the Kruskal-Wallis one-way analysis of variance. Again groups were considered significantly different when the p value was lower than 0.05. All data are expressed as mean6SD. All the analyses were performed in Matlab (The MathWorks Inc., Natick, MA, USA) and R (R Foundation for Statistical Computing, Vienna, Austria).

ACKNOWLEDGMENTS

Authors acknowledge financial support from the National Institute of Health (NIH-USA U01 CA 143062-01, Radiomics of NSCLC), the CTMM framework (AIRFORCE project, grant 030-103), EU 6th and 7th framework program (METOXIA, EURECA, ARTFORCE), euroCAT (IVA Interreg - www.eurocat.info), Kankeronderzoekfonds Limburg from the Health Foundation Limburg and the Dutch Cancer Society (KWF UM 2011-5020, KWF UM 2009-4454). Authors also acknowledge financial support from the QuIC-ConCePT project (Grant Agreement No. 115151).

REFERENCES

1. Jemal, A. et al. Global cancer statistics. CA Cancer J Clin 61, 69–90 (2011).
2. van Baardwijk, A. et al. Mature results of an individualized radiation dose prescription study based on normal tissue constraints in stages I to III non-smallcell lung cancer. J Clin Oncol 28, 1380–1386 (2010).
3. Steenbakkers, R. J. et al. Observer variation in target volume delineation of lung cancer related to radiation oncologist-computer interaction: a 'Big Brother' evaluation. Radiother Oncol 77, 182–190 (2005).
4. Van de Steene, J. et al. Definition of gross tumor volume in lung cancer: interobserver variability. Radiother Oncol 62, 37–49 (2002).
5. Bowden, P. et al. Measurement of lung tumor volumes using three-dimensional computer planning software. Int J Radiat Oncol Biol Phys 53, 566–573 (2002).
6. Caldwell, C. B. et al. Observer variation in contouring gross tumor volume in patients with poorly defined non-small-cell lung tumors on CT: the impact of 18FDG-hybrid PET fusion. Int J Radiat Oncol Biol Phys 51, 923–931 (2001).
7. van Baardwijk, A. 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 Biol Phys 68, 771–778 (2007).
8. Steenbakkers, R. J. et al. Reduction of observer variation using matched CT-PET for lung cancer delineation: a three-dimensional analysis. Int J Radiat Oncol Biol Phys 64, 435–448 (2006).
9. De Ruysscher, D. PET-CT in radiotherapy for lung cancer. Methods Mol Biol 727, 53–58 (2011).
10. Nestle, U. et al. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-Small cell lung cancer. J Nucl Med 46, 1342–1348 (2005).
11. Cheebsumon, P. et al. Assessment of tumour size in PET/CT lung cancer studies: PET- and CT-based methods compared to pathology. EJNMMI Res 2, 56 (2012).
12. Daisne, J. F. et al. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. Radiother Oncol 69, 247–250 (2003).
13. Dehmeshki, J., Amin, H., Valdivieso, M. & Ye, X. Segmentation of pulmonary nodules in thoracic CT scans: a region growing approach. IEEE Trans Med Imaging 27, 467–480 (2008).
14. Schaefer, A. et al. PET-based delineation of tumour volumes in lung cancer: comparison with pathological findings. Eur J Nucl Med Mol Imaging 40, 1233–1244 (2013).
15. Hatt, M. et al. Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. Int J Radiat Oncol Biol Phys 77, 301–308 (2010).
16. Wu, K. et al. PET CT thresholds for radiotherapy target definition in non-smallcell lung cancer: how close are we to the pathologic findings? Int J Radiat Oncol Biol Phys 77, 699–706 (2009).
17. Egger, J. et al. GBM volumetry using the 3D Slicer medical image computing platform. Sci Rep 3, 1–7 (2013).
18. Kumar, V. et al. Radiomics: the process and the challenges. Magn Reson Imaging 30, 1234–1248 (2012).
19. Lambin, P. et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48, 441–446 (2012).
20. Buckler, A. J., Bresolin, L., Dunnick, N. R. & Sullivan, D. C. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. Radiology 258, 906–914 (2011).
21. Buckler, A. J. et al. Quantitative imaging test approval and biomarker qualification: interrelated but distinct activities. Radiology 259, 875–884 (2011).
22. Greco, C., Rosenzweig, K., Cascini, G. L. & Tamburrini, O. Current status of PET/ CT for tumour volume definition in radiotherapy treatment planning for nonsmall cell lung cancer (NSCLC). Lung Cancer 57, 125–134 (2007).
23. Sonke, J. J. & Belderbos, J. Adaptive radiotherapy for lung cancer. Semin Radiat Oncol 20, 94–106 (2010).
24. van Baardwijk, A. et al. Individualized radical radiotherapy of non-small-cell lung cancer based on normal tissue dose constraints: a feasibility study. Int J Radiat Oncol Biol Phys 71, 1394–1401 (2008).

25. Ye, X., Beddoe, G. & Slabaugh, G. Automatic Graph Cut Segmentation of Lesions in CT Using Mean Shift Superpixels. Int J Biomed Imaging 2010, 1–14 (2010).
26. Daisne, J.-F. o. et al. Tri-dimensional automatic segmentation of PET volumes based on measured source-to-background ratios: influence of reconstruction algorithms. Radiotherapy and Oncology 69, 247–250 (2003).
27. Wanet, M. et al. Gradient-based delineation of the primary GTV on FDG-PET in non-small cell lung cancer: a comparison with threshold-based approaches, CT and surgical specimens. Radiother Oncol 98, 117–125 (2010).
28. Rios Velazquez, E. et al. A semiautomatic CT-based ensemble segmentation of lung tumors: comparison with oncologists' delineations and with the surgical specimen. Radiother Oncol 105, 167–173 (2012).
29. Gu, Y. et al. Automated Delineation of Lung Tumors from CT Images Using a Single Click Ensemble Segmentation Approach. Pattern Recognit 46, 692–702 (2013).
30. Aerts, H. J. et al. Identification of residual metabolic-active areas within individual NSCLC tumours using a pre-radiotherapy (18)Fluorodeoxyglucose-PET-CT scan. Radiother Oncol 91, 386–392 (2009).

CHAPTER 9

Robust radiomics feature quantification using semiautomatic volumetric segmentation

Published in: PLoS ONE 9(7); 2014

Robust radiomics feature quantification using semiautomatic volumetric segmentation

Chintan Parmar, Emmanuel Rios Velazquez*, Ralph Leijenaar, Mohammed Jermoumi, Sara Carvalho, Raymond H. Mak, Sushmita Mitra, B. Uma Shankar, Ron Kikinis, Benjamin Haibe-Kains, Philippe Lambin, Hugo J.W.L. Aerts*

*These authors contributed equally to this work

ABSTRACT

Purpose

Due to advances in the acquisition and analysis of medical imaging, it is currently possible to quantify the tumor phenotype. The emerging field of Radiomics addresses this issue by converting medical images into minable data by extracting a large number of quantitative imaging features. One of the main challenges of Radiomics is tumor segmentation. Where manual delineation is time consuming and prone to inter-observer variability, it has been shown that semi-automated approaches are fast and reduce inter-observer variability. In this study, a semiautomatic region growing volumetric segmentation algorithm, implemented in the free and publicly available 3D-Slicer platform, was investigated in terms of its robustness for quantitative imaging feature extraction.

Materials and methods

Fifty-six 3D-Radiomics features, quantifying phenotypic differences based on the tumor intensity, shape and texture, were extracted from the computed tomography images of twenty lung cancer patients. These Radiomics features were derived from the 3D-tumor volumes defined by three independent observers twice using 3D-Slicer, and compared to manual slice-by-slice delineations of five independent physicians in terms of intra-class correlation coefficient (ICC) and feature range.

Results

Radiomics features extracted from 3D-Slicer segmentations had significantly higher reproducibility (ICC= 0.85 ± 0.15, p= 0.0009) compared to the features extracted from the manual segmentations (ICC= 0.77 ± 0.17). Furthermore, we found that features extracted from 3D-Slicer segmentations were more robust, as the range was significantly smaller across observers (p= 3.819^{e-07}), and overlapping with the feature ranges extracted from manual contouring (boundary lower: p= 0.007, higher: p= 5.863^{e-06}).

Conclusions

Our results show that 3D-Slicer segmented tumor volumes provide a better alternative to the manual delineation process, as they are more robust for quantitative image feature extraction. Therefore, 3D-Slicer can be employed for quantitative image feature extraction and image data mining research in large patient cohorts.

INTRODUCTION

Lung cancer affects approximately 1.6 million people worldwide every year [1]. The majority of lung cancer cases are non-small cell lung cancer (NSCLC), which has substantially poor prognosis and low survival rates [2].

Medical imaging is one of the major disciplines involved in oncologic science and treatment. By assessing human tissues non-invasively, imaging is extensively used for the detection, diagnosis, staging, and management of lung cancer. Due to the emergence of personalized medicine and targeted treatment, the requirement of quantitative image analysis has risen along with the increasing availability of medical data. Radiomics addresses this issue, and refers to the high throughput extraction of a large number of quantitative and mineable imaging features, assuming that these features convey prognostic and predictive information [3,4]. It focuses on optimizing quantitative imaging feature extraction through computational approaches and developing decision support systems, to accurately estimate patient risk and improve individualized treatment selection and monitoring.

Quantitative imaging features, extracted from medical images, are being extensively examined in clinical research. Several studies have shown the importance of imaging features for treatment monitoring and outcome prediction in lung and other cancer types [5-7]. For example, Ganeshan et al. assessed tumor heterogeneity in terms of imaging features extracted from routine computed tomography (CT) imaging in NSCLC, and reported their association with tumor stage, metabolism [8], hypoxia, angiogenesis [9] and patient survival [10]. Furthermore, several studies have uncovered the underlying correlation between gene expression profiles and radiographic imaging phenotype [11,12]. This kind of radiogenomic analysis has raised the utility of medical image descriptors in clinical oncology by projecting them as potential predictive biomarkers [13,14].

To ensure the reliability of quantitative imaging features, accurate and robust tumor delineation is essential. Tumor segmentation is one of the main challenges of Radiomics, as manual delineation is prone to high inter-observer variability and represents a time-consuming task [3,4]. This makes the requirement of (semi)automatic and efficient segmentation methods evident. It has been shown that semiautomatic tumor delineation methods are better alternatives to manual delineations [15,16]. Recently, we have shown that for NSCLC, semiautomatic segmentation using 3D-Slicer (a free open source software platform for biomedical imaging research) reduces inter-observer variability and delineation uncertainty, compared to manual segmentation [17]. During the evaluation of quantitative imaging features as prognostic or predictive factors, it is essential to determine their variability with respect to the tumor delineation process. We hypothesize that quantitative imaging features extracted from semi-automatically segmented tumors have lower variability and are more robust compared to features extracted from manual tumor delineations.

In this study we analyzed the robustness of imaging features derived from semi-automatically and manually segmented primary NSCLC tumors in twenty patients. We extracted fifty-six CT 3D-Radiomics features from 3D-Slicer segmentations made by three independent observers, twice, and compared them to the features extracted from manual delineations provided by five independent physicians. As 3D-Slicer is publicly available and easily accessible by download, it can have a large application in Radiomics to extract robust quantitative image features, and be employed for high-throughput data mining research of medical imaging in clinical oncology.

MATERIAL AND METHODS

CT-PET scans of NSCLC patients

The imaging data was acquired at MAASTRO Clinic in The Netherlands, as reported previously by Baardwijk et al [25]. In short, twenty patients with histologically verified non-small cell lung cancer, stage IB-IIIB, were included in this study. All patients received a diagnostic whole body positron emission tomography (PET)-computed tomography (CT) scan (Biograph, SOMATOM Sensation 16 with an ECAT ACCEL PET scanner; Siemens, Erlangen, Germany). Patients were instructed to fast at least six hours before administration of ^{18}F-fluoro-2-deoxy-glucose (FDG) (MDS Nordion, Liège, Belgium), followed by physiologic saline (10 mL). After the injection of FDG, the patients were encouraged to rest for a period of 45 minutes. Next, free-breathing PET and CT images were acquired. The whole thorax spiral CT scan was acquired with intravenous contrast. The PET images were obtained in 5-min bed positions. The complete data set was then reconstructed iteratively with a reconstruction increment of 5 mm. This study was approved by the local Medical Ethics Committee (Maastricht University Medical Center) and according to the Dutch law. As it was a retrospective study the requirement for informed consent was waived.

Semiautomatic segmentation in 3D Slicer

For the semiautomatic segmentation, the GrowCut algorithm implemented in 3D-Slicer was used (www.slicer.org). GrowCut is an interactive region growing segmentation strategy. Given an initial set of label points the algorithm automatically segments the remaining image by using cellular automation. The algorithm uses a competitive region growing approach and is considered to provide good accuracy and speed for both the 2D and 3D image segmentation. For N-class segmentation the algorithm needs N initial sets of labeled pixels (one set corresponding to each class) from the user. Based on these, the algorithm automatically generates the region of interest (ROI), which is the convex hull of the user-labeled pixels with an additional margin. Next, it iteratively labels all the remaining pixels in the ROI using the user-given pixel labels. Pixel labeling is done using a weighted similari-

ty score, which is a function of the neighboring pixel weights. An unlabeled pixel is labeled corresponding to the neighboring pixels that have the highest weights. The algorithm converges when all the pixels in the ROI have unchanged labels across several iterations.

3D-Slicer provides a graphical user interface (GUI) as the frontend and an efficient algorithm as the backend for the GrowCut segmentation. After loading the patient data, the process begins with the initialization of the foreground and background by marking the area inside and outside the tumor region. Next, the Growcut automatic competing region-growing algorithm gets activated, and segments the ROI into foreground and background regions. Thereafter, background and the surrounding isolated foreground pixels are removed following visual inspection.

Manual tumor delineations

Five physicians manually delineated the gross tumor volume (GTV) of the primary tumor based on fused PET-CT images using standard delineation protocol [which includes fixed window-level settings of both CT (lung W 1,700; L –300, mediastinum W 600; L 40) and PET scan (W 30,000; L 15,000) 2,7,22]. Radiation oncologists were mutually blind of each other's delineations. The primary GTV was defined for each patient based on combined CT and PET information along the axial plane. The physicians were given transversal, coronal, sagittal and 3D views simultaneously. A treatment planning system (XiO; Computer Medical System, Inc., St. Louis, MO) was used for performing delineations.

Image processing and feature extraction

All image data were loaded and analyzed in Matlab R2012b (The Mathworks, Natick, MA) using an adapted version of CERR (Computational Environment for Radiotherapy Research)[26], extended with in-house developed Radiomics image analysis software to extract imaging features.

From the five manual and the six 3D-Slicer segmentations, we extracted fifty-six 3D-Radiomics features for the computed tomography scans. See figure 1 for an illustration of the employed methodology. A mathematical description of all features is shown in Supplement I. The radiomics features were divided in three groups: (I) tumor intensity, (II) shape, and (III) texture. The tumor intensity features consisted of features describing histogram of voxel intensity values contained within the volume of interest (VOI). Geometric features were calculated, describing the three-dimensional shape and size of the lesions. Textural features describing patterns or spatial distribution of voxel intensities, were calculated from gray level co-occurrence (GLCM) [27] and gray level run-length (GLRLM) matrices respectively [28]. Determining texture matrix representations requires the voxel intensity values within the VOI to be discretized. This step not only reduces image noise, but also normalizes intensities across all patients, allowing for a direct comparison of all calculated textural features between patients. Texture matrices were determined considering 26-connected voxels (i.e. voxels were considered to be neighbors in all 13 directions

in three dimensions), and a distance of one voxel between consecutive voxels was set for computing co-occurrence and gray level run-length matrices. Features derived from co-occurrence and gray level run-length matrices were calculated by averaging their value over all 13 considered directions in three dimensions. Overall, the extracted imaging features comprised 15 features describing tumor intensity, 8 shape features and 33 textural features.

Statistical analysis

Intra-class correlation coefficient (ICC) was calculated in order to quantify the feature reproducibility. The ICC is a statistical measure, ranging between 0 and 1, indicating null and perfect reproducibility, respectively. In order to determine the ICC for inter-observer segmentations, variance estimates were obtained from two-way mixed effect model of analysis of variance (ANOVA). McGraw and Wong [29] defined ICC in case 3A to measure the absolute agreement as,

$$ICC = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n}(MS_C - MS_E)}$$

ICC values for intra-observer segmentations were obtained from one-way analysis of variance (ANOVA). It is defined using case 1 of McGraw and Wong [29] as,

$$ICC = \frac{MS_R - MS_W}{MS_R + (k-1)MS_W}$$

Where MS_R = mean square for rows, MS_W = mean square for residual sources of variance, MS_E = mean square error, MS_C = mean square for columns, k = number of observers involved and n = number of subjects. R package IRR (inter rater reliability) was used for ICC computation [30].

Wilcoxon rank-sum test was used to compare the reproducibility of image features derived from manual and 3D-Slicer segmentations methods. Two methods were considered to be significantly different when the p-value was lower than 0.05. All data are expressed as mean ± SD. All the analyses were performed in Matlab (The MathWorks Inc., Natick, MA, USA) and R (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

In order to assess the robustness of 3D-Slicer segmentation on CT imaging for quantitative image feature extraction, we assessed fifty-six 3D-radiomics features quantifying I) tumor intensity, II) tumor shape, and III) tumor texture (**Fig. 1**, Supplement I online). From twenty lung cancer patients we extracted the radiomics features from 3D-volumes defined by

three independent observers twice using 3D-Slicer, and compared them to manual delineations by five independent radiation oncologists.

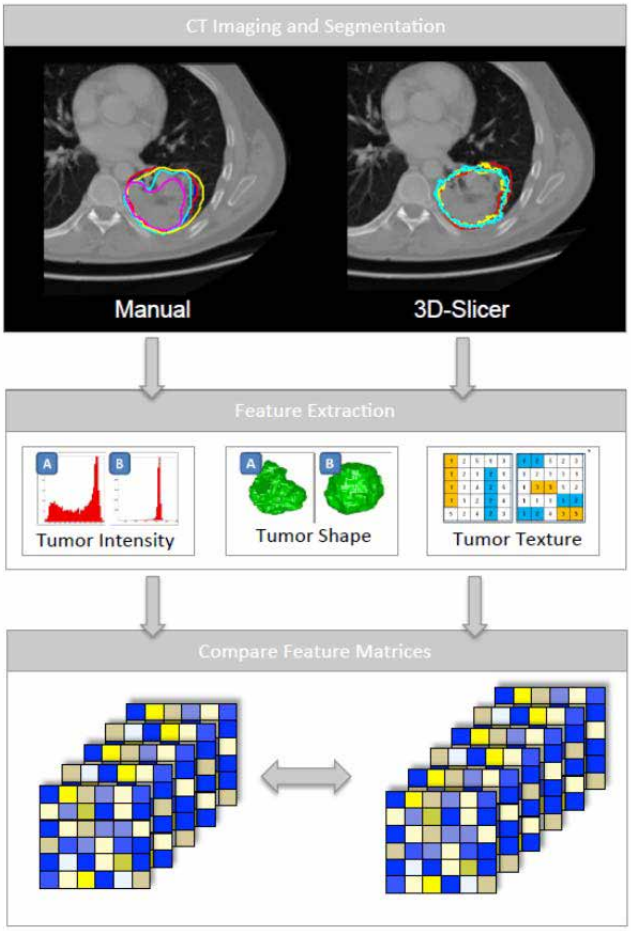


Figure 1

Schematic diagram depicting the overview of the analysis. A: First, we performed five manual delineations and six 3D-Slicer segmentations (three observers twice) on twenty lung tumors. B: Second, fifty-six radiomics features quantifying tumor intensity, texture and shape were extracted from these segmentations. C: Third, the resulting feature matrices were compared for robustness of the feature values.

Since two 3D-Slicer segmentations from each of the three observers were considered for the analysis, the six 3D-Slicer segmentations were divided in to two sets, each having three segmentations (one from each observer). We calculated the intra-class correlation coefficient (ICC) for the radiomics features extracted from these two sets of three 3D-Slicer segmentations and five manual delineations. We observed that the radiomics features extracted from 3D-Slicer segmentations, had significantly higher reproducibility (avg.

of two 3D-Slicer segmentation sets ICC = 0.85 ± 0.15) as compared to the features extracted from the manual segmentations (ICC = 0.77 ± 0.17) (p = 0.0009, **Fig. 2**).

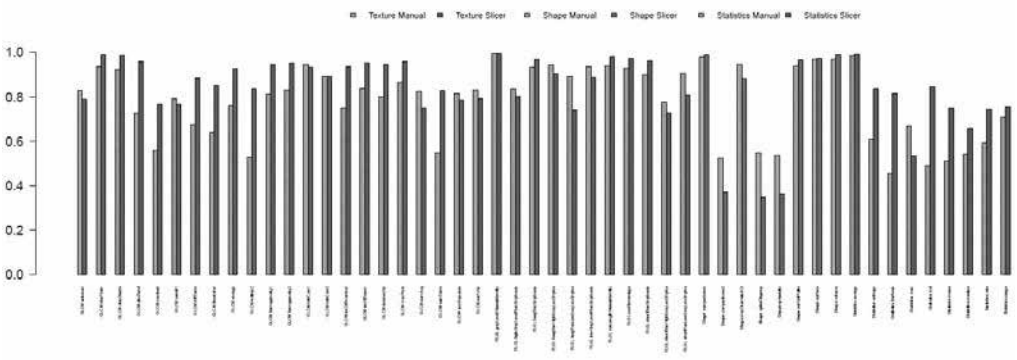


Figure 2

Feature wise comparison of Intra-class correlation coefficients (ICC) between manual and 3D-Slicer segmentations. A: First order statistics features. B: Shape based features. C Textural features.

Overall 38 out of the 56 features (68%) showed higher ICC values for 3D-Slicer segmentations as compared to the manual ones. ICC values for all the assessed features are reported in Supplement II (online). To evaluate the robustness against multiple algorithmic initializations of the same observer, we computed ICC for the three intra-observer 3D-Slicer segmentation sets, each having two 3D-Slicer segmentations from the same observer. High ICC values (avg. of three intra-observer 3D-Slicer segmentation sets ICC = 0.90 ± 0.17) were observed for intra-observer segmentation groups. **Fig. 3** depicts the ICC values corresponding to the inter-observer manual delineations and intra- & inter-observer 3D-Slicer segmentations.

Intensity statistics and textural features showed significantly higher reproducibility (two sided Wilcoxon test p = 0.0006, p = 0.009, respectively) for 3D-Slicer based segmentations (avg. inter-observer ICC = 0.82 ± 0.13, ICC= 0.88 ± 0.09, respectively) as compared to manual delineations (ICC = 0.63 ± 0.16, ICC= 0.82 ± 0.12, respectively). No statistically significant difference (two sided Wilcoxon test p = 0.31) was observed in ICC values for shape based features between the manual (ICC = 0.80 ± 0.22) and semiautomatic (avg. inter-observer ICC= 0.75 ± 0.31) groups. Fourteen out of 15 statistical features (93%), and 20 out of 33 textural features (67%), showed higher reproducibility (higher ICC) for 3D-Slicer segmentations as compared to manual delineations. For shape based descriptors there was no clear winner between the two segmentation strategies as 4 out of 8 (50%) features turned out having higher ICC for 3D-Slicer segmentations.

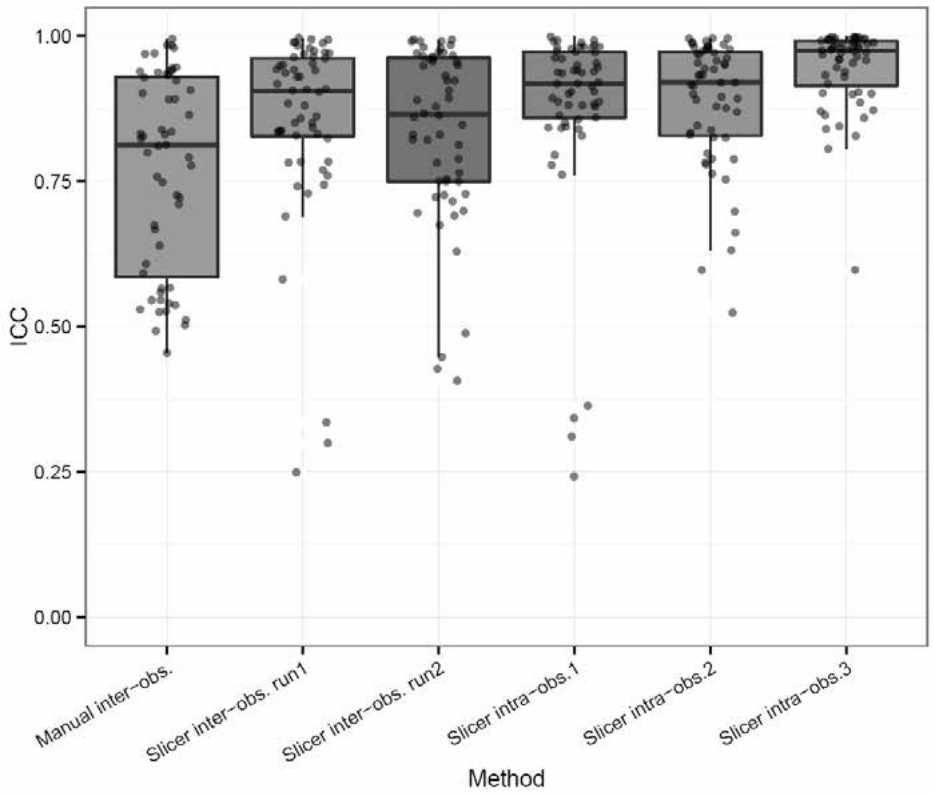


Figure 3

Box-plot comparing intra- and inter-observer reproducibility (ICC) of radiomics features. High inter- and intra- observer reproducibility (ICC) was observed for 3D-Slicer segmentations compared to the inter-observer reproducibility (ICC) of manual delineations. From left the first box refers to the manual inter-observer reproducibility (ICC), second and third boxes refer to the inter-observer reproducibility (ICC) of two different 3D-Slicer segmentation runs. Remaining three boxes refer to the intra-observer reproducibility (ICC) of 3D-Slicer segmentations.

We next classified the 56 features into three groups according to their ICC, as (I) having a high (ICC≥0.8), (II) medium (0.8>ICC≥0.5), or (III) low (ICC<0.5) reproducibility (Supplement II online). For manual delineations, 52% of all the assessed features had high, 45% had medium, and 3% had low reproducibility on the other hand for 3D-Slicer based semiautomatic segmentations, 70% features had high, 25% had medium, and 5% had low reproducibility. Therefore, reproducibility of the features was, in general, higher for 3D-Slicer segmentations.

Furthermore, it becomes important to determine whether the features extracted from semiautomatic segmentations capture the same tumor image properties as with manual delineations. Therefore, we compared the normalized range for all features between these two segmentation groups (**Fig. 4**). We normalized every feature value with respect to all 11 (5 manual + 6 3D-Slicer) segmentations, using Z-score normalization. We

observed that the features extracted from 3D-Slicer based segmentations, spread over significantly smaller range across observers as compared to those of the manual delineations (two sided Wilcoxon test p = 3.819e-07). Moreover, the features derived from 3D-Slicer segmentations overlapped in range with those of the manual delineations, as the lower(higher) limit(s) being significantly higher(lower) for the 3D-Slicer features (two sided Wilcoxon test p = 0.007, p = 5.863e-06). This corroborates that the feature set, extracted from both the semiautomatic and manual strategies, correspond to similar tumor image characteristics, with the features from 3D-Slicer having less variability across observers.

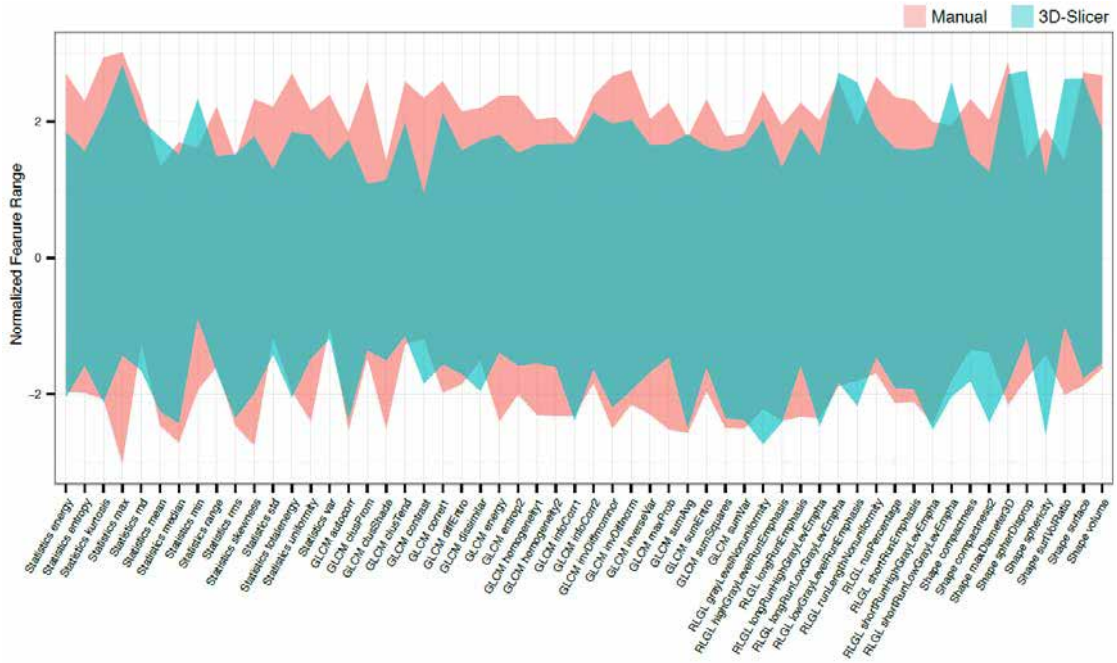


Figure 4

Comparison of normalized feature range between manual and 3D-Slicer segmentation groups. Radiomics features derived from 3D-Slicer segmentations had significantly smaller and overlapping range compared to that from manual delineations.

DISCUSSION

Medical imaging is considered as one of the fundamental building blocks of clinical oncology. It is routinely used for cancer staging, treatment planning, and treatment response monitoring. Furthermore, recent developments in computational imaging, data mining and predictive analysis have broadened the scope of the imaging in clinical oncology. For example, quantitative imaging features extracted from CT images have been shown to predict 78% of the gene expression variability in hepatocellular carcinoma [11]. In a similar

study, image descriptors, extracted from contrast enhanced MRI images of glioblastoma patients, predicted immunohistochemical identified protein expression patterns [12,18]. Recent computational approaches for image quantification, such as Radiomics, hypothesize that image descriptors extracted from tumor regions are associated with the risk of adverse events after treatment and could provide improved prognostic information for patient management [3,4].

Accurate and efficient tumor segmentation is one the main challenges for the extraction of robust quantitative imaging features [4]. Manual segmentation suffers from high inter-observer variability and is time consuming [19]. It has been reported that semi-automatic segmentation strategies, as compared to manual delineation can improve tumor segmentation by reducing uncertainty as well as time [15,17,19]. These studies focused on tumor volumes while comparing semiautomatic and manual segmentation methods. However, tumor segmentation should also be evaluated in terms of the reliability of radiomics features derived from the volume of interest (VOI).

In this study, we investigated the robustness of quantitative imaging features, extracted from 3D-Slicer tumor segmentations, as compared to those, extracted from manual tumor delineations. Overall 3D-Slicer based semiautomatic segmentation method produced more reproducible radiomics features ($p = 0.0009$). We also analyzed different feature groups for their reproducibility, and observed that the difference in ICC, for intensity statistics and textural features, was statistically significant ($p= 0.0006$, $p= 0.0094$, respectively) between the two segmentation strategies. The shape features, however did not significantly differ in reproducibility between the two strategies ($p = 0.31$). We also analyzed intra– and inter–observer reproducibility for 3D Slicer based semiautomatic segmentations. Three independent observers segmented each tumor twice, with different algorithmic initialization. Image descriptors demonstrated high intra-observer reproducibility for 3D-Slicer segmentations, which indicates their robustness over different seed point initializations. We also observed high inter-observer reproducibility in image descriptors for semiautomatic segmentations. Further reduction of inter-observer variability could be achieved by improving the semiautomatic segmentation strategy, i.e., by reducing observer interaction. Fully automatic methods requiring minimum user interaction, that may solve the complex problem of accurately defining the tumor boundaries, particularly in the case of large tumors with pleural attachment, are still a matter of investigation [20]. Although, current investigation shows that 3D-Slicer segmentation provides a more robust alternative to manual contouring. Furthermore, as 3D-Slicer is publicly available and easily accessible by download, we expect its large utility in the field of quantitative imaging.

Recently the reproducibility of quantitative image features has been evaluated against repetitive test-retest CT image scans, acquired within fifteen minutes time interval, and was used to select the most informative radiomics features [4]. This work was expanded by Hunter et al, to evaluate the robustness of CT image features over three different imaging machines for identifying high quality multi-machine robust radiomics fea-

tures [21]. In both these studies, since the NSCLC tumors were segmented by a single observer (by using a semiautomatic segmentation), the inter-observer reproducibility of the imaging features could not be evaluated. Leijenaar et al, have analyzed the stability to FDG-PET image features with respect to test-retest scans and inter-observer delineations independently and reported a strong correlation between them [22]. Although they quantified the radiomics PET-based features for manual delineation stability, they did not compare it with that of semiautomatic tumor segmentations. No previous study, in our knowledge, has evaluated the reproducibility of quantitative CT-based imaging features in NSCLC, with respect to tumor segmentation methods.

One of the limitations of our study is not being able to associate these image descriptors with patient outcome due to cohort size and unavailability of clinical data. It would be interesting to investigate the effects of manual and semiautomatic segmentations on the image descriptor based prognostic performance. We hypothesize that more robust features having a stronger association with patient outcome are the most important biomarkers and play a vital role in high throughput data-mining research like Radiomics. Besides segmentation methods, other sources of variation should also be considered while evaluating quantitative image features. For instance, Galavis et al. investigated the variability in quantitative image descriptors due to different image acquisition modes and reconstruction parameters [23]. It has also been shown that different ways of image discretization influence the variability of textural features [24]. Although image acquisition, reconstruction and delineation protocols are typically standardized in the clinical practice, there still exists significant variation between imaging studies. Standardized protocols using semiautomatic segmentation tools are also warranted. Therefore, imaging features should be selected based on their robustness towards these sources of variation as well as their prognostic performance.

In conclusion, 3D-Slicer based semiautomatic segmentation significantly improves the robustness of radiomics feature quantification and thus could serve as a potential alternative to the time consuming manual segmentation process. 3D-Slicer can have a large application in radiomics research to extract robust quantitative image features, and be employed for high-throughput data mining research of medical imaging in clinical oncology.

REFERENCES

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, et al. (2011) Global cancer statistics. CA: A Cancer Journal for Clinicians 61: 69-90.
2. van Baardwijk A, Wanders S, Boersma L, Borger J, Öllers M, et al. (2010) Mature results of an individualized radiation dose prescription study based on normal tissue constraints in stages I to III non–small-cell lung cancer. Journal of Clinical Oncology 28: 1380-1386.

3. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, et al. (2012) Radiomics: Extracting more information from medical images using advanced feature analysis. European Journal of Cancer 48: 441-446.
4. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, et al. (2012) Radiomics: the process and the challenges. Magnetic Resonance Imaging 30: 1234-1248.
5. Vaidya M, Creach KM, Frye J, Dehdashti F, Bradley JD, et al. (2012) Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. Radiotherapy and Oncology 102: 239-245.
6. El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, et al. (2009) Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recognition 42: 1162-1171.
7. Tixier F, Le Rest CC, Hatt M, Albarghach N, Pradier O, et al. (2011) Intra-tumor heterogeneity on baseline 18 F-FDG PET images characterized by textural features predicts response to concomitant radio-chemotherapy in esophageal cancer. Journal of Nuclear Medicine (JNM) 52: 369-378.
8. Ganeshan B, Abaleke S, Young RC, Chatwin CR, Miles KA (2010) Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. Cancer Imaging 10: 137-143.
9. Ganeshan B, Goh V, Mandeville HC, Ng QS, Hoskin PJ, et al. (2013) Non–small cell lung cancer: histopathologic correlates for texture parameters at CT. Radiology 266: 326-336.
10. Ganeshan B, Panayiotou E, Burnand K, Dizdarevic S, Miles K (2012) Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. European Radiology 22: 796-802.
11. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, et al. (2007) Decoding global gene expression programs in liver cancer by noninvasive imaging. Nature Biotechnology 25: 675-680.
12. Zinn PO, Majadan B, Sathyan P, Singh SK, Majumder S, et al. (2011) Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme. PLoS One 6: e25451.
13. Buckler AJ, Bresolin L, Dunnick NR, Sullivan DC (2011) Quantitative imaging test approval and biomarker qualification: interrelated but distinct activities. Radiology 259: 875-884.
14. Buckler AJ, Bresolin L, Dunnick NR, Sullivan DC (2011) A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. Radiology 258: 906-914.
15. Rios Velazquez E, Aerts HJ, Gu Y, Goldgof DB, De Ruysscher D, et al. (2012) A semiautomatic CT-based ensemble segmentation of lung tumors: Comparison with oncologists' delineations and with the surgical specimen. Radiotherapy and Oncology 105: 167-173.
16. Heye T, Merkle EM, Reiner CS, Davenport MS, Horvath JJ, et al. (2013) Reproducibility of Dynamic Contrast-enhanced MR Imaging. Part II. Comparison of Intra-and Interobserver Variability with Manual Region of Interest Placement versus Semiautomatic Lesion Segmentation and Histogram Analysis. Radiology 266: 812-821.
17. Rios Velazquez E, Parmar C, Jermoumi M, Mak RH, van Baardwijk A, et al. (2013) Volumetric CT-based segmentation of NSCLC using 3D-Slicer. Scientific Reports 3: DOI: 10.1038/srep03529.
18. Zinn PO, Sathyan P, Mahajan B, Bruyere J, Hegi M, et al. (2012) A novel volume-age-KPS (VAK) glioblastoma classification identifies a prognostic cognate microRNA-gene signature. PLoS One 7: e41522.
19. Egger J, Kapur T, Fedorov A, Pieper S, Miller JV, et al. (2013) GBM Volumetry using the 3D Slicer Medical Image Computing Platform. Scientific Reports 3.
20. Gu Y, Kumar V, Hall LO, Goldgof DB, Li C-Y, et al. (2012) Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach. Pattern Recognition 46: 692-702.
21. Hunter LA, Krafft S, Stingo F, Choi H, Martel MK, et al. (2013) High quality machine-robust image features: Identification in nonsmall cell lung cancer computed tomography images. Medical Physics 40: DOI:10.1118/1111.4829514.
22. Leijenaar RT, Carvalho S, Velazquez ER, Van Elmpt WJ, Parmar C, et al. (2013) Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability. Acta Oncologica 52: 1391-1397.
23. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R (2010) Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncologica 49: 1012-1016.

24. Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, et al. (2012) Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. Journal of Nuclear Medicine 53: 693-700.
25. Van Baardwijk A, Bosmans G, Boersma L, Buijsen J, Wanders S, et al. (2007) 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. International Journal of Radiation Oncology* Biology* Physics 68: 771-778.
26. Deasy JO, Blanco AI, Clark VH (2003) CERR: a computational environment for radiotherapy research. Medical Physics 30: 979-985.
27. Haralick RM, Shanmugam K, Dinstein IH (1973) Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics SMC-3: 610-621.
28. Galloway MM (1975) Texture analysis using gray level run lengths. Computer Graphics and Image Processing 4: 172-179.
29. McGraw KO, Wong S (1996) Forming inferences about some intraclass correlation coefficients. Psychological methods 1: 30-46.
30. Gamer M, Lemon J, Fellows I, Singh P (2013) IRR: Various coefficients of interrater reliability and agreement. R package version 0.84. CRAN: http://www.r-project.org.13. Chen AY, Halpern M (2007) Factors predictive of survival in advanced laryngeal cancer. Arch Otolaryngol Head Neck Surg 133:1270–1276

CHAPTER 10

Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach

Published in: Nature Communications. 5:4006; 2014.

Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach

Hugo J.W.L. Aerts, Emmanuel Rios Velazquez*, Ralph T.H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Cavalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. René Leemans, Joseph O. Deasy, Andre Dekker, John Quackenbush, Robert J. Gillies, Philippe Lambin*

*These authors contributed equally to this work

ABSTRACT

Human cancers exhibit strong phenotypic differences that can be visualized noninvasively by medical imaging. Radiomics refers to the comprehensive quantification of tumour phenotypes by applying a large number of quantitative image features. Here we present a radiomic analysis of 440 features quantifying tumour image intensity, shape and texture, which are extracted from computed tomography data of 1,019 patients with lung or head-and-neck cancer. We find that a large number of radiomic features have prognostic power in independent data sets of lung and head-and-neck cancer patients, many of which were not identified as significant before. Radiogenomics analysis reveals that a prognostic radiomic signature, capturing intratumour heterogeneity, is associated with underlying gene-expression patterns. These data suggest that radiomics identifies a general prognostic phenotype existing in both lung and head-and-neck cancer. This may have a clinical impact as imaging is routinely used in clinical practice, providing an unprecedented opportunity to improve decision-support in cancer treatment at low cost.

INTRODUCTION

Medical imaging is one of the major factors that have informed medical science and treatment. By assessing the characteristics of human tissue non-invasively, imaging is often used in clinical practice for oncologic diagnosis and treatment guidance[1-3]. A key goal of imaging is 'personalized medicine', were treatment is increasingly tailored based on specific characteristics of the patient and their disease[4].

Much of the discussion of personalized medicine has focused on molecular characterization using genomic and proteomic technologies. However, as tumors are spatially and temporally heterogeneous, these techniques are limited. They require biopsies or invasive surgeries to extract and analyze what are generally small portions of tumor tissue, which do not allow for a complete characterization of the tumor. Imaging has great potential to guide therapy because it can provide a more comprehensive view of the entire tumor and it can be used on an on-going basis to monitor the development and progression of the disease or its response to therapy. Further, imaging is non-invasive and is already often repeated during treatment in routine practice, on the contrary of genomics or proteomics, which are still challenging to implement into clinical routine.

The most widely used imaging modality in oncology is x-ray computed tomography (CT), which assesses tissue density. Indeed, CT images of lung cancer tumors exhibit strong contrast reflecting differences in the intensity of a tumor on the image, intra tumor texture, and tumor shape (**Fig.1a**). However, in clinical practice, tumor response to therapy is only measured using 1 or 2 dimensional descriptors of tumor size (RECIST and WHO, respectively)[5]. While a change in tumor size can indicate response to therapy, it often does not predict overall or progression free survival[6,7]. Although some investigations have characterized the appearance of a tumor on CT images, these characteristics are typically described subjectively and qualitative ("moderate heterogeneity", "highly spiculated", "large necrotic core"). However, recent advances in image acquisition, standardization, and image analysis, allow for objective and precise quantitative imaging descriptors that could potentially be used as non-invasive prognostic or predictive biomarkers.

Radiomics is an emerging field that converts imaging data into a high dimensional mineable feature space using a large number of automatically extracted data-characterization algorithms[8, 9]. We hypothesize that these imaging features capture distinct phenotypic differences of tumours and may have prognostic power and thus clinical significance across different diseases. Here we assess the clinical relevance of 440 radiomic features, many of which currently have no known clinical significance, in seven independent cohorts consisting of 1,019 lung cancer and head-and-neck cancer patients. Two data sets are used to assess the stability of the features, four data sets to assess the prognostic value of radiomic features on lung cancer patients and head-and-neck cancer patients, and one data set for association with gene-expression profiles of lung cancer patients (Fig. 2). Our results reveal that radiomics data contain strong prognostic information in both lung and head-and-neck cancer patients, and are associated with the underlying

gene-expression patterns. These results suggest that radiomics decodes a general prognostic phenotype existing in multiple cancer types. Radiomics can have a large clinical impact, as imaging is used in routine practice worldwide, providing a method that can quantify and monitor phenotypic changes during treatment.

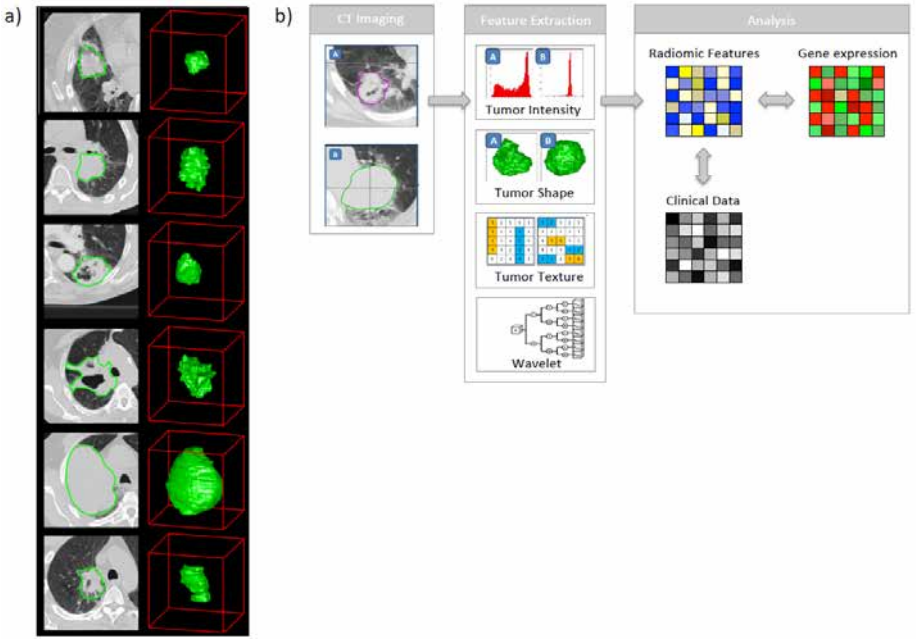


Figure 1

*Extracting Radiomics data from images. (**a**) Tumors are different. Example computed tomography (CT) images of lung cancer patients. (**b**) Strategy for extracting Radiomics data from images.*

RESULTS

First, we defined 440 quantitative image features describing tumor phenotype characteristics by: I) tumor image intensity, II) shape, III) texture and IV) multi-scale Wavelet (**Fig.1b**, Supplement I online).

To investigate radiomic expression patterns we extracted radiomic features from the Lung1 dataset, consisting of 422 NSCLC cancer patients (**Fig.2**). Unsupervised clustering revealed clusters of patients with similar radiomic expression patterns (**Fig.3**). We compared the three main clusters of patients with clinical parameters (**Fig.3b**), and found significant association with primary tumor stage (T-stage; p <1x10^{-24}) and overall stage (p<1x10^{-6}), wherein cluster I was associated with lower stages. N-stage (lymph node) and M-stage (metastasis), however, showed no correspondence with the radiomic expression patterns (p=0.27, and p=0.73 respectively).

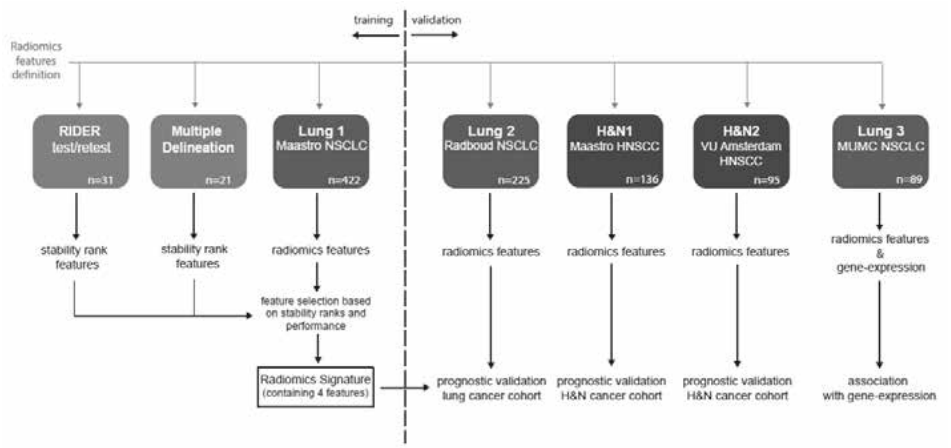


Figure 2

*Analysis workflow. The defined radiomic features algorithms (more information **Supplement I**) were applied to seven different datasets (more information **Supplement II**). Two datasets were used to calculate the feature stability ranks, RIDER test/retest and Multiple Delineation respectively (both orange). The Lung1 dataset was used as training dataset. Lung2, H&N1, and H&N2 were used as validation datasets. The Lung3 dataset was used for association of the radiomic signature with gene expression profiles.*

Furthermore, a significant association with histology (p=0.31x10-3) was observed, wherein squamous cell carcinoma showed a higher presence in cluster II. Looking at the representation of the feature groups (**Fig.3c**), there was no correspondence between the feature group and radiomic expression patterns.

The analysis was divided in training and validation phases (**Fig.2**). For the training phase, we first explored feature stability determined in both test-retest and inter-observer setting. Using the publicly available RIDER[10] dataset, consisting of 31 sets of test-retest CT-scans that were acquired approximately 15 minutes apart, we tested how consistent the radiomic features were between the test and retest scan. The multiple delineation dataset, where five oncologists delineated lesions on CT scans from 21 patients[11], was used to test the stability of the radiomic features to variation in manual delineations.

For each feature we compared the stability ranks for test-retest and multiple delineation with prognosis in the Lung1 training dataset. Although the stability ranks did not use any information about prognosis, in general, features with higher stability for test retest and delineation inaccuracies showed higher prognostic performance (**Extended Data Figure 1**). This is possibly due to reduced amount of noise in the stable features and supports the use of stability ranks for feature selection.

The possible association of radiomic features with survival was then explored by Kaplan-Meier survival analysis. For training we used the Lung1 dataset, and for validation the Lung2, H&N1, H&N2 datasets (**Fig.2**). The radiomic features were not normalized on any dataset, and only the raw values were used that were directly computed from the DICOM images.

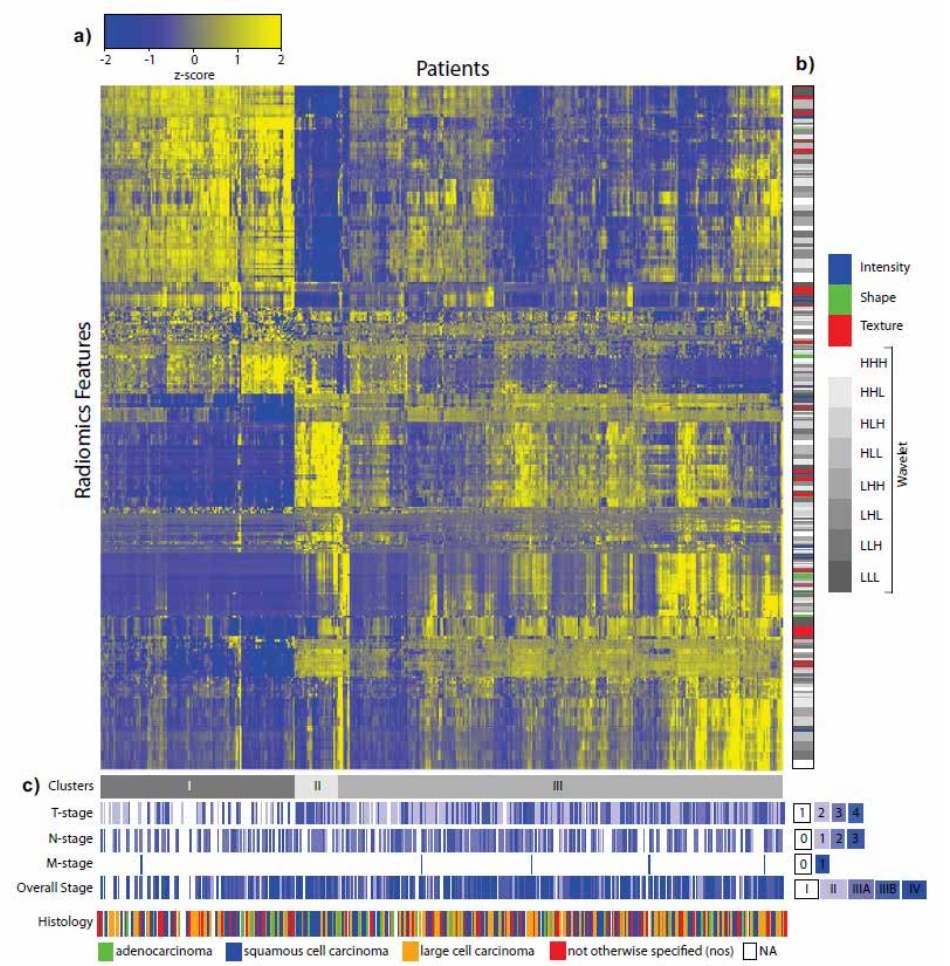


Figure 3

*Radiomics heat map. (**a**) Unsupervised clustering of lung cancer patients (Lung1 set, n=422) on the y-axis and radiomic feature expression (n=440) on the x-axis, revealed clusters of patients with similar radiomic expression patterns. (**b**) Clinical patient parameters for showing significant association of the radiomic expression patterns with primary tumor stage (T-stage; $p<1x10^{-24}$), overall stage ($p<1x10^{-6}$), and histology ($p=0.31x10^{-3}$). (**c**) Correspondence of radiomic feature groups with the clustered expression patterns.*

To ensure a completely independent validation, the median value of each feature was computed on the training Lung1 dataset, and locked for use as a threshold in the validation datasets in order to assess the survival differences without retraining (**Fig.2**). In **Extended Data Figure 2** we show Kaplan-Meier survival curves for four representative features. Features describing heterogeneity in the primary tumor were associated with worse survival in all four datasets. Also, patients with more compact/spherical tumors had better survival probability.

Overall, the median threshold derived from Lung1 yielded a significant survival difference for 238 features (54% of in total 440, FDR 10%) in the Lung2 validation dataset. Furthermore, there was a significant survival difference for 135 features (31%) in H&N1 and for 186 features in H&N2 (42%). Sixty-six (15%) of the features derived from Lung1 were significant for survival in all three-validation datasets (Lung2, H&N1, and H&N2).

To test the multivariate performance of a radiomic signature, we used the workflow depicted in **Extended Data Figure 3**. We focused our analysis on the 100 most stable features, which were determined by averaging the stability ranks of RIDER dataset and Multiple Delineation dataset. To remove redundancy within the radiomic information, we select the single best performing radiomic feature from each of the four feature groups, and combined these top four features into a multivariate Cox proportional hazards regression model for prediction survival.

The resulting radiomic signature consisted of I) "Statistics Energy" (**Supplement I** Feature 1.1) describing the overall density of the tumor volume, II) "Shape Compactness" (Feature 2.2) quantifying how compact the tumor shape is, III) "Gray Level Nonuniformity" (Feature 3.25) a measure for intra-tumor heterogeneity, and IV) Wavelet "Gray Level Nonuniformity HLH" (Feature Group4), also describing intra-tumor heterogeneity after decomposing the image in mid-frequencies. The weights of each of the features in the signature were fitted on the training dataset Lung1. The performance of the four feature radiomic signature was validated in the datasets Lung2, H&N1, and H&N2 (**Fig.4a**) using the concordance index (CI), which is a generalization of the area under the ROC-curve[12].

The radiomic signature had good performance on the Lung2 data (CI=0.65, p=$2.91x10^{-09}$), and a high performance in H&N1 (CI=0.69, p=$7.99x10^{-07}$) and H&N2 (CI=0.69, p=$3.53x10^{-06}$). Although volume had a good performance in all datasets, the radiomic signature performed significantly better, suggesting that radiomic features contain relevant, complementary information for prognosis (Extended Data Table 1). Furthermore, combining the radiomic signature with volume was significantly better than volume alone in all datasets.

Comparing the radiomic signature to the TNM staging[13], we see that the signature performance was better in both Lung2 and H&N2 and comparable in H&N1. Importantly, combining the radiomic signature with TNM staging showed a significant improvement in all datasets, compared with TNM staging alone. Furthermore, we assessed if the radiomics signature preserved the significant prognostic performance compared to the treatment patients received. We found that the signature preserved its prognostic performance for all treatment groups (radiation, or concurrent chemo-radiation), for both Lung and H&N cancer patients (see **Extended Data Table 2)**, demonstrating the complementary value of radiomics for each treatment type.

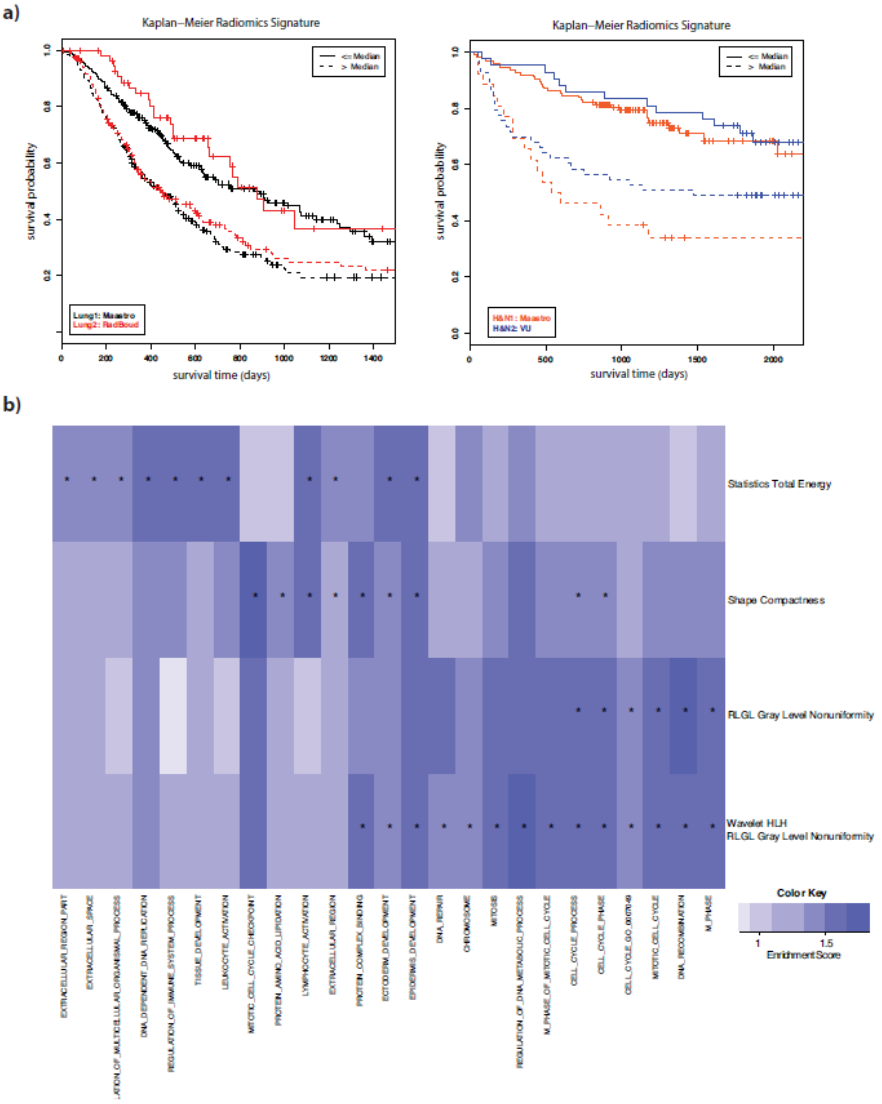


Figure 4

(a) *Radiomic signature performance. Performance of the radiomic signature on the lung cancer datasets (left) and the head and neck cancer datasets (right). The signature was built on the Lung1 data (n=422). The signature had good performance in the Lung2 (CI=0.65, p= $2.91x10^{-09}$, n=225), and a high performance in H&N1 (CI=0.69, p=$7.99x10^{-07}$, n=136) and H&N2 (CI=0.69, p=$3.53x10^{-06}$, n=95) validation datasets.* ***(b)*** *Association of radiomic signature features and gene expression using Gene Set Enrichment Analysis (GSEA). Gene sets that have been significantly enriched for at least one of the four-radiomic features are indicated with an asterisk. The corresponding normalized enrichment scores (NES), GSEA's primary statistic, for all radiomic signature features is displayed in a heat-map, where light blue means low, and dark blue means high NES.*

Human papillomavirus (HPV) is an important determinant in head and neck cancer patients, especially those with oropharyngeal carcinoma for prognosis and may guide future treatment selection. We did not find a significant association between radiomic signature prediction and HPV status in a combined analysis in the H&N1 and H&N2 dataset (*p*=0.17, **Extended Data Table 2**). However, we found that the signature preserved its prognostic performance in the HPV negative group (CI=0.66), consisting of the majority of patients (76%, n=130), demonstrating the complementary value of Radiomics to HPV screening.

To assess the association between the radiomic signature and the underlying biology, we compared the radiomic signature with gene-expression profiles (Lung3 dataset, **Fig.2**) using gene-set enrichments analysis (GSEA)[1,14]. We found significant associations between the signature features and gene-expression patterns (**Fig.4b**). Further, the radiomic features are significantly associated with different biologic gene-sets, demonstrating that radiomic features probe different biologic mechanisms. It is noteworthy that both intratumor heterogeneity features in the signature (Feature III and IV) were strongly correlated with cell cycling pathways, indicating an increased proliferation for more heterogeneous tumors.

DISCUSSION

Medical imaging is one of the major factors informing medical science and treatment. Its potential resides in its ability to assess the characteristics of human tissue non-invasively, and therefore is routinely used in clinical practice for oncologic diagnosis and treatment guidance and monitoring.

However, traditionally, medical imaging has been a subjective or qualitative science. Recent advances in medical imaging acquisition and analysis, allow the high-throughput extraction of informative imaging features to quantify the differences that oncologic tissues exhibit in medical imaging.

Radiomics applies advanced computational methodologies to medical imaging data, to convert medical images into quantitative descriptors of oncologic tissues[8].

In this study, we analyzed 440 radiomic features quantifying tumor phenotypic differences based on its image intensity, shape and texture. In a large dataset of 1019 lung and head and neck cancer patients, of which we extracted radiomic features on computed tomography images, we found that a large number of radiomic features have prognostic power, many of which its prognostic implication have not been described before. Furthermore, our integrated analysis showed that features selected based on their stability and reproducibility were also the most informative features, which indicates the power of integrating independent datasets for radiomic feature selection and model building.

We showed as well that a radiomic signature, capturing intra-tumor heterogeneity, was strongly prognostic and validated in three independent datasets of lung and head and neck cancer patients, and was associated with gene-expression profiles. To avoid any form of over-fitting or bias, we performed a robust statistical validation: only one radiomics signature (containing 4 radiomic features) was validated in data of 545 patients in independent validation datasets (**Figure 2** and **Extended Data Figure 3**). The four features were selected based on feature stability and prognostic performance in the discovery dataset only.

The top performing feature "Gray Level Nonuniformity" (Group3, number 3.25) and the most dominant features in the radiomic signature (feature III and IV), quantified intra-tumor heterogeneity. Indeed, it is often hypothesized that intra-tumor heterogeneity is exhibited on different spatial scales, for example at the radiological, macroscopic, cellular, and the molecular (genetics) level. Radiological tumor phenotype characteristics may thus be useful to investigate the underlying evolving biology. It is known that multiple subclonal populations co-exist within tumors, reflecting extensive intra-tumoral "somatic evolution"[15,16]. This heterogeneity is a clear barrier to the goal of personalized therapy based on molecular biopsy-based assays, as the identified mutations and gene-expression does not always represent the entire population of tumor cells[17,18]. Radiomics circumvents this by assessing the comprehensive 3D tumor bulk. The study presented here probes heterogeneity and demonstrates corresponding clinical importance in two cancer types. Furthermore, we demonstrated association of intra-tumor heterogeneity with proliferation, a general hallmark of cancer.

Overall, the lung-derived radiomic signature had better performance in head and neck compared to lung cancer. One reason could be that head and neck images were acquired with head immobilization, whereas lung images were acquired with free-breathing and are affected by patient movement or respiration, resulting in relatively more image noise. Nonetheless, our results show that the radiomic signature could be transferred from lung to head and neck cancer, which suggests that the signature identifies a general prognostic tumor phenotype.

Our method provides a non-invasive (and therefore with no risk of infection or complications that accompany tissue biopsies), fast, low cost, and repeatable way of investigating phenotypic information, potentially speeding up the development of personalized medicine. Furthermore, we show that the radiomic signature is significantly associated with the underlying gene-expression patterns, suggesting that inter-patient differences of gene-expression are large than intra-patient differences.

The clinical impact of our results are illustrated by the fact that it advances knowledge in the analysis and characterization of tumors in medical images, previously not done, and provides knowledge currently not used in the clinic. We showed the complementary performance of Radiomic features with TNM staging for prediction of outcome, which illustrates the clinical importance of our findings as TNM is routinely used in the clinic. Currently, the TNM staging system is used for risk stratification and treatment

decision-making. However, the TNM staging system is primarily based on resectability of the tumor, while a larger number of NSCLC patients will receive primary treatment with radiotherapy either alone or combined with chemotherapy. Therefore, the TNM staging system is insufficient for risk stratification of this group of patients, in particular to make the decision between curative treatment (concomitant radio-chemotherapy) or palliative treatment especially in elderly patients, a growing issue in western countries. Our results show that the radiomics signature is performing better in independent cohorts than the TNM classification. In future clinical trials this inexpensive method can be used as well for pretreatment risk stratification (e.g. high, low risk).

Furthermore, we have shown for the first time the translational capability of radiomics in two cancer types (lung and head and neck cancer). These results indicate that radiomics quantifies a general prognostic cancer phenotype that likely can broadly be applied to other cancer types. Similar observations have been made in gene-expression studies where signatures are prognostic across different diseases[19].

Analysis of image features applied to medical imaging has been a largely studied field and extensive literature exists. However, the majority of previous work describes the use of imaging features focused in the detection of small nodules in for example mammograms or chest CT/PET scans, or in the differential diagnosis of malignant versus benign nodules (Computed Aided Diagnostics). However, applications and methodologies are distinct from our study. Quantitative imaging for personalized medicine is a recent field, with a limited number of publications [12,20-27]. The main clinical question of this research is not the diagnosis, but how to extract more useful information from the tumor phenotype that can be used for personalized medicine.

Therefore, we assessed the association of radiomics with clinical factors, prognosis, and gene-expression levels, using large amounts of features and with external and independent validation cohorts of patients. The most important message in our manuscript is that there is prognostic and biologic information enclosed in routinely acquired CT imaging and was evident in two cancer types.

It is known that variability in image acquisition exists across hospitals and that this is a reality in clinical practice. However, in our analysis we used data directly generated from the scanner and the features were calculated from the RAW imaging data, without any pre-processing or normalization. As there was no correction by cohort or scanner type, this illustrates the translational potential of our results and it is a strong argument in favor of a multi-centric application of radiomics. The radiomics signature had strong prognostic power in these independent datasets generated in daily clinical practice. Furthermore, we expect that with better standardization and imaging protocols, the power of radiomics will even further improve. Among others, the Quantitative Imaging Network (QIN) of the National Institute of Health (NIH), as well as the quantitative imaging biomarker alliance (QIBA), investigates future directions, by performing phantom studies and discussing with vendor's open and standardized protocols for image acquisition [2,3].

Due to the large availability of non-invasive imaging performed routinely in a large number of cancer patients, and the automated feature algorithms, the results of this work could stimulate further research of image-based quantitative features. Also, we presented evidence that the defined radiomic feature-metrics are platform independent, though this should be studied further, and can potentially be applied to other image modalities, such as magnetic resonance imaging (MRI), or positron emission tomography (PET). This approach can have a large impact as imaging is routinely used in clinical practice, worldwide, in all stages of diagnoses and treatment, providing an unprecedented opportunity to improve medical decision support.

METHODS

Radiomics Features: In Supplement I the algorithms of radiomic features are described in detail. In short, we defined 440 radiomic image features that describe tumor characteristics and can be extracted in an automated way. In total 440 distinct features
were defined, divided in four groups: I) tumor intensity, II) shape, III) texture, and IV) wavelet features. The first group quantified tumor intensity characteristics using first-order statistics, calculated from the histogram of all tumor voxel intensity values. Group 2 consists of features based on the shape of the tumor (e.g. sphericity or compactness of the tumor). Group 3 consists of textual features that are able to quantify intra-tumor heterogeneity differences in the texture that is observable within the tumor volume. These features are calculated in all 3-dimensional directions within the tumor volume, thereby taking the spatial location of each voxel compared to the surrounding voxels into account. Group 4 calculates intensity and textural features from wavelet decompositions of the original image, thereby focusing the features on different frequency ranges within the tumor volume. All feature algorithms were implemented in Matlab.

Datasets: In Supplementary II the datasets are described in detail. In short, we applied a radiomic analysis to six image datasets (see overview in Fig. 2).

- RIDER: This dataset consists of 31 non-small cell lung cancer (NSCLC) patients with two CT-scans acquired approximately 15 min apart[10]. We used this dataset to assess stability of the features for test retest.
- Multiple Delineation: This dataset consists of 21 NSCLC patients were the tumor volume was delineated manually on CT/PET scans by five independent oncologists[11]. We used this dataset to assess stability of the features for delineation inaccuracies.
- Lung1: This datasets consist of 422 NSCLC patients that were treated at MAASTRO Clinic, The Netherlands. For these patients CT scans, manual delineations,

clinical and survival data was available. We used this dataset to assess the prognostic value of the radiomic features and to build a radiomic signature.

- Lung 2: This datasets consists of 225 NSCLC patients that were treated at Radboud University Nijmegen Medical Centre, The Netherlands. For these patients CT-scans, manual delineations, clinical, and survival data was available. We used this dataset to validate the prognostic value of the radiomic features and signature in an independent NSCLC cohort.
- H&N1: This dataset consists of 136 head and neck squamous cell carcinoma (HNSCC) patients treated at MAASTRO Clinic, The Netherlands. For these patients CT-scans, manual delineations, clinical, and survival data was available. We used this dataset to validate the prognostic value of the radiomic features and signature in HNSCC patients.
- H&N2: This dataset consists of 95 HNSCC patients treated at the VU Univeristy Medical Center Amsterdam, The Netherlands. For these patients CT-scans, manual delineations, clinical, and survival data was available. We used this dataset to validate the prognostic value of the radiomic features and signature in a second cohort of HNSCC patients.
- Lung 3: This dataset consists of 89 NSCLC patients that were treated at MAASTRO Clinic, The Netherlands. For these patients pre-treatment CT-scans, tumor delineations and gene expression profiles were available. We used this dataset to associate imaging features with gene-expression profiles.

The discovery Lung1 dataset, consisting of CT images for 422 NSCLC patients, and the Lung3 dataset consisting of CT images and gene-expression profiling for 89 NSCLC patients, are publicly available at www.cancerdata.org.

Data Analysis: An overview of the analysis is shown in Figure 2. The analysis was divided in training and validation phases. For the training phase, we first explored feature stability determined in both test-retest and inter-observer setting. The RIDER and Multiple Delineation datasets were used to assess stability of the features to select the most informative features for further investigation. Using the RIDER test retest dataset, we tested the stability of the radiomic features between test and retest[10]. For each patient, we extracted the radiomic features from both scans. A stability rank was calculated for each feature, using the intra-class correlation coefficient (ICC), where a lower ICC rank corresponds to a more stable feature.

We assessed the feature stability for delineation inaccuracies using a Multiple Delineation dataset[11]. All radiomic features were computed for five delineations per patient, and a stability rank per feature was calculated using the Friedman test. The Friedman test is a non-parametric repeated measurement test for a non-Gaussian population. A rank of 1 indicated the most stable feature for delineation inaccuracies and 440 the least stable feature.

All 440 radiomic features were extracted for the Lung1, Lung2, H&N1, and H&N2 datasets. The radiomic features were not normalized on any dataset, and only the raw values were used that were directly computed from the DICOM image.

To explore the association of the radiomics features with survival we used Kaplan-Meier analysis in a training and validation phase. To ensure a completely independent validation, the median threshold of each feature on the Lung1 dataset was computed, and then this threshold was used in the validation datasets (Lung2, H&N1, and H&N2) to split the survival curves. We used the G-rho rank test for censored survival data to test for significant difference between the two survival curves. P-values were corrected for multiple testing using by controlling the false discovery rate (FDR) of 10%, the expected proportion of false discoveries amongst the rejected hypotheses.

To assess the multivariate performance of radiomic features we build a signature. We selected the 100 most stable features, determined by averaging the stability ranks of RIDER dataset and Multiple Delineation dataset. Next, we computed the performance in the Lung 1 dataset of each of the selected 100 features using the concordance index (CI)[12]. This measure is comparable to the Area Under the Curve (AUC) but can also be used for Cox regression analysis. From each of the four feature groups, we selected the single best performing feature for prognosis in the Lung1 dataset, and combined these top four features into a multivariate Cox proportional hazards regression model for prediction survival. The weights of the model were fitted on the Lung1 dataset. We applied the radiomic signature to the validation datasets Lung2, H&N1, and H&N2, and performance was assessed with the CI. To calculate significance between two models we used a bootstrap approach, for 100 times we calculated the CI of both models from 100 random selected samples. The Wilcoxon test was used to assess significance.

A similar approach was used to assess if the signature had significant power, compared with random (CI=0.5). We used a bootstrap approach, for 100 times we calculated the CI of the radiomics signature based on 100 random selected samples with correct outcome data, as well as on 100 random chosen samples with random outcome data. This process was repeated 100 times. The Wilcoxon test was used to assess significance, between the two distributions.

To assess the complementary effect of the signature with clinical parameters, we build a new model with the prediction of the signature as one input and the clinical parameter as the other input. The weight of the clinical parameter was fitted on the training dataset Lung1.

To assess the association of the radiomic signature with gene expression we used the Lung3 dataset. Gene expression of 89 patients was measured on Affymetrix chips with the custom chipset HuRSTA_2a520709 for 21766 genes. Expression values were normalized with the RMA algorithm5 in the Affy package in Bioconductor. For each of the four features in the radiomic signature, we calculated the Spearman rank correlation to gene expression and used the corresponding p-values to obtain a rank of genes representing high to low agreement. Each of these gene ranks were used to perform a pre-ranked ver-

sion of Gene Set Enrichment Analysis (GSEA)[14] on the C5 collection of MSigDB[28], which contains gene sets associated with specific GO terms. We only regarded gene sets of size 15 to 500. Local false-discovery-rates were calculated on the normalized enrichment scores (NES), GSEA's primary statistic, and only gene sets enriched with an FDR of <= 20% were retained. Fig. 4B displays gene sets that have been significantly enriched (FDR <= 20%) for at least one of four radiomic features (indicated by an asterisk). The corresponding absolute NES in all of the four features are given color-coded, where light blue means low and dark blue means high NES.

ACKNOWLEDGMENTS

Authors acknowledge financial support from the National Institute of Health (NIH-USA U01 CA 143062-01, Radiomics of NSCLC), the CTMM framework (AIRFORCE project, grant 030-103), EU 6th and 7th framework program (METOXIA, EURECA, ARTFORCE), euroCAT (IVA Interreg - www.eurocat.info), and the Dutch Cancer Society (KWF UM 2011-5020, KWF UM 2009-4454). Authors also acknowledge financial support from the QuIC-ConCePT project (Grant Agreement No. 115151).

REFERENCES

1. Kurland, B. F. *et al.* Promise and pitfalls of quantitative imaging in oncology clinical trials. *Magn Reson Imaging* **30,** 1301–1312 (2012).
2. Buckler, A. J., Bresolin, L., Dunnick, N. R., Sullivan, D. C.Group. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. *Radiology* **258,** 906–914 (2011).
3. Buckler, A. J. *et al.* Quantitative imaging test approval and biomarker qualification: interrelated but distinct activities. *Radiology* **259,** 875–884 (2011).
4. Lambin, P. *et al.* Predicting outcomes in radiation oncology—multifactorial decision support systems. *Nat Rev Clin Oncol* **10,** 27–40 (2013).
5. Jaffe, C. C. Measures of response: RECIST, WHO, and new alternatives. *J Clin Oncol* **24,** 3245–3251 (2006).
6. Burton, A. RECIST: right time to renovate? *The Lancet Oncology* **8,** 464–465 (2007).
7. Birchard, K. R., Hoang, J. K., Herndon, J. E. & Patz, E. F. Early changes in tumor size in patients treated for advanced stage nonsmall cell lung cancer do not correlate with survival. *Cancer* **115,** 581–586 (2009).
8. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48,** 441–446 (2012).
9. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn Reson Imaging* **30,** 1234–1248 (2012).
10. Zhao, B. *et al.* Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non–Small Cell Lung Cancer. *Radiology* **252,** 263–272 (2009).
11. van Baardwijk, A. *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. *International journal of radiation oncology, biology, physics* **68,** 771–778 (2007).
12. Harrell, F. E. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. (2001).
13. Compton, C. C. *et al. AJCC Cancer Staging Atlas*. (Springer, 2012).

14. Subramanian, A. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102,** 15545– 15550 (2005).
15. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467,** 1114–1117 (2010).
16. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366,** 883–892 (2012).
17. Gerlinger, M. & Swanton, C. How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. *Br. J. Cancer* **103,** 1139–1143 (2010).
18. Kern, S. E. Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. *Cancer Res.* **72,** 6097–6101 (2012).
19. Starmans, M. H. W. *et al.* Independent and functional validation of a multi-tumourtype proliferation signature. *Br. J. Cancer* **107,** 508–515 (2012).
20. Nair, V. S. *et al.* Prognostic PET 18F-FDG uptake imaging features are associated with major oncogenomic alterations in patients with resected non-small cell lung cancer. *Cancer Res.* **72,** 3725–3734 (2012).
21. Diehn, M. *et al.* Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc Natl Acad Sci USA* **105,** 5213–5218 (2008).
22. Segal, E. *et al.* Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat Biotechnol* **25,** 675–680 (2007).
23. Tixier, F. *et al.* Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med* **52,** 369–378 (2011).
24. Naqa, El, I. *et al.* Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit* **42,** 1162–1171 (2009).
25. Ganeshan, B., Panayiotou, E., Burnand, K., Dizdarevic, S. & Miles, K. Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. *European Radiology* **22,** 796–802 (2011).
26. Ganeshan, B., Skogen, K., Pressney, I., Coutroubis, D. & Miles, K. Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis: Preliminary evidence of an association with tumour metabolism, stage, and survival. *Clinical Radiology* **67,** 157–164 (2012).
27. Gevaert, O. *et al.* Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data--methods and preliminary results. *Radiology* **264,** 387–396 (2012).
28. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27,** 1739–1740 (2011).

CHAPTER 11

General discussion and future perspectives

GENERAL DISCUSSION

In the past decade, it has become increasingly clear that cancer is a very complex and heterogeneous genetic disease, where patients with the same diagnosis respond differently to a specific treatment. Large efforts have been oriented towards the development of tailored treatments that can incorporate patient specific characteristics and therefore improve the therapeutic ratio.

Similarly, technological advances led to the development of an ever increasing number of ways to characterize cancer, for example, with non-invasive imaging, or invasively with blood samples and tissue biopsies from which a number of molecular quantities can be measured, individually (e.g. with quantitative polymerase chain reaction) or simultaneously in hundreds, thousands (i.e. microarrays, imaging) and more recently, in millions of molecular entities (DNA sequencing). As a result, the abundance of treatment options and patient specific characteristics, has brought the challenge of Big Data into health care.[1]

It is becoming increasingly clear that cancer and biomedical research are being transformed into computational sciences, where the discovery of evidence largely depends on our ability to collect, analyse and interpret large amounts of biomedical data. There is a growing consensus about the need for new quantitative methods to address the complexities associated with the explosion of medical data.

In radiation oncology, the development of decision support systems, currently focuses on the use of machine learning and statistical methods to derive prognostic and predictive models of outcome.[2-7] Indeed, a large body of literature describes a number of prognostic and predictive models of outcome, based on factors related to the disease and treatment, however, a clear methodology for the assessment of their robustness, accuracy and usefulness is still underway.[4,8]

In the second chapter of this thesis, we discussed the existent prognostic and predictive models in radiation oncology, enumerating the main predictive factors and their challenges in relation to the different data sources, i.e., clinical, treatment (both investigated in Part 1), imaging and molecular factors (investigated in Part 2).

The work presented in this thesis showed that multivariate prognostic models can help medical professionals to determine patient's prognosis more accurately. The use of multivariate models based on clinical patient characteristics and in advanced imaging traits yield better predictions than the traditional staging system TNM and can be validated in external patient cohorts. These models can also be presented in graphical easy-to-use interfaces that facilitate their interpretation and use. However, although this thesis presents the basis of decision support systems in radiation oncology, particularly in lung and head and neck cancer, there are several challenges that must be addressed before these models can be used in radiation oncology practice in a daily basis. Nevertheless, we believe that the integration of validated decision support systems in radiation oncology is inevitable, and will stimulate advances in shared decision making.

Decision support systems in Radiation Oncology

The focus of **Chapter 2** was to define the methodological basis for the development of decision support systems, as well as to draw the steps required to facilitate their use and integration in the clinical settings.

Shared decision making is crucial because any decision made with respect to the patient's treatment, implies identifying the right balance between the benefits and harms. Thus, we proposed that a DSS should ideally simultaneously predict local control, survival, toxicity, quality of life and cost. Currently, most predictive models are able to predict only one of these aspects at the time.

This is simply, because the models are not trained in large enough datasets that include all outcomes of interest and large sub-groups of patients with different treatments. Current prediction models also suffer from uncertainty due to missing data, uncertainty in the studied variables of interest, size and quality of the datasets and intrinsic patient heterogeneity. One should bear in mind that analyses in retrospective cohorts have the advantages of including larger unselected patient cohorts, but with the expense of lower data quality and likely more missing data, as opposed to high quality data from controlled clinical trials.

These sources of uncertainty should be incorporated when providing an individualized prediction for a specific patient. A confidence interval should be always provided when making individualized predictions. However, this is itself a matter of debate as one could ask whether makes sense to indicate an area of uncertainty in an already uncertain prediction. Still, a confidence interval should always accompany a model prediction, to facilitate its interpretation as well as to discern whether its predictions provide meaningful and statistically sound differences in clinical outcomes.[9-11]

Furthermore, the current stage of decision support systems relies on the analysis of single features, or combinations of features from a single data source. It is expected that the combination of multimodal variables into multivariate models, will provide a more comprehensive view of the patient's response to treatment and will increase its ability to accurately predict the patient's response.[12,13]

Altogether, this emphasizes the need for standardized, systematic, multimodal data collection and storage in the routine clinical practice, as has been recently proposed.[14] Systematic medical data collection and storage or the addition of new parameters to existent predictive models, will not per se increase prediction accuracy, but it will facilitate the continuous learning in larger and more diverse patient datasets, which in turn, could improve prediction accuracy.

This is supported by rapid learning,[1,15] a concept that allows knowledge extraction from retrospective data, its constant update with new datasets or data sources, that incorporate the up to date medical evidence and knowledge, to eventually adapt treatment protocols accordingly and provide personalized decision support.

Large multicentre efforts have been started in this direction, for instance, the Computer-Aided Theragnostics network,[16] GENEPI,[17] the Radiogenomics Consortium,[18] ALLEGRO,[19] ULICE[20] and the Cancer Genome Atlas,[21] are examples of international initiatives to systematically collect data and tissue of routine oncology patients. These initiatives will allow the external validation of predictive models and facilitate research reproducibility and the evaluation of new hypotheses.[22] Furthermore, efforts on data quality assurance programmes, semantic standardization and secure data transfer have been initiated.[23] In conclusion, despite the many current challenges on the development of validated decision support systems, numerous steps have been started to incorporate health information systems in the routine clinical care, which we believe, will improve the efficiency and efficacy of cancer care.

PROGNOSTIC MODELS BASED ON CLINICAL PREDICTORS

Prediction of residual metabolic activity in NSCLC

Residual metabolic activity within the primary tumor after treatment, has been proposed as a surrogate for survival in NSCLC patients. Indeed, several studies have shown that patients with residual metabolic active areas after treatment, have considerably poorer prognosis compared to patients without.[24,25] In Chapter 3, we studied the relationship of clinical, demographic and tumor characteristics, with metabolic response in NSCLC patients.

Our results showed that the assessment of tumor metabolic activity on a post-treatment FDG-PET-CT scan is a useful tool to determine persistent disease and showed that residual disease is a validated surrogate for survival.

The proposed multivariate model yielded a relatively high prediction accuracy (AUC of 0.71; 95% CI, 0.65-0.76), however the lack of validation in an external cohort is a clear limitation to this study. Additionally, we were not able to analyze potential prognostic variables such as molecular markers or imaging surrogates [33-35] that may improve the ability of the presented model to predict the post- treatment failure.

Other studies from our group have investigated if the location of metabolic-active areas after treatment are the most radio-resistant areas, and whether they can be identified before treatment starts using a PET-CT scan.[26] They have shown that high uptake areas before treatment largely co-localize with the residual active areas after treatment. They also hypothesize that these recurrent areas are an indication of treatment resistance. The prognostic model proposed in **Chapter 3**, could be coupled with the pre-treatment identification of areas with high metabolic activity, and therefore, provide an assessment of a patient chances of residual disease as well as its location, which could provide a stronger indication for dose escalation by dose painting,[27] in patients with high risk of residual disease.

In **Chapter 3**, we confirmed that by using an FDG-PET-CT scan, patients who do not respond to radiotherapy can be identified early in the course of their treatment, and the proposed prognostic model could assist clinicians in deciding whether a patient with predicted residual disease is a candidate for dose escalation in a clinical trial. However further validation is warranted.

Prognostic nomograms in head and neck cancer

To elucidate the influence of clinical, patient and tumor characteristics on outcome of patients with laryngeal carcinoma, in **Chapter 4**, we investigated and developed a prognostic nomogram for survival and local control. This nomogram, developed in a large cohort of unselected laryngeal cancer patients, and validated in four external datasets, was able to provide individualized predictions of outcome and furthermore, was able to assign patients to clearly distinct risk groups.

An important step to show the usefulness of prognostic nomograms such as the one presented in **Chapter 4**, is to show, as was the case here that the prognostic nomograms perform better than models based on TNM.

A limitation of this study, due to its retrospective nature, which again highlights the importance of systematic data collection for rapid learning, was the lack of co-morbidity data in these patients. It is known that co-morbidity in laryngeal carcinoma patients is high and influences overall survival more than the cancer itself. In future studies, co-morbidity should be investigated as prognostic factor.

This limitation also goes along the lines that a prognostic model solely based on clinical features is too limited to allow predictions with high accuracy and to allow clinical decision making. A prospective multicentre trial with systematic collection of tissue, imaging studies, clinical data as well as normal tissue reactions, would allow the incorporation of biologic and imaging prognostic factors into the baseline clinical nomograms, and would facilitate validation and extension of the current results.

Nevertheless, the ability of the presented nomogram to stratify patients into clearly distinct risk groups, could already potentially be used to customize more aggressive strategies for patients with high risk of relapse.

Furthermore, ongoing international efforts have been made to validate the presented nomogram in RTOG dataset of laryngeal carcinoma patients that compared concurrent chemotherapy and radiotherapy for organ preservation in advanced larynx cancer.[28] If the presented nomogram validates also in the RTOG-9111 dataset, we could have an argument in favour that learning on retrospective datasets, despite the inherent issues on data quality, is useful and the derived knowledge can be applied to datasets coming from controlled clinical trials.

In **Chapter 5**, a similar study was carried out to identify the most important determinants of treatment outcome in patients with oropharyngeal carcinoma.

The combination of the most important prognostic factors in a multivariate model, including HPV status, comorbidity and smoking, known important prognostic factors in this patient population, yielded high predictive performances, for overall survival and progression free survival.

Once again, the combination of individual important predictors in multivariate models, generated significantly better predictions than using TNM or HPV alone. The 95% CI of the model predictions were significantly better than those obtained with TNM alone or, for example, HPV status alone, which underlines the importance of multifactorial prediction models.

Similar to other studies[29] we observed that overall survival and progression-free survival were significantly better for patients with an HPV-positive OPSCC compared to patients with an HPV-negative OPSCC with the 5-year overall survival rates of 82 % in the HPV-positive group and 39% in the HPV-negative group.

A limitation in our study, inherent to its retrospective nature is the lack of standardization in which data has been collected over the years. This naturally affects the quality of data with respect to patient reported smoking and alcohol consumption as well as completeness of data regarding comorbidity. Nevertheless, the model predictions were validated in an external patient cohort.

This highlights the increasing need for systematic routine patient care data collection, warehouse and semantic inter-operable data retrieval systems, to assure improved and standardized data retrieval and allow external applicability.[30,31]

This model performance is acceptable for clinical support, particularly due to the clear distinction in risk groups, in both cohorts; however it is still far from optimal. Combining clinical parameters with an important OPSCC biomarker such as HPV, is a first step into developing validated decision support systems in head and neck cancer; however we anticipate that adding other features, such as diagnostic and molecular imaging, and other important biomarkers such as EGFR or CAIX will increase the model accuracy.[32]

Finally, we consider that is important to disseminate and to facilitate the use of prediction models. Therefore, the published prognostic models, are made available on the www.predictcancer.org website, along with published prognostic models in other cancer sites.[3,5,33] In this way, radiation oncologists, everywhere, can enter data of new patients and obtain an online calculation of probability of outcome and risk group stratification. Naturally, treatment decisions cannot be solely based on the predictions of these models, and therefore, they are intended as supportive information. Sufficient information on the model development, required input data and interpretation is provided in the www.predictcancer.org website, to enable the correct interpretation of the model's predictions.

Radiomics: advanced feature extraction from medical images

In Part 1 of this thesis, we anticipated that the combination of multimodal prognostic factors in multivariate models will improve the prognostic and predictive information as compared to factors from individual sources. In Part 2 of this thesis, we presented the Radiomics concept. The central hypothesis of the Radiomics approach, is that more information of what is currently used, can be quantified from conventional medical images through the application of advanced image analysis algorithms.

We hypothesized that quantitative parameters extracted from routine medical images convey prognostic information and capture underlying genomic and proteomic patterns at the imaging level.

Although the analysis of medical images is not novel, since a large body of literature exists, for instance in computer-aided diagnosis; however CAD has focused mainly on the detection of small nodules in e.g. mammograms, or the differential diagnosis between malignant and benign nodules.[34,35] Nevertheless, CAD literature already describes a large number of quantitative imaging features.

The novel idea in the Radiomics approach is to define the methodological workflow that starts with the acquisition of a high quality scan, the robust definition of the target of interest, the high-throughput quantification of imaging traits and finally the association of the most informative imaging traits with treatment outcomes or biologic data.

Chapter 6, provides an overview of the existent work on CT, MR and PET images, quantifying imaging patterns and associating them with treatment outcome in ovarian, esophagus, cervix and head and neck cancer.[36-38]

Pioneering work on radiogenomics, a field that associates imaging features with gene expression patterns, has demonstrated that major differences in gene expression patterns within a tumor are correlated with qualitative radiographic findings in hepatocellular carcinoma and glioblastoma multiforme.[39]

The discussed literature in **Chapter 6**, opens the question whether moving radiological sciences to a computational science with the high-throughput extraction of quantitative imaging features on routine imaging improves the ability of currently used imaging parameters to predict or monitor response to treatment.

Multicenter efforts have been initiated to address and mitigate the different sources of noise in the Radiomics workflow,[40,41] as well as to confirm experimentally the Radiomics hypothesis, that is, to establish a causal relationship between gene expression patterns and imaging features.[42]

The work presented in this dissertation, provides the initial basis of Radiomics applied to radiation oncology. In **Chapters 7 and 8** we evaluated algorithms to robustly segment lung tumors in CT images.

A step further, in **Chapter 9**, we compared quantitative imaging features extracted from CT images using semiautomatic segmentation with features extracted from manual tumor delineations. Our results show that semiautomatically segmented tumor volumes provide

a better alternative to the manual delineation process, as they are more robust for quantitative image feature extraction.

In **Chapter 10**, to our knowledge, the largest integrated Radiomics analysis is presented, quantifying phenotypic differences in CT images of more than 1000 patients with lung or head and neck cancer.

Semi-automatic lung tumor segmentation

In **Chapter 7 and 8** of this work, we evaluated semi-automatic methods to segment lung tumors in NSCLC patients. Both methods were compared against manual delineations of multiple radiation oncologists, and against the largest diameter measured on the resected tumor, considered the ground truth.

It has been widely documented that human interpretation of visual imaging information remains as the largest source of variability for target definition. Especially for lung cancer, high intra and inter observer variability has been observed for target definition.[43,44]

We believe that the algorithms evaluated here have clinical value since both provided segmentations statistically equivalent to those delineated by the clinical experts and also agreed largely with the pathological tumor dimensions. Furthermore, they also reduced drastically the delineation time as compared to manual delineations.

The algorithm evaluated in **Chapter 8** presents an additional advantage as it is open source, and freely available with download, and thus, accessible to a wider community.
A step further in the analysis, in **Chapter 8**, we demonstrated that semi-automatic segmentation showed significantly lower volume variability and smaller uncertainty areas, as compared to the CT/PET manual delineations. Both algorithms showed robustness towards user initialization, however it became apparent that the human factor is not fully negligible and human interaction is still the largest source of variation, even with the aid of semi-automatic contouring tools. Future studies should be oriented towards minimizing human interaction in the delineation process. Nevertheless, a medical expert should supervise auto-segmentation algorithms in all cases.

It is important to acknowledge that in future studies, 4D CT scans should be employed to diminish the blurring effects produced by tumor motion. We also anticipate the continuing efforts in this field will be on the development of segmentation algorithms that combine CT and PET information, although not all centers are equipped with PET scanners.

The results presented in **Chapters 7 and 8**, provide sufficient evidence allow the use of these methods as a starting point for treatment planning delineations and in high-throughput data mining research, such as Radiomics, where manual tumour delineations are often not available, or represent a considerable time consuming bottleneck.

Radiomics features and volumetric segmentation

Accurate tumor delineation is also essential to ensure the reliability of quantitative imaging features. If imaging features were to be used as prognostic or predictive factors, it is essential to determine their variability with respect to the tumor delineation process.

In **Chapter 9**, we followed the results obtained in **Chapter 7 and 8**, and evaluated whether quantitative imaging features extracted from CT images using semiautomatic tumor segmentation are more robust than those extracted from manual tumor delineations. Radiomics features extracted from 3D-Slicer segmentations had significantly higher reproducibility compared to the features extracted from the manual segmentations. Furthermore, we found that features extracted from 3D-Slicer segmentations were more robust, as the range was significantly smaller across observers, and overlapping with the feature ranges extracted from manual contouring.

These results show that 3D-Slicer segmented tumor volumes provide a better alternative to the manual delineation process, as they are more robust for quantitative image feature extraction.

A limitation in this study was not being able to associate these image descriptors with patient outcome due to cohort size and unavailability of clinical data, to investigate the hypothesis that more robust features have a stronger association with patient outcome. This hypothesis is evaluated however in **Chapter 10**, the first radiomics analysis in over 1000 patients with several independent validation datasets.

Radiomics: decoding the tumor phenotype by non-invasive imaging

Medical imaging is one of the major disciplines that have informed medical science and treatment. However, although medical images of tumors exhibit strong phenotypic differences between patients, only simple one dimensional descriptors of size or uptake as in the case of PET imaging, are currently used.

In **Chapter 10**, we explored 440 radiomic features in 1019 cancer patients demonstrating that quantitative imaging biomarkers have strong prognostic performance in large independent cohorts of lung and head and neck cancer, and are associated with the underlying gene expression patterns. We also showed that the features with a highest reproducibility against manual delineation and test-retest, have the strongest prognostic value.

Overall, the results presented in **Chapter 10**, indicate that Radiomic features decode a prognostic phenotype that could be translated from lung to head and neck cancer, which may also generalize to other tumor types. These results indicate that radiomics quantifies a general prognostic cancer phenotype that likely can broadly be applied to other cancer types.

Furthermore, we showed the complementary and improved performance of Radiomic features compared to TNM staging for prediction of outcome, which illustrates the clinical importance of our findings as TNM is routinely used in the clinic.

Currently, the TNM staging system is used for risk stratification and treatment decision-making. However, the TNM staging system is primarily based on resectability of the tumor, while 30% - 35% of NSCLC patients will receive primary treatment with radiotherapy either alone or combined with chemotherapy. Therefore, it is widely known that the TNM staging system is inadequate for risk stratification of this group of patients, in particular to make the decision between curative treatment (concomitant radio-chemotherapy) or palliative treatment especially in elderly patients, a growing issue in western countries. Our results show that the radiomics signature is performing better in independent cohorts then the TNM classification. In future clinical trials this inexpensive method can be used as well for pretreatment risk stratification (e.g. high, low risk).

It is expected that Radiomics will have an impact in personalized medicine, as it provides complementary phenotypic information compared to genomic information.
In comparison to biopsy based genomic assays, non-invasive imaging and radiomics describe the phenotype of the entire tumor, and provide complementary information to biopsy-based assays that probe only small portions of tumor tissue, and do not allow for a complete characterization of the tumor. Furthermore, acquiring a radiomics signature is non-invasive and adds no additional costs for the clinical application of our methods, as opposed to genomics or proteomics, which are still challenging to implement into clinical routine, due to the high costs and invasiveness, especially in lung cancer.
Finally, due to the large application of non-invasive imaging in cancer patients in almost all hospitals worldwide, the results of this work will stimulate the development of image-based biomarkers and their evaluation in retrospective and prospective studies.[41,45,46]

FUTURE PERSPECTIVES

The key factor in the future development of decision support systems is standardization. Standardization of data through the application of protocols and semantic-interoperability through different centers; and the creation of a framework for reproducible data sharing, research and predictions. Locally this can be started with centralizing and making transparent all data and methodologies employed in the development of prognostic or predictive models. This also will allow the re-use of data for validation purposes or for the analysis of new hypothesis.[30]

Across centers, several initiatives have been initiated for the systematic collection of patient information of various data types.[16,20,21] These efforts will allow external validation of predictive algorithms and facilitate research reproducibility and the evaluation of new research questions.
The Computer Aided Theragnostics Network, aims to create a regional infrastructure to enable standardized data sharing through the participant medical centers and to apply and develop decision support software by the continuous learning in shared databases.[16]

This initiative has rapidly expanded outside the region to centers in the USA, EU, China and Australia.

Infrastructures like euroCAT, will enable a crucial step for quality assurance of data collection and standardization, which is the semantic interoperability through different centers and the automated storage of patient, tumor and treatment characteristics. In the future, this will tackle all the drawbacks inherent to retrospective data collection and analysis. Furthermore, imaging data from centralized imaging repositories in each participant center, can be decentralized and shared through standardized de-identification, image annotation and storage, which adds imaging as an extra dimension for the development of prognostic algorithms. All these efforts can be coupled with the already established mechanisms for the standardization of imaging studies.[13,47,48]

The results of part 1, particularly the prognostic nomograms for patients with laryngeal and oropharyngeal carcinomas showed that outcome prediction in these patients is improved with these prognostic nomograms.

The outlook for these nomograms will be the addition of other features, such as diagnostic and molecular imaging features, or other important biomarkers such as EGFR or CAIX will increase the model accuracy. Furthermore, these models should also be validated in other external datasets. These models set the basis for incorporating decision support in the clinic, to eventually stratify patients according to their estimated risk using a nomogram, and their assigned eligibility for different treatment approaches.[49] Ongoing trials are evaluating treatment de-intensification or escalation based on the patient estimated good or poor prognosis i.e. NCT01663259.[50,51]

In summary, the outlook on the development of decision support systems relies in assessing their clinical usefulness, standardizing the methodologies employed in the model development, standardizing data collection and annotation and eventually supporting the design of clinical trials. These crucial steps are the basis of validated decision support systems, which, will stimulate developments in rapid learning and enable advances in shared decision making.

In Part 2, we stated that Radiomics analyses can have a large impact in personalized medicine. This is because it provides complementary information compared to clinical and genomic data, and the extensive use of non-invasive imaging in all stages of cancer management brings endless possibilities to the evaluation and development of imaging biomarkers.

With Radiomics as well, a key issue is the standardization of imaging acquisition protocols and efforts to reduce human error in target definition. In radiation oncology, image acquisition has already been largely standardized, however this is not the case for diagnostic imaging. Future studies should elucidate the impact of voxel size, acquisition parameters, contrast agents and algorithm settings in the extraction of imaging features.

We expect that with better standardization and protocols, the prognostic power of radiomics will even further improve. The Quantitative Imaging Network (QIN) of the NIH

investigates future directions, by looking at phantom studies and discussing with vendor's open and standardized protocols for image acquisition.

It is also crucial to assure transparency in the implementation of feature extraction algorithms. Therefore, we have developed a feature extraction pipeline implemented in the open source CERR package,[52] to facilitate its dissemination and access to the scientific community.

The results presented in **Chapter 10**, demonstrated that advanced imaging features are prognostic for survival in NSCLC and HNSCC and associated with gene expression in NSCLC. These results stimulate the evaluation of radiomic analysis in other cancer sites for example i.e. rectum, glioblastoma multiforme; the evaluation of other imaging modalities such as MR and PET, the investigation of changes in radiomic features during treatment and shortly after treatment and eventually the radiomics analysis of normal tissues. The possibilities are endless. A crucial step in the validation of imaging biomarkers will be to investigate their association with DNA mutation profiles, gene expression profiles and patient outcomes. These analysis are now possible due to the advances in DNA sequencing, however, also bring new methodological challenges to the analysis of these biomedical BigData.

Multicenter efforts have been initiated for the development and evaluation of imaging biomarkers, including their pre-clinical and clinical evaluation, and their association with underlying biological aspects.[41,45,53]

In conclusion, these studies show how the analysis of existent routine clinical and imaging data, will facilitate personalized medicine in radiation oncology.

REFERENCES

1. Abernethy, A.P.*, et al.* Rapid-learning system for cancer care. *J Clin Oncol* **28**, 4268-4274 (2010).
2. Borst, G.R.*, et al.* Standardised FDG uptake: A prognostic factor for inoperable non-small cell lung cancer. *European Journal of Cancer* **41**, 1533-1541 (2005).
3. Dehing-Oberije, C.*, 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 Biol Phys* **74**, 355-362 (2009).
4. Steyerberg, E.W.*, et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128-138 (2010).
5. Valentini, V.*, 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* **29**, 3163-3172 (2011).
6. van Stiphout, R.G.*, 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* **98**, 126-133 (2011).
7. Dekker, A., Dehing-Oberije, C., De Ruysscher, D. & al., e. Survival Prediction in Lung Cancer Treated with Radiotherapy: Bayesian Networks vs. Support Vector Machines in Handling Missing Data. in *IEEE Computer Society* 494–497 (Los Alamitos, CA, USA, 2009).

8. Lambin, P.*, et al.* Predicting outcomes in radiation oncology--multifactorial decision support systems. *Nat Rev Clin Oncol* **10**, 27-40 (2013).
9. Iasonos, A., Schrag, D., Raj, G.V. & Panageas, K.S. How to build and interpret a nomogram for cancer prognosis. *J Clin Oncol* **26**, 1364-1370 (2008).
10. Vickers, A.J. Prediction models: revolutionary in principle, but do they do more good than harm? *J Clin Oncol* **29**, 2951-2952 (2011).
11. Ludbrook, J. Outlying observations and missing values: how should they be handled? *Clin Exp Pharmacol Physiol* **35**, 670-678 (2008).
12. Hudson, T.J.*, et al.* International network of cancer genome projects. *Nature* **464**, 993-998 (2010).
13. Gillies, R.J., Anderson, A.R., Gatenby, R.A. & Morse, D.L. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clinical Radiology* **65**, 517-521 (2010).
14. Friedman, C. & Rigby, M. Conceptualising and creating a global learning health system. *International Journal of Medical Informatics* **82**, e63-e71 (2013).
15. Ginsburg, G.S., Staples, J. & Abernethy, A.P. Academic medical centers: ripe for rapid-learning personalized health care. *Sci Transl Med* **3**, 101cm127 (2011).
16. http://www.eurocat.info.
17. De Ruysscher, D.*, et al.* First report on the patient database for the identification of the genetic pathways involved in patients over-reacting to radiotherapy: GENEPI-II. *Radiother Oncol* **97**, 36-39 (2010).
18. West, C.*, et al.* Establishment of a Radiogenomics Consortium. *Int J Radiat Oncol Biol Phys* **76**, 1295-1296 (2010).
19. Ottolenghi, A., Smyth, V. & Trott, K.R. The risks to healthy tissues from the use of existing and emerging techniques for radiation therapy. *Radiat Prot Dosimetry* **143**, 533-535 (2011).
20. Kessel, K.A.*, et al.* Connection of European particle therapy centers and generation of a common particle database system within the European ULICE-framework. *Radiat Oncol* **7**, 115 (2012).
21. http://cancergenome.nih.gov/.
22. Deasy, J.O.*, et al.* Improving normal tissue complication probability models: the need to adopt a "data-pooling" culture. *Int J Radiat Oncol Biol Phys* **76**, S151-154 (2010).
23. Digital Agenda for Europe [cited 2012 May 1]. Available from: http://ec.europa.eu/information_society/digital-agenda/index_en.htm
24. Mac Manus, M.P.*, et al.* Metabolic (FDG-PET) response after radical radiotherapy/chemoradiotherapy for non-small cell lung cancer correlates with patterns of failure. *Lung Cancer* **49**, 95-108 (2005).
25. Decoster, L.*, et al.* Complete metabolic tumour response, assessed by 18-fluorodeoxyglucose positron emission tomography (18FDG-PET), after induction chemotherapy predicts a favourable outcome in patients with locally advanced non-small cell lung cancer (NSCLC). *Lung Cancer* **62**, 55-61 (2008).
26. Aerts, H.J.*, et al.* Identification of residual metabolic-active areas within individual NSCLC tumours using a pre-radiotherapy (18)Fluorodeoxyglucose-PET-CT scan. *Radiother Oncol* **91**, 386-392 (2009).
27. Bentzen, S.M. & Gregoire, V. Molecular Imaging–Based Dose Painting: A Novel Paradigm for Radiation Therapy Prescription. *Seminars in Radiation Oncology* **21**, 101-110 (2011).
28. Forastiere, A.A.*, et al.* Concurrent chemotherapy and radiotherapy for organ preservation in advanced laryngeal cancer. *The New England journal of medicine* **349**, 2091-2098 (2003).
29. Petrelli, F., Sarti, E. & Barni, S. Predictive value of HPV in oropharyngeal carcinoma treated with radiotherapy: An updated systematic review and meta-analysis of 30 trials. *Head Neck* (2013).
30. Lambin, P.*, et al.* 'Rapid Learning health care in oncology' - An approach towards decision support systems enabling customised radiotherapy'. *Radiother Oncol* **109**, 159-164 (2013).
31. Roelofs, E.*, et al.* Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. *Radiother Oncol* **108**, 174-179 (2013).
32. Lassen, P., Overgaard, J. & Eriksen, J.G. Expression of EGFR and HPV-associated p16 in oropharyngeal carcinoma: Correlation and influence on prognosis after radiotherapy in the randomized DAHANCA 5 and 7 trials. *Radiother Oncol* (2013).
33. http://www.predictcancer.org.

34. Dehmeshki, J., Amin, H., Valdivieso, M. & Ye, X. Segmentation of pulmonary nodules in thoracic CT scans: a region growing approach. *IEEE Trans Med Imaging* **27**, 467-480 (2008).
35. Armato, S.G., 3rd, Giger, M.L. & MacMahon, H. Automated detection of lung nodules in CT scans: preliminary results. *Med Phys* **28**, 1552-1561 (2001).
36. O'Connor, J.P.*, et al.* Enhancing fraction predicts clinical outcome following first-line chemotherapy in patients with epithelial ovarian carcinoma. *Clin Cancer Res* **13**, 6130-6135 (2007).
37. Westerterp, M.*, et al.* Esophageal cancer: CT, endoscopic US, and FDG PET for assessment of response to neoadjuvant therapy--systematic review. *Radiology* **236**, 841-851 (2005).
38. Tixier, F.*, et al.* Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med* **52**, 369-378 (2011).
39. Rutman, A.M. & Kuo, M.D. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. *Eur J Radiol* **70**, 232-241 (2009).
40. Kumar, V.*, et al.* Radiomics: the process and the challenges. *Magn Reson Imaging* **30**, 1234-1248.
41. Buckler, A.J., Bresolin, L., Dunnick, N.R. & Sullivan, D.C. A collaborative enterprise for multi-stakeholder participation in the advancement of quantitative imaging. *Radiology* **258**, 906-914 (2011).
42. Lambin, P.*, et al.* Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer* **48**, 441-446 (2012).
43. Steenbakkers, R.J.*, et al.* Observer variation in target volume delineation of lung cancer related to radiation oncologist-computer interaction: a 'Big Brother' evaluation. *Radiother Oncol* **77**, 182-190 (2005).
44. Greco, C., Rosenzweig, K., Cascini, G.L. & Tamburrini, O. Current status of PET/CT for tumour volume definition in radiotherapy treatment planning for non-small cell lung cancer (NSCLC). *Lung Cancer* **57**, 125-134 (2007).
45. Buckler, A.J.*, et al.* Quantitative imaging test approval and biomarker qualification: interrelated but distinct activities. *Radiology* **259**, 875-884 (2011).
46. Gatenby, R.A., Grove, O. & Gillies, R.J. Quantitative imaging in cancer evolution and ecology. *Radiology* **269**, 8-15 (2013).
47. Boellaard, R.*, et al.* FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur J Nucl Med Mol Imaging* **37**, 181-200 (2010).
48. Boellaard, R.*, et al.* The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials. *Eur J Nucl Med Mol Imaging* **35**, 2320-2333 (2008).
49. O'Sullivan, B.*, et al.* Deintensification candidate subgroups in human papillomavirus-related oropharyngeal cancer according to minimal risk of distant metastasis. *J Clin Oncol* **31**, 543-550 (2013).
50. http://clinicaltrials.gov/.
51. Ang, K.K.*, et al.* Human papillomavirus and survival of patients with oropharyngeal cancer. *N Engl J Med* **363**, 24-35 (2010).
52. Deasy, J.O., Blanco, A.I. & Clark, V.H. CERR: a computational environment for radiotherapy research. *Medical physics* **30**, 979-985 (2003).
53. http://www.quic-concept.eu/.

Summary

SUMMARY

The development of decision support systems to predict patients outcome in radiation oncology is needed to facilitate clinical shared decision making. Therefore, this thesis investigates the development of prognostic models for lung and head and neck cancer patients to identify patients at different risk levels before treatment.
The work of this thesis is divided in two sections; in the first section, this thesis investigated the integration of patient clinical and treatment characteristics in lung and head and neck cancer to derive prognostic models of outcome. In the second part, we investigated the use of advanced quantitative imaging features, extracted from conventional medical imaging, many of which are not currently used, to improve patient's prognostic information in lung and head and neck cancer.

The concept of decision support systems in radiation oncology is introduced in **Chapter 1**, along with a general introduction of the work presented in this thesis.
In **Chapter 2**, a comprehensive review of current factors that have been associated with outcomes in radiation oncology is presented, as well as a thorough discussion of the methodology needed for the development of prediction models.

Part 1: Clinical predictors

This section presents the development of predictive models using patient clinical and treatment characteristics. Here, we investigated what factors have a relevant association with patient's outcome. Also, we evaluated whether these models can be validated in external datasets and how their prediction accuracy compares with currently used factors to assess patient's prognosis.

In lung cancer, early identification of patients with residual metabolic activity is essential as this enables selection of patients who could potentially benefit from additional therapy. **Chapter 3** presents a study in which we evaluated the most important patient, tumor and treatment factors associated with residual metabolic activity after treatment. Metabolic response assessment has been associated with survival and treatment failure. This study was performed in a MAASTRO dataset of 101 NSCLC patients.

Chapters 4 and 5 are focused on the development of predictive nomograms in head and neck cancer. **Chapter 4** describes the development of a prognostic nomogram for the prediction of overall survival and local control in laryngeal carcinoma patients treated with radiotherapy. It also shows the validation of the same prognostic tool in four external datasets. This model has been made publicly available on www.predictcancer.org. In **Chapter 5,** a similar nomogram was developed for oropharyngeal carcinoma patients for prediction of overall survival and progression-free survival. This model combines important patient and tumor characteristics with the human papilloma virus status, an established important

prognostic factor in this patient population. It also shows a comparison of the developed nomogram with TNM and HPV status alone, and its validation in an independent patient cohort.

Part 2: Radiomics: Extracting more information from medical images using advanced feature analysis

This second part of the thesis deals with the potential of extracting advanced quantitative features from medical images for outcome prediction. We investigated whether there is more information in medical imaging than what is currently used, and if quantitative imaging traits are prognostic in lung and head and neck cancer.
In **Part 2** of this thesis we proposed a methodology for high-throughput extraction of quantitative imaging parameters, evaluated methods for robust target definition and assessed the prognostic value of these imaging parameters in lung and head and neck cancer cohorts.

Chapter 6, puts forward the concept of Radiomics: the high-throughput extraction of large amounts of image features from radiographic images. This review addresses the hypothesis, methodological aspects, and challenges underlying the Radiomics approach.
Towards the extraction of reproducible quantitative imaging features, **Chapter 7**, evaluates the relevance of a semiautomatic CT-based segmentation method, by comparing it to manual delineations made by radiation oncologists and to pathological tumor measurements considered as “gold standard” in NSCLC patients.

Chapter 8, evaluates an open source, freely available method for lung tumors segmentation, and evaluates its usefulness by comparing it again, against the gold standard pathological measurements and radiation oncologists delineations, and a step further, examining whether its use reduces variability during tumor segmentation.
Following these results, **Chapter 9** evaluates whether quantitative imaging features extracted from semi-automatically segmented tumors have lower variability and are more robust compared to features extracted from manual tumor delineations. This study analyzes the robustness of imaging features derived from semi-automatically and manually segmented primary NSCLC tumors in twenty patients.

Chapter 10, presents an analysis of 440 quantitative imaging features quantifying phenotypic differences based on tumor appearance, i.e., shape, intensity and texture, in CT images of more than 1000 patients with lung or head and neck cancer. In this study we found that a large number of radiomic features have prognostic power in independent data sets, many of which were not identified as significant before. Radiogenomics analysis revealed that a prognostic radiomic signature, capturing intratumour heterogeneity, is associated with underlying gene-expression patterns. These results can have a clinical impact as imaging is routinely used in clinical practice, providing an unprecedented opportunity to improve decision-support in cancer treatment at low cost.

Finally, in Chapter 11, the results presented in this thesis are discussed along with its outlook and future perspectives. These studies show how the analysis of existent routine clinical and imaging data, will facilitate personalized medicine in radiation oncology. However, the key factor in the future development of decision support systems is standardization. Standardization of data through the application of protocols and semantic-interoperability through different centres; and the creation of a framework for reproducible data sharing, research and predictions.

Valorization Addendum

VALORIZATION ADDENDUM

The central idea of the work included in this thesis is to extract knowledge from medical data (of various sources), which in turn could be used to guide medical decisions in the complex process of identifying the most effective treatment for a cancer patient. Thus, valorization of the knowledge derived from this work is essential.

The simplest and perhaps most effective way to share and make this knowledge available to a wider scientific and medical community is by making the entire knowledge discovery process publicly available, thus, additionally to publishing this research in international peer-reviewed journals, data, software and developed models are made publicly available. Below specific examples are mentioned.

Part 1

This part of the thesis focused on the development of prediction models for radiation oncology. These models are of interest to physicians and health professionals who want to have processed information about the risks and benefits of a cancer treatment, in terms of response and follow-up outcomes.

The model presented in Chapter 4, which has been validated in independent patient cohorts has been made available in the website www.predictcancer.org. There, the patient population and statistical analysis on which the model is based are described.
Any physician around the world can access the model web version and obtain a risk probability for an individual patient. These estimates are based on individual information entered about patients and their tumors. The estimates are provided on printed sheets in simple graphical and text formats. The site, however, is not intended for patients, but for healthcare professionals that can discuss the model outcomes with a patient, oriented towards assisting shared decision-making in radiation oncology.
Furthermore, the model presented in chapter 4, was published in a special 100 anniversary issue of the Radiotherapy and Oncology journal and was featured in AuntMinnie, a radiology and radiation oncology news site: http://www.auntminnie.com/-index.aspx?sec=sup_n&sub=roc&pag=dis&ItemID=96200&wf=1236.

The model presented in Chapter 5 of this thesis, will be made available in the www.predictcancer.org website after publication. All of these models can also be implemented as a smartphone app, that physicians can access offline as well.
This prediction website facilitates knowledge dissemination, and we believe that our findings are relevant for a broad scientific community because we anticipate that validated decision-support systems will be fully integrated in the clinic, with data and knowledge being shared in a standardized, instant and global manner.

Following this, data in which the models are based is also publicly available in www.cancerdata.org. Thus, scientific publications, data and developed models are made publicly available, in an effort to widespread knowledge and to facilitate the use of deci-

sion support systems in radiation oncology, often a bottle-neck, since many complex models are often not easily accessible nor easy to use or interpret. This also will further promote the concept of shared-decision making in radiation oncology and personalized medicine.

Part 2

This part of the focuses on the use of quantitative imaging characteristics extracted from standard clinical scans to predict patient's outcome in lung and head and neck cancer.

In the Radiomics methodology proposed in this thesis, three independent components are targets for valorization. The first one is on the methods employed to segment tumors on CT images. The method employed in Chapter 7 is an already commercial product and therefore its application is limited. The method evaluated in chapter 8 is part of an open-source, image analysis software, easily accessible by download, which facilitates its use and dissemination.

The second component, is the software and know-how developed to extract quantitative imaging features from medical images. This can be requested already in the www.radiomics.org website. More interestingly, this second step of the analysis could be coupled with the open source 3D-Slicer software, which would allow image segmentation, visualization and quantification using a single open-source tool, with a wide range of image analysis tools available additionally from 3D-Slicer.

The third step would be integrating the modelling outcomes or model output steps as well into the same platform. This as a whole, would be an integrated radiomic analysis software that could be implemented in treatment planning workstations in the clinic.

Furthermore, the work presented in chapter 10, allowed for the first time the evaluation of more than 400 quantitative imaging features in over a 1000 patients with lung and head and neck cancer, with both training and several independent validation datasets. Current publications investigating imaging features with survival and gene-expression have only limited sample sizes, often without decent sample sizes or lacking validation cohorts. We have shown for the first time the true independent validated impact of radiomics in both lung and head and neck cancer, showing the translational capability of our findings.

Radiomics allows the discovery of imaging biomarkers of low cost, easy to perform, non-invasive and that captures the 3D complexity of solid tumors. By applying the radiomics software it is now possible to reveal biological information from standard clinical scans. The radiomics impact on personalized medicine, relies on the ability to predict the behavior of the tumor before the treatment begins.

This method can reveal more about tumor behavior than traditional clinical methods for tumor staging, just by processing the scans through radiomics software. Therefore, chapter 10 represents a novel high-throughput analysis of imaging features, applying computational methodologies to the analysis of radiographic images. Particularly in cancer

management, a large amount of the information that can be extracted from medical images is not currently used, and we propose here a methodology to take advantage of that information, applicable on routine cancer imaging data.

This emphasizes is large potential impact in personalized medicine and in cancer treatment worldwide, since there are millions of imaging studies, stored in medical centers throughout the world, waiting to be analyzed and mined for relevant information on patient's prognosis. Our radiomic software could also be coupled to automated data retrieval systems, such as the infrastructure proposed by EUROCAT (www.eurocat.info), allowing feature quantification coupled with accession to standardized imaging studies.

Additionally, with the publication of the study presented in Chapter 10, we shared the largest publicly available imaging dataset to the cancer imaging archive (TCIA) of 511 NSCLC patients of which 89 with gene-expression data. This again, will facilitate reproducibility of research and allow the investigation of other methodologies and hypotheses in the same data.

This chapter received attention from the media, including a press-release from the Maastricht University Medical Centre (http://maastrichtumchoofdsite.createsend1.com/t/ViewEmail/t/CADC497D2FEB9C5D) and an internal press-release in the Dana-Farber Cancer Institute in Boston, USA. This study was also featured in the outlook issue of the prestigious scientific magazine Nature, in a special issue on the outlook of lung cancer: http://www.nature.com/nature/journal/v513/n7517_supp/full/513S4a.html.

Acknowledgments

THANK YOU

To begin with, I would like to mention my supervisor Philippe Lambin. Philippe, although your agenda is fully saturated and your work involves managing a large amount of people, it's always possible to find time to discuss burning hot issues or get guidance from you. Since I came to Maastro Clinic as a biomedical engineering student (when you welcomed me with a greeting in Spanish) you showed interest and always tried to motivate me in my research. Despite that the experience after a few years tells me that that motivation is not always realistic, it's however the impulse that beginners need, to believe in their work and in its importance.

From you I learned that effective communication is essential to be successful in research, and that science is not only the pure idea of science that people has, but it's also a business that involves, marketing, networking, etc. In the CAT meeting I always appreciated your innovative vision, a data driven approach of let's do it all, your skill to connect ideas into projects, to identify niches and to have always a further question. Something very important is that newcomers or beginners in the CAT meeting always were treated with respect and attention, even if they had no idea of what they were talking about, which gives you confidence to talk in public and to discuss work with more experienced colleagues. Unfortunately, our relationship at work did not have a personal connection, which is understandable given the characteristics of your work.

Second, I would like to thank the person that welcomed me in Maastro Clinic and with whom I have had the longest collaboration since then, Hugo Aerts. Since I came to Maastro clinic for a master graduation project you have shown confidence, interest and value in working with me. Although I saw you in those early days as a formal supervisor, that relationship then evolved towards a more horizontal and relaxed relationship where I can discuss things with you in a very open and direct manner, perhaps due to your Dutch culture. I also developed a personal connection that involved talking not only about work related topics, but basically about anything, and that's important to develop a sense of confidence and open communication at work.

The Radiomics project, which started with exploring a few simple parameters in a re-reused small dataset, has grown dramatically and has also influenced me to the point of joining your recently started enterprise in Boston. This again, showed me your confidence and trust in working with me. Working with you taught me how to think big in science and I hope this dynamic communication and collaboration will continue indefinitely. Thanks for welcoming me at Maastro and a few years later again in Boston.

To my co-promotors, Frank Hoebers and Andre Dekker. Frank, I would like to thank you for the collaboration and support in the second part of my PhD with the head and neck projects. As a supervisor, I like your discipline and organized manner of following the progress of things, your open communication and the fact that you showed interest also per-

sonally. It has been a pleasure working with you. Your previous experience in other institutes also opened and facilitated the doors for collaboration with our projects in head and neck cancer.

I admire your constancy and love for running. I was surprised that you made a shorter time than me in the 10k Paralleloop in Brunssum, however I didn't know at that time of your shape and skills as a runner. Until you showed up I had an unbeaten record at the Paralleloop (no, Patrick isn't faster). My right knee recovered completely however, so we may have the chance to have another run. Perhaps the Boston marathon? Ok, I'm overreacting.

Andre, even though it was only a few times, I always found it very pleasant to work with you. Perhaps because you express a relaxed way of working, with patience, and always mastering the subjects. Since my first times as audience in the refereer avond, I knew that when it was my turn, I had to be prepared for someone like you. Also I tried to absorb from you ideas for presenting research in an exciting way and for doing research not only in the correct way, but also in a neat way. I hope to continue collaborating with you in the future.

To the assessment committee:

Thank you Prof. Ramaekers, Dr. Kooi, Prof. Kaanders and Dr. Beets for taking part of my assessment committee and for your time and effort reviewing this thesis.

To my colleagues and friends at Maastro Clinic:

To the participants of the CAT meeting (also later re-named to Georgi's meeting): Cary, Dirk, Georgi, Lucas, Hugo, Ralph, Sara, Andre, Erik, Wouter, Ruud, Johan, Ragu, Frank, Steven, Karen, plus a few others. It was a pleasure to participate with you in the fresh and exciting Tuesday morning meetings at 09:45ish. This was basically the place where you grow as a scientist, because ideas are challenged, you are presented with questions, and you may even end up nervously making drawings in the board to "clarify" your ideas. Thank you all for the discussions (and follow up discussions after the meeting), for the quick and almost always empty rondvraag sessions and for your readiness to end up the meeting as soon as possible once your turn to speak had passed (with the exception of Georgi ;)).

Wouter, goede middag. It was my pleasure working with you. I enjoyed how you often chaired the CAT meeting, with a to-the-point-ok-done practical Dutch approach!.. although always with full follow-up notes. I considered you always as the senior scientist of Maastro and as someone that could be reached for objective and valuable advice.

Thanks to the data management team for their support with collecting data. Also, many thanks to the kind receptionists downstairs who would always be happy to help retrieving follow-up data, even if it was in a very short notice and with high urgency, somehow it's

always like that. And also for providing me with an extra pass the many times I forgot mine. I'm sure that eventually you learned my name.

Thank you Erik, Andre, Lucas and Wouter for all the help with IT trouble-shooting and data retrieval related problems. I learned a lot from you and for the things I didn't learn I can always ask you again ;).

Esther Troost, thank you for your help and patience with contouring lung tumors. It was a pleasure working together with you.

Jos de Jong, thank you for your support and for sharing your famous up-to-date-most-complete Jos de Jong database. That database was the engine and source for the head and neck projects I was involved in. It was also a pleasure working with you. Thanks for your sheer interest in the Radiomics part of the projects.

Piet van den Ende, ik vond het altijd leuk om met je te praten in the koffie kamer over de red socks voetbal ploeg, waar ik voor wat tijd heb gespeeld en waar de vrienden van jou zoon ook spellen, over locale evenementen in Maastricht, over het weer, over het prachtige Maastricht. Altijd in het Nederlands, wat ik uitstekend vond anders blijft het moeilijk om te oefenen. We hebben niet veel samen gewerkt maar bedankt voor het maken van een leuke werk plek.

Lars, I enjoyed talking with you about instruments, in a beginner amateur drummer enthusiast to expert direction. I enjoyed a lot the couple of times I got to see your band playing in the Café Forum.

To the Radiomics team: Ralph, Sara and Chintan. Thank you for joining the Radiomics team, which keeps on growing since. It has been a pleasure working with you. Ralph, thanks for greatly improving our slow and primitive code, and for the exciting but also sometimes going nowhere discussions that we had over a bunch of radiomics related issues. These discussions were essential to make the code bullet-proof (almost) and to assure that all algorithms were correctly implemented and that things made sense or at least that's what we convinced each other of. Thanks for your support trouble-shooting radiomics code. It's been great working with you. Too bad that by living far away you missed the other side of things happening in Maastricht.

Thanks to the Maastro Lab guys who helped me in the early days of my PhD, when I was passing through a time in which I thought I was done with computers and wanted to learn cancer biology and lab techniques. Thank you Ludwig, Kasper, Natasja, Sarah, Barry, Maud, for teaching me basic lab techniques, lab discipline and for allowing me to work with you in an improvised project. However, doing cancer biology science requires more than enthusiasm and several weeks of learning. I was hoping I could continue joining the lab days out even though I was no longer there. Do you still have my lab journal by the way?

Thank you to all the co-authors and external collaborators with whom I had the chance to work together. Your input is the core of this thesis. Thank you for all the rounds of critical

remarks and also the non-critical ones, for the revisions, the discussions, the improvements and for your time and effort. Thank you as well for providing data and guidance. Thanks to Ernst-Jan Speel, Jos Straetmans, Bernd Kremer, Kim de Ruyck, Michelle Rietbergen, Ruud Brakenhoff, C. Rene Leemans, Derek Rietveld, Johan Bussink, Vincenzo Valentini, Robert Gillies, Yuhua Gu, Rene Korn, amongst others.

To the heavy-header friends (you know who you are): I wonder what forces and unrelated succession of events came together to put in the same working place such bunch of talented, cultured and likeminded entrepreneurs. Looking back you realize that that connection and combination of personalities was unique, that great things came out of it and that it now looks like those are the never happening again (in that way) golden years. Thanks guys for the many beer, Belgian beer, mezcal, jagger, whiskey, tequila, beer to cool down again, evenings. I owe you a round.

Rudi, it went from, ooh no, again this boring nerdy in the train to verrekte #$!, or Ruudi Ik vind jou leuk. Rudi, thanks for everything. I discussed and got help from you about everything: nomograms code, trees, stats, languages, trees, Dutch traditions, photoshop, trees, music, trees. Of course many of those discussions and classified information would leak out through your fully maintained wikileaks channel. The most fun time at Maastro started when the upper management decided to put us together in the same desk and in the best section (the cool dudes' one) in the office. You were always happy to help me with heavy duty work, with a smile and without complains. I miss it, we should have continued working together. Shall I spend a few months in Oxford?

Guillaume, our relationship also went from a friendly-tense-competing this new guy pisses me off, to I love you man. Probably the richest part of our friendship was centered in the philosophical discussions about many aspects of life (although 97% of the times it was the same subject), the mutual enjoyment of things and the fact that we found every small part of Maastricht a beautiful source of inspiration. It has been a fantastic time in Maastricht. I can't believe I almost cried when you were leaving to Canada for Christmas holidays but not when I left to Boston and said bye to my girlfriend.

Paddy, dear Paddy. I don't know how to start. When you came to Maastricht I thought I should help this lost, distracted, wandering around, older looking Canadian, and before I knew it, I thought, wait a minute, could it be, did I just add one more to my one-wolf wolf-pack? We have been through so many things together you wouldn't even know, I can't list all that here because of space and because I can't mention them, I really can't, but it was great. I love you man. Ah, thanks for being my paranymph too.

Mark, Marky-Marky. Thanks buddy for adding a complete different layer to the group, the one of the wise, piss-off, Irish flavor. I'm sorry that I didn't understand what you said almost all the time, even though you always had a sharp comment. Thanks for the very insightful and inspiring trip in your home city. It's been a pleasure and I hope we do it again. It wasn't so nice that you never made again that cherry port risotto.

Dr. Sean Walsh. After moving to Boston and starting a new job in a new city, I often remind myself of you, of your positive, cheerful and very open attitude. It was great to see how you quickly became part of an otherwise very closed group. Your input as experienced and only post-doc quickly became evident in the group: we drank more, partied more, some smoked more, everything more. I enjoyed a lot the +10k running sessions in the winter, without hoodie and the MAC sessions. It's been a pleasure hombre hanging around with you.

New-Mathieu and Gab. Although we co-existed at Maastro for a short period of time, it was great meeting you guys and hanging out. Thanks for the very fun trip to Warsaw, Gab did you get married at the end? NewMat, thanks for the countless excursions to the center of Maastricht at night, to see what was happening in there. See you soon in North America.

Also thanks to all the office colleagues for sharing and maintaining a pleasant work environment: Ralph, Sara, Hoda, Skadi (we should have lunch together once Skadi), Celine, Karen, Scott, Adriana, Shane, Stefan, Georgi, Davide, Francesco, students, temps.

To my Maastricht friends:

Pablo, gracias por el par de años de excelente compañerismo y amistad en Maastricht. Siempre vi nuestra causa como la de los primeros exploradores de cualquier lugar antes desconocido, decididos a conquistar el lugar y absorber su, su cultura. Creo que debimos haber extendido un par de años más el doctorado. Igual que con Patrick, gracias por las innumerables sesiones de trabajo pesado en Maastricht, las pausas con café muertos de risa en la universidad y las muchas veces que casi le prendimos fuego a la ciudad, literalmente. No por nada al final te apodamos "the beast".

Lo mejor de todo fue la facilidad con la que intercambiamos amigos mutuos. Gracias por introducirme a otro par de personajes y grandes amigos: Leo y Fabien.

Leo and Fabien, thank you for the great times and drinks in Maastricht and through different European cities. It was a fantastic time, I hope we repeat one of those trips again soon. Wankers, youff.

Luis, gracias también por la excelente amistad en Maastricht. Sé que al principio pensabas, ah inche par de güeyes sangrones, pero al final se convirtió en una gran amistad. Siempre disfruté mucho y extraño los partidos de futbol, creo que es de lo que más extraño de Maastricht, las tardes de Falstaff, seguidas por un par de mezcales en Take Five, con discusiones alargadas de futbol, de libros, de música, de gente, de lugares. Gracias también por las excelentes cenas y los almuerzos norteños tan reparadores de la resaca. Me gustaría que mi familia sea como la tuya en el futuro.

Bart van den Bogaard, vriend, I also don't know where to begin. It was probably an immediate click when I met you in the party in Eindhoven where we improvised the mister blue, mister orange and so on story, like many other stories. Your easy going and open personality never ceased to amaze me. I just truly enjoyed every single activity in which we engaged together. You were a challenging friend because you always pushed the limits

however pushing the limits was what made things happen. I just feel that we didn't expand our recruiting Bombeiros company enough and that we should continue doing so. Your friendship was also a door to the Dutch culture, many of whose habits now I make of my own. Also, thanks to you I got to meet another great set of Dutch wankers: Roel, Jochum and Emiel, and the lieutenant Kyle S. Herman. Thank you guys for all the trips, evenings through different cities in the Netherlands and for a fantastic surprise farewell. We still need to finish that evening. De uilen: wooohooo!. Ah, Bart thanks for accepting and taking very seriously the position of paranymph.

Arthur and Clara, two of my early and best friends in Boston. To you I owe the last push to finish writing this thesis. The evening that we had pasta, wine and polish vodka and that I concentrated in finalizing the work meanwhile you constantly interrupted me with a new offer to drink. Cheers, good luck with things back in Europe and hope to see you soon.

A mi familia, madre, papa, hermanos. Gracias por el apoyo constante desde los 14 años cuando me corrieron de la casa. Aunque para muchos no es fácil estar lejos de su casa, la libertad y la confianza que ustedes me dieron ayudo a que haya podido estar tanto tiempo lejos. Aunque los extraño, les agradezco infinitamente que me hayan empujado a ver otros lugares, a viajar, a vivir otras cosas. El viaje del 2005 a Europa, del que me dieron una regañiza por mi conducta inapropiada y libertina, me abrió los ojos y fue el impulso que me hizo ir a Holanda dos años después.

Monsters inc, querida Saskia. To you I really cannot thank enough. Thank you for gracefully performing the positions of Emmanuel's girlfriend, friend, moderator, assistant, stylist, thesis acknowledgments censurer, adventures friend, cooking teacher, advisor, bad jokes/comments moderator, replacement of my male friends when they aren't there, educator, etc, etc. I'm not sure if you contributed to this thesis, maybe yes, with making me work in the weekends or when I didn't feel like, however you definitely have greatly contributed to my life, to my experience in the Netherlands and I'm sure also, to the love and admiration that I feel for your country.

Thanks for pushing me to move to Boston and for joining me here, it's the beginning of a very long and very exciting trip together. Te llevo para que me lleves.

Soon you will have to take care of two children!

And to the Netherlands, thank you for welcoming me and integrating me into your country, for so much fun, experiences, people and everything else in the last few years. I miss you. Het was echt erg gezellig!

Curriculum Vitae

CURRICULM VITAE

Emmanuel Rios Velazquez was born on September 17th 1985 in Coroneo, Guanajuato, Mexico. At the age 14, he moved to Mexico City where he studied high school, at the National Polytechnic Institute, from which he obtained the title of Chemistry Laboratory Technician in 2003. He studied Biomedical Engineering in the same institution, particularly interested in the development of computer controlled electronic medical devices. In 2007, after obtaining the Bachelor's degree in Biomedical Engineering, he obtained the Royal Dutch Shell Scholarship, to follow a Master in Biomedical Engineering at the Eindhoven University of Technology (TU/e). His Master degree was primarily focused in the computational modelling of biological processes and medical imaging.

Part of his program included an internship at the Biomodeling and Bioinformatics group of the TU/e, where he studied the dynamics of the hypothalamic-pituitary-adrenal system and its relation with metabolic syndrome. Also, he performed an externship at the Biophysics Lab of the Maastricht University and the SoftMatter CryoTEM Research Unit of the TU/e, working on the quantitative analysis of microscopy images of angiogenesis in mice models.

His master thesis project brought him to the radiation oncology department (MAASTRO) of the Maastricht University, where he worked on the pioneering steps on the use of imaging parameters to predict outcome in lung cancer patients, in a joint project between the Systems Biology group of the TU/e and the Maastro Clinic.

In 2009, he obtained the master degree in Biomedical Engineering, and continued his master graduation work and his scientific education at the Maastricht University, under the supervision of Prof. Dr. Philippe Lambin. Then the initial master graduation work was extended over the years to his PhD dissertation work and led to multi-centric nationwide and international scientific collaborations.

He currently works as postdoctoral researcher at the Computational Imaging and Bioinformatics Laboratory at the Dana Farber Cancer Institute and Harvard Medical School, in Boston, USA.

Grants and awards

- KWF travel grant for research visit to Dana-Farber Cancer Institute: "Radiogenomics in head and neck cancer: linking molecular profiles with noninvasive imaging".
- Best clinical poster award. European Society for Therapeutic Radiology and Oncology. 29th Conference
- Royal Dutch Shell Scholarship. Eindhoven University of Technology. Scholarship awarded based on academic and social commitment.
- Institutional Scholarship "Alfredo Harp Helu". Scholarship awarded based on academic excellence.

List of publications

SCIENTIFIC PUBLICATIONS

1. Hugo J.W.L. Aerts*, **E. Rios Velazquez***, R. Leijenaar, C. Parmar, B. Haibe-Kains, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. 5:4006 (2014), IF = 10.47

2. **E. Rios Velazquez**, F. Hoebers, et al. Externally validated HPV-based prognostic nomogram for oropharyngeal carcinoma patients yields more accurate predictions than TNM staging. Accepted for publication in Radiotherapy and Oncology.

3. Kim De Ruyck, Fréderic Duprez, Liesbeth Ferdinande, **Emmanuel Rios Velazquez**, et al. A let-7 microRNA polymorphism in the KRAS 3'-UTR is prognostic in oropharyngeal cancer. Cancer Epidemiology, 2014. Epub ahead of print.

4. Chintan Parmar*, **Emmanuel Rios Velazquez***, Ralph Leijenaar, Mohammed Jermoumi, Sara Carvalho, et al. Robust Radiomics Feature Quantification using Semiautomatic Volumetric Segmentation. PLoS ONE 9(7): 2014. IF = 3.1

5. Olya Grove, Anders E. Berglund, Matthew B. Schabath, Andre Dekker, **Emmanuel Rios Velazquez**, et al. Computed tomographic imaging biomarkers and prognosis of lung adenocarcinomas. Under review

6. P.Lambin, R. van Stiphout, M. Starmans, **E. Rios Velazquez**, A. Dekker, et al. Predicting outcomes in radiation oncology —multifactorial decision support systems. Nat Rev Clin Oncol. 2013 Jan;10(1):27-40, IF = 15.03

7. Gu Y, Kumar V, Hall LO, Goldgof DB, Li CY, Korn R, Bendtsen C, **Velazquez ER**, et al. Automated delineation of lung tumors from CT images using a single click ensemble segmentation approach. Pattern Recognit. 2013 Mar 1;46(3):692-702, IF = 3.21

8. **Velazquez ER*,** Parmar C*, Jermoumi M, Mak RH., et al. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. Sci. Rep. 3, 3529 (2013), IF = 5.07

9. Lambin P, Roelofs E, Reymen B, **Velazquez ER**, Buijsen J, et al. 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy. Radiother Oncol. 2013 Oct;109(1):159-64, IF = 4.52

10. Carvalho S, Leijenaar RT, **Velazquez ER**, Oberije C, Parmar C, et al. Prognostic value of metabolic metrics extracted from baseline positron emission tomography images in non-small cell lung cancer. Acta Oncol. 2013 Oct;52(7):1398-404. IF = 2.86

11. Leijenaar RT, Carvalho S, **Velazquez ER**, van Elmpt WJ, Parmar C, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013 Oct;52(7):1391-7. IF = 2.86

12. Hoebers F*, **Velazquez ER***, Troost E, van den Ende P, Kross K, Lacko M., et al. Definitive radiation therapy for treatment of laryngeal carcinoma: Impact of local relapse on outcome and implications for treatment strategies. Strahlenther Onkol 2013 Oct;189(10):834-41. IF = 4.16

13. **E. Rios Velazquez**, H. J. W. L. Aerts, Y. Gu, et al. An automatic CT-based ensemble segmentation of lung tumors: comparison with oncologists' delineations and validation with surgical specimen. Radiother Oncol. 2012, 105:2, 167-173, IF = 4.52

14. P. Lambin, **E. Rios Velazquez**, R. Leijenaar, S. Carvalho, R. G.P.M. van Stiphout, P. Granton, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012 Mar; 48(4):441-6. 2012, IF = 5.25

15. Egelmeer A***, Velazquez ER***, de Jong JM, Oberije C, Geussens Y, Nuyts S, et al. Development and validation of a nomogram for prediction of survival and local control in laryngeal carcinoma patients treated with radiotherapy alone: A cohort study based on 994 patients. Radiother Oncol. 2011 Jul;100(1):108-15, IF = 4.52

16. **Velazquez ER**, Aerts HJ, Oberije C, De Ruysscher D, Lambin P. Prediction of residual metabolic activity after treatment in NSCLC patients. Acta Oncologica. 2010, 49:7, 1033-1039.2010, IF = 2.86