

Hand-crafted and deep radiomics for the management of advanced cancer stages

Citation for published version (APA):

Keek, S. A. (2023). *Hand-crafted and deep radiomics for the management of advanced cancer stages*. [Doctoral Thesis, Maastricht University]. Maastricht University. <https://doi.org/10.26481/dis.20231012sk>

Document status and date:

Published: 01/01/2023

DOI:

[10.26481/dis.20231012sk](https://doi.org/10.26481/dis.20231012sk)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Hand-Crafted and Deep Radiomics For the Management of Advanced Cancer Stages



Simon Keek

Hand-crafted and deep radiomics for the management of advanced cancer stages

Simon Andreas Keek

ISBN 978-94-6469-571-7

Cover design by Kate Lediakhova | www.linkedin.com/in/kate-lediakhova

Design by ProefschriftMaken

Printed by Proefschriftmaken

copyright © Simon Keek 2023

Hand-crafted and deep radiomics for the management of advanced cancer stages

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht,
op gezag van de Rector Magnificus, Prof. Dr. Pamela Habibović
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op donderdag 12 oktober 2023 om 16.00 uur

door

Simon Andreas Keek

Promotor

Prof. Dr. P. Lambin

Copromotores

Dr. H.C.A. Woodruff

Dr. L.E.L. Hendriks

Beoordelingscommissie

Prof. dr. V.B.W. Schrauwen-Hinderling (voorzitter)

Prof. Dr. B. van Ginneken (Radboud Universitair Medisch centrum, Nederland)

Dr. M.M.H. Hochstenbag

Prof. dr. L. Holloway (UNSW Sydney, Australia)

This thesis was partially funded by the Dutch technology Foundation STW (grant n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs

This thesis was also partially funded by a Lung Foundation grant, n° 11.1.18.250.

Content page

Chapter 1.	Introduction and outline of thesis	9
Chapter 2.	A review on radiomics and the future of theragnostics for patient selection in precision medicine <i>The British journal of radiology, 2018</i>	47
Chapter 3.	A prospectively validated prognostic model for patients with locally advanced squamous cell carcinoma of the head and neck based on radiomics of computed tomography images <i>Cancers, 2021</i>	67
Chapter 4.	Computed tomography-derived radiomic signature of head and neck squamous cell carcinoma (peri)tumoral tissue for the prediction of locoregional recurrence and distant metastasis after concurrent chemo-radiotherapy <i>PLoS One, 2020</i>	117
Chapter 5.	Investigation of the added value of CT-based radiomics in predicting the development of brain metastases in patients with radically treated stage III NSCLC <i>Therapeutic advances in medical oncology, 2022</i>	153
Chapter 6.	Predicting adverse radiation effects in brain tumors after stereotactic radiotherapy with deep learning and radiomics <i>Frontiers in Oncology, 2022</i>	191
Chapter 7.	Towards texture accurate slice interpolation of medical images using PixelMiner <i>Computers in Biology and Medicine, 2023</i>	225
Chapter 8.	Discussion and future perspectives	249
Chapter 9.	Impact statement	265
Chapter 10.	Summary	271
Chapter 11.	Samenvatting	277
Chapter 12.	Acknowledgements	283
Chapter 13.	Curriculum Vitae	289
Chapter 14.	List of publications	295

CHAPTER 1

1

Introduction and outline of thesis



1.1 Cancer

The worldwide cancer incidence and mortality estimates in 2012 were 14.1 and 8.2 million, respectively [1]. In 2020, the incidence and mortality were 19.3 and 10 million, respectively, which are increases of approximately 37% and 22% in 8 years, indicating that the burden of cancer on healthcare is increasing worldwide [2].

As most head and neck squamous cell carcinoma (HNSCC) and non-small cell lung cancer (NSCLC) tumors show no early symptoms, they are commonly diagnosed at an advanced stage [3-6], which drastically lowers the chance of survival. Similarly, symptomatic brain metastases (BM) are associated with a decreased quality of life (QoL) and poor prognosis (regardless of the primary tumor) [7], and usually have no curative treatment options available. Therefore, we chose to focus on these types of cancer in this thesis. HNSCC has a large variety of subtypes, induced by differences in genetic profiles, and could benefit from a personalized medicine approach [8]. NSCLC, especially advanced disease, is associated with a high risk of BM [9, 10]. However, it is currently not possible to reliably predict which patients with NSCLC will develop BM, which could help in clinical decision-making. For the patients that do develop BM, stereotactic radiotherapy (SRT) is an increasingly used treatment modality [11, 12], but it is currently likewise not possible to reliably predict which patients may develop radiation necrosis (RN), a common side effect of this treatment. For all three types of cancers, artificial intelligence (AI) could be used to improve prognosis, using information available in clinical imaging, e.g., scans already performed for staging procedures, that is currently not used in the clinic. Identifying patients with HNSCC with high or low chance of survival might aid in clinical decision-making, or could be used to stratify patients in clinical trials according to prognosis. For NSCLC, predicting the risk of BM would allow to identify patients who could benefit from treatment such as prophylactic cranial irradiation (PCI). For BM, identifying patients with high risk of RN could allow the clinician to opt for systemic therapy instead, if systemic therapy is a reasonable treatment option. This thesis will therefore focus specifically on AI applications in advanced stages of these types of cancer. To further demonstrate that there is a need for these applications, the paragraphs below will first give some background on current staging procedures and (selection for) treatment options.

1.2 Clinical diagnosis and staging

Diagnosis and management of cancer depends on the cancer type. In general, a tissue sample is first obtained to determine pathological classification [13]. The type of procedure to obtain this tissue sample depends on the type of cancer, the location in the body, and the health status of the patient.

A clinical staging workup is performed to determine the extent of the disease. This is mainly performed through the tumor-nodes-metastasis (TNM) staging system. This system describes the primary tumor location, size, and invasiveness (T-status), the number of local lymph nodes that the cancer has spread to (N-status), and finally the presence and extent of cancer metastases (M-status). This combination of T-, N-, and M-status results in an overall cancer staging, ranging from I to IVA/B. A TNM stage can be clinical (c-stage, based on imaging) and pathological (p-stage, after surgery). Stage I tumors are generally small and have not spread yet to other parts of the body, while stage IV tumors have metastasized. It follows that patients with higher stages generally have poorer prognosis, and may not be eligible to the same treatment as lower stages.

In addition to size and location, the genetic and mutational status of the tumor are commonly used for deciding treatment, and sometimes also for staging. For oropharyngeal HNSCC, human papillomavirus (HPV) status is included in American Joint committee of cancer (AJCC) 8th edition staging [14]. For NSCLC, tumor programmed death ligand1 (PD-L1) status, as well as the presence of several oncogenic drivers such as anaplastic lymphoma kinase (ALK) fusions, B-Raf V600 mutations, epidermal growth factor receptor (EGFR) mutations and ROS1 rearrangements influence the choice for treatment in metastatic disease and increasingly also in earlier disease stages [15-18]. Furthermore, information regarding patient characteristics and preferences is also taken into consideration by determining patient performance score, comorbidities, and the expected QoL for and risk of complications for the treatment choices available. Lastly, patient specific clinical and biological factors are regarded, such as smoking and alcohol consumption status, hemoglobin level, sex, and age [19-24].

1.3 Treatment

Treatment is dependent on the type of cancer and the stage of cancer. For smaller, easily resectable tumors, patients are usually treated with surgery. For patients with higher but potentially still curable cancer stages, often a combination of radiotherapy (RT) or surgery with chemotherapy (CTx) is performed, either concurrently or sequentially. Other treatments are immunotherapy or targeted therapy, depending on specific mutational and genetic markers.

RT is a treatment wherein ionizing radiation is delivered to the tumor, which destroys cells and aims to shrink or remove the tumor entirely. Because of their fast cell division and growth, tumor cells are extra susceptible to radiation that can damage these processes. The extent to which a tumor reacts to radiation is dependent on the cancer type, as different cancers have shown to have different radio sensitivities [25]. Furthermore, the

size of the tumor is important, as well as the presence of hypoxia in the core of the tumor, which can negatively affect the effectiveness of RT [26, 27].

The ionizing radiation has to travel through the body to reach the intended target, and will pass through the rest of the body after having reached the target. This causes healthy tissues before and behind the target to also be affected by a (similar) dose of ionizing radiation, which will damage healthy tissues and may lead to toxicity [28]. To minimize the dose to healthy tissues and maximize the target dose, the radiation is delivered to the tumor at multiple angles of exposure intersecting in the region of interest (ROI). This causes the dose to be focused on the ROI, and spares the rest of the body. The delivery of the dose to the ROI needs to be very precise, as missing the target means a large dose is delivered to healthy tissue instead. If the intended target is close to organs at risk (OAR), this may cause severe and long-lasting (if not permanent) side effects to the patient [29-31]. Therefore, monitoring of the tumor location, size, and shape throughout the entirety of the treatment is very important.

1.4 Specific cancers

An important part of clinical decision-making is the estimation how a patient will respond to a certain treatment. Examples include an estimation of the survival chance, the chance of recurrence, or the risk of toxicity. Despite improvements in the identification of prognostic and predictive factors as well as in treatment itself, many challenges remain as discussed below for stage III-IVB (locally advanced and advanced) HNSCC, stage III (locally advanced) NSCLC, and BM (regardless of the primary tumor).

1.4.1 Head and neck cancer

HNSCC is a type of cancer that develops in the cells of the mucosal tissues of the upper digestive and respiratory tract, which includes laryngeal, hypopharyngeal, oropharyngeal, nasopharyngeal, and oral cavity cancer. HNSCC is the sixth most prevalent cancer worldwide [32, 33]. Over the last decade, a striking increase in oropharyngeal HNSCC incidences has been observed in the Western world, mostly affecting generally young and healthy patients. This is due to HPV infection, which is a risk factor for oropharyngeal HNSCC (3). One important note is that while HPV infection increases the risk of (oropharyngeal) HNSCC, patients with HPV-positive cancer have in general a better treatment response than those that are HPV-negative. Other risk factors common to HNSCC in general include (excessive) alcohol consumption and smoking. Figure 1 shows an example of a stage IVA oropharynx HNSCC tumor.

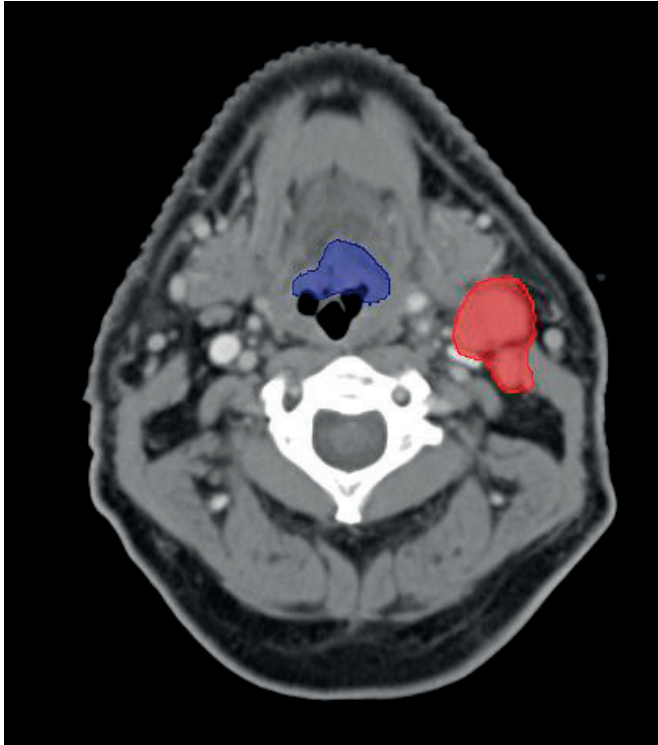


Figure 1. CT of Stage IVA oropharynx HNSCC patient, with the primary tumor outlined in blue and a lymph node tumor outlined in red.

HNSCC is generally associated with a poor prognosis, as most are found at advanced stages (stage III-IVB). The 5-year median survival chance of patients with advanced HNSCC is 25-60% (4). Neoadjuvant immunotherapy has shown promising results in studies investigating improvement of survival outcome for advanced HNSCC (8-10). However, as immunotherapy seems eligible for only 20% of investigated patients (5), barely any increase in 5-year survival has been observed. Lastly, HNSCC is one of the most psychologically damaging types of cancers, primarily because of late effects of the cancer and its treatment, including disfigurement, swallowing difficulties, and pain [34-36].

Diagnosis according to the AJCC tumor staging of HNSCC starts with determining the primary tumor location, determination of the size of the primary tumor, the extent of invasiveness in surrounding tissues, and the presence of secondary tumor masses (7). This is performed through clinical examination in the form of fiberoptic endoscopy and by taking biopsies, and through imaging. Imaging involves computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and ultrasound (US). CT and MRI are used to determine tumor size and extent of invasiveness, and the metastasis to regional lymph nodes, while PET is used to determine if metastasis occurred

to distant locations. Lastly, US is used to assess the lymph nodes through US-guided fine-needle-aspiration, to perform pathological confirmation (6).

Treatment of HNSC is currently either RT, CTx, surgery, or any combination of these three strategies. CTx for advanced HNSCC is mainly cisplatin-based, but cetuximab and carboplatin are often used as well (6). However, HNSCC has a large variety in response to treatment, induced by the different tumor locations and genetic profiles of the cancer. This means patients could benefit greatly from personalized treatment paths, which would require accurate prognostic tools. These tools could be used to stratify patients into risk-groups. For advanced HNSCC, having three identifiable risk-groups (a low, medium, and high-risk group) would allow for selection of those patients who could benefit from more intense treatment, or identify patients who could be spared from treatment.

1.4.2 Non-small cell lung cancer

Lung cancer describes a number of histological subtypes of cancer originating in the lungs. These subtypes are commonly divided into two groups: the histological subtyping small cell lung cancer (SCLC), and all subtypes that are NSCLC [37]. NSCLC comprises over 80% of all lung cancers [38]. The main NSCLC subtypes are adenocarcinoma and squamous cell carcinoma [6]. Adenocarcinoma describes cancers that form in glandular tissues of the body, while squamous cell carcinomas describe cancers that originate from squamous cells that form the surface of the skin, and line the respiratory and digestive tracts. Over 50% of patients with NSCLC present with metastatic disease (stage IV) and are considered incurable, and around half of the patients without metastasis upfront (stage I-III), progress to metastatic disease despite curative intent therapy [3]. The long-term survival of patients with metastatic NSCLC is in general very poor, with a 5-year survival rate generally below 10%, although targeted therapies and immune therapies are changing this paradigm with 5-year survival rates reaching 30-60% [39-41].

Recommended staging procedures for NSCLC depend on the suspected stage, but generally include a fluorodeoxyglucose ^{18}F (^{18}F -FDG) PET and CT chest and upper abdomen, often brain imaging (preferably MRI, otherwise CT) and if necessary mediastinal staging with EUS/EBUS or even mediastinoscopy. Approximately 21% of patients are diagnosed with stage III (locally advanced) NSCLC [42]. Standard treatment for patients with unresectable stage III NSCLC is concurrent chemoradiation (CCRT) [43]. If a patient is incapable of receiving this treatment due to comorbidities or suboptimal performance status, or if the tumor volume (and as a result the radiation field) is too large, sequential CRT or only radical RT are an option. As of 2018, adjuvant immunotherapy (durvalumab) has become the standard of care for patients treated with CCRT, without disease progression after CCRT [44]. Unfortunately, despite radical intent therapy, up to 30% of the patients with stage III NSCLC will develop BM during the course of their disease [45]. PCI reduces incidence

of BM with a relative risk of 0.33, but without improving OS in NSCLC and at a cost of neurotoxicity [45]. Durvalumab reduces this incidence with approximately 50%, but still BM remains a major problem in NSCLC as the overwhelming majority of these patients cannot be treated with curative intent anymore, while QoL decreases with symptomatic BM [7]. Being able to determine which patients are at high risk of BM may help in clinical decision-making, offering patients with high BM risk PCI. However, conventional risk assessment of BM is currently lacking, and more advanced and accurate models are required. Figure 2 shows an example of a stage III NSCLC tumor (primary tumor, involved lymph nodes not depicted).

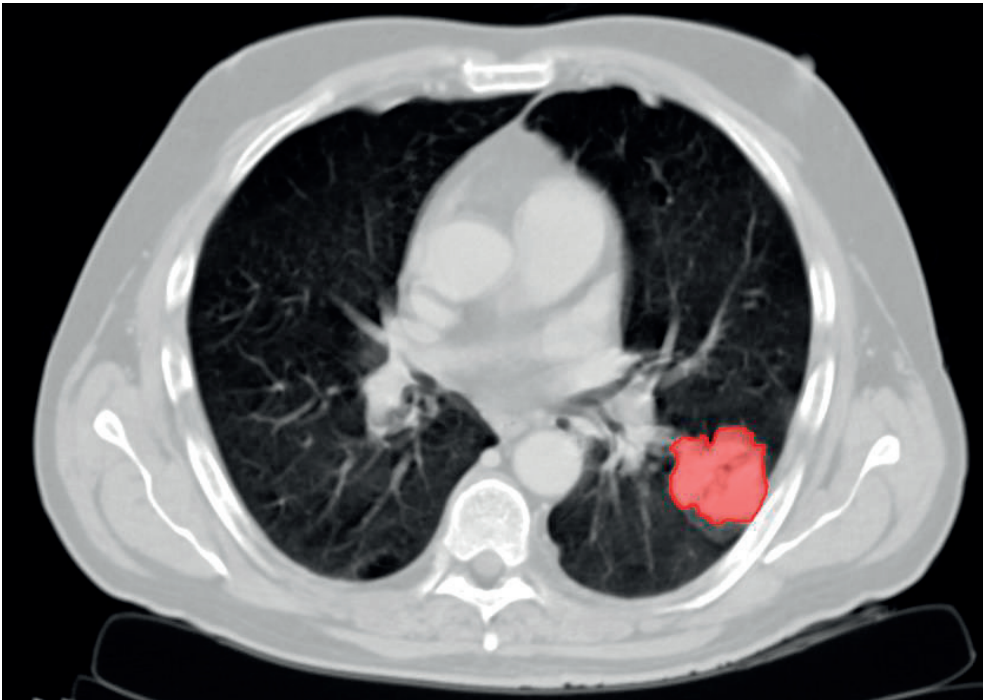


Figure 2. CECT of Stage III NSCLC patient, with the primary tumor outlined in red (N2 lymph nodes not shown).

1.4.3 Brain metastasis

BM means that cancer has spread from a different part of the body to the brain. BM are more common than primary brain tumors, with reported incidences of 8.3-11 out of 100.000 for brain metastases compared to 6.6 out of 100.000 for primary brain tumors [46-48]. The most common primary tumor locations are lung cancer, breast cancer, and melanoma, which have a cumulative risk of 20-40% to develop BM and which add up to 67-80% of all BM cases [49-51].

In general, patients with BM have a poor prognosis and little curative options. The prognosis and optimal treatment of BM are dependent on the primary location, and the median survival times reported generally fall within an 8-16 month range [52]. The diagnosis specific graded prognostic assessment (ds-GPA) is one of the tools to estimate expected survival time for patients with BM, dependent on patient age, primary tumor, Karnofsky performance score (KPS), the number of metastases, and mutational statuses [53]. For lung cancer specifically, an updated tool that includes molecular markers (Lung-molGPA) exists [54]. Treatment options are local therapy (RT or surgery, or a combination), or systemic therapy. The treatment decision is based on symptoms, the number and volume of BM, the presence of extracranial metastases, KPS or similar performance status, the options available for local and systemic therapy, and the patient's preferences. RT has two different treatments available: whole brain radiotherapy (WBRT) and stereotactic RT (SRT) [11]. WBRT delivers radiation to the entire brain over a relatively long time, to radiate the present tumors and any invisible lesions that may be present. However, WBRT has shown to have low success in treating BM effectively, and it can induce (severe) neurological side effects. SRT instead delivers high doses of radiation with high precision to the BM in one or several fractions. SRT has a higher success than WBRT in treating a limited number of BM [55, 56], and does not expose the rest of the brain to radiation which prevents neurological degradation. However, SRT carries a risk of complications in the form of adverse radiation effects (ARE) [57]. One example of an ARE is RN, which is a late side effect that occurs when the high radiation dose delivered during SRT inadvertently has been delivered to nearby healthy tissue, which causes reversible or irreversible necrotic scarring of this tissue [58]. RN can be asymptomatic, in which case no treatment is necessary, or present with severe neurological side-effects requiring treatment through steroids, vascular endothelial growth factor (VEGF) inhibitors, or local interventions (laser interstitial thermal therapy or surgery) [59]. It would therefore be ideal to be able to determine which patients are at higher risk of RN to be able to counsel them, as it may lead to the decision to opt for systemic therapy instead if this is a reasonable treatment alternative (e.g. in patients with NSCLC that are eligible for targeted therapy). Figure 3 shows an example of a BM.

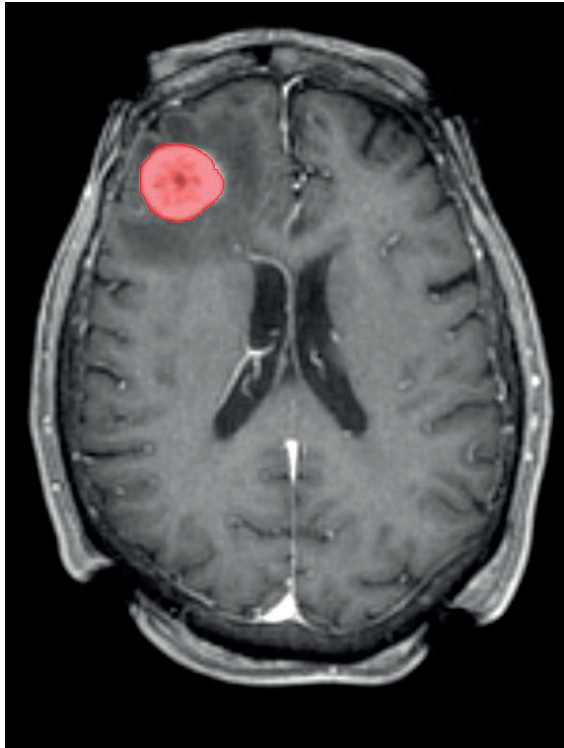


Figure 3. T1 gadolinium enhanced MRI of a brain with a BM, with the BM outlined in red.

Purpose of this thesis

To summarize, advanced stages of NSCLC and HNSCC as well as BM pose several problems in the patient journey: during the diagnosis, the effectiveness of treatment, the prediction of disease relapse, and late toxicity after treatment. For HNSCC stratifying patients in survival risk-groups, even with the implementation of 8th edition staging, remains difficult for stage III-IVB patients. For NSCLC, even though effective treatments exist to lower the risk of BM, due to the possible side-effects of these treatments determining which patients are at a high risk of BM is needed, which is currently not possible clinically. Last, for BM determining which patients are at risk of severe RN before delivering SRT is important, as this information may be used for risk stratification, informing the patient, or even opting for different treatment.

What these patients have in common is the use of medical imaging, either for diagnosis, staging, or for treatment planning purpose. Common medical imaging modalities include PET, CT, and MRI. The imaging modality used depends mainly on the location of the tumor and the aim of the procedure (finding distant metastases, providing local details etc.). Currently, these images are generally not used other than for the aforementioned

purposes. Quantitative image analysis of these medical images may allow for the identification of phenotypical subtypes of tumors that could be investigated for their correlation to certain clinical outcomes.

1.5 Medical Imaging

Medical imaging is the capturing of imaging data of patients, both exterior and interior, for diagnostic, prognostic, and intervention purposes. Medical images can be either two-dimensional (2D) or three-dimensional (3D), depending on the imaging modality. A 2D image extends in the x-direction and the y-direction, expressed through number of pixels, while the z-direction is generally expressed by the number of 2D slices that combine to form a 3D image. Higher numbers of pixels or slices, or to be more precise the number of pixels or slices per unit length, mean more information is captured at a higher resolution. What follows is a brief description of the imaging modalities mentioned or analyzed in this thesis: PET, CT, and MRI. PET

PET is performed through tomography, which is an imaging technique where waves penetrate the imaging target, and images are then produced through reconstruction of the degree of absorption of these waves. PET is an imaging technique which is able to relay information regarding the patient's metabolism and as such requires a patient to be injected with a tracer, which is a radioactive substance that gets absorbed in different gradations by different parts of the body. A commonly used tracer for cancer is ^{18}F -FDG, which is a sugar with an attached positron-emitting radionuclide that gets absorbed by tissues during metabolism. As cancer cells have a very fast metabolism, relatively more FDG gets absorbed by these cells. The radionuclide emits photons, which can be detected by detectors located around the patient in the PET-scanner. Body parts with higher uptake of the tracer will show with more intensity, which is expressed through the semi quantitative standardized uptake value (SUV). This is the ratio of the local measured radioactivity and the body concentration of the injected radioactivity, which is displayed on a ^{18}F -FDG-PET scan. Common downsides of PET are low resolution of the images compared to techniques such as CT and MRI, and variability in tracer uptake between patients and due to differences in protocols.

1.5.1 CT

Similarly to PET, CT is a tomography imaging technique. Specifically, CT is an x-ray imaging procedure that creates 2D "slices" of the patient, which are combined to form a 3D image. The slices are created by sending a narrow beam of x-rays into the patient, which absorbed at different rates in the body depending on the tissues it encounters and are measured by detectors opposite to the transmission source. By combining the attenuation of the

x-rays over multiple angles, a 2D image of the patient can be reconstructed. The calibrated attenuation coefficient, or radio density, of each tissue is expressed in Hounsfield units (HU), with -1000 HU representing air and 0 HU representing water. Pixel values in CTs are therefore quantitative measures that correspond to the densities of the displayed tissues.

CT scans are usually used for imaging regions of the body with high contrast, for example within the lungs, or when bone structures need to be displayed. To better display soft tissues, intravenous contrast agents are injected, creating contrast-enhanced CT (CECT) scans. These contrast agents absorb x-rays, and as they are absorbed at different rates by tissues can increase contrast compared to non-CECT scans. Most CT modalities included in this thesis are CECT images. Furthermore, a combination of PET and CT (PET/CT) can be used, combining the information on the tissues through CT and of the metabolic activity of these tissues through PET.

1.5.2 MRI

MRI is an imaging procedure that uses the energy of shifting magnetic fields to produce images. MRI employs a strong constant magnetic field over the entire body, which aligns all the protons in the body in the same direction. A radiofrequency (RF) pulse is then applied, which causes the protons to spin out of equilibrium. When this new field is turned off, the atoms return from an excited stage to a relaxed state in the original magnetic field, which releases energy measured by the MRI scanner. Depending on the density of the hydrogen atoms, tissues will release less or more energy, resulting in a different intensity value for different tissues on the images.

A particular setting of the RF pulse or the magnetic field gradients is called a sequence, which all have their unique image types. The main types of sequences used clinically are T1-weighted, and T2-weighted. The main difference between these sequences is that T2-weighted images display a higher intensity for water compared to T1-weighted images. The intensity values obtained through MRI are not calibrated and hence cannot be connected to meaningful physical or chemical characteristics of tissues - they are therefore not strictly quantitative.

1.5.3 Clinical role of medical imaging

Medical imaging has become a fundamental aspect of routine clinical practice, and is used for a multitude of purposes. For RT treatment-planning, images are needed to accurately deliver radiation to the ROI. For many diseases, such as cancer, ischemic stroke, Alzheimer's disease, lung emphysema, radiological images are used to determine the presence and extent of the disease. This is done traditionally by expert radiologists, oncologists, and other medical experts by visually grading the disease on an image. Sometimes, semi-quantitative measures such as Response Evaluation Criteria in Solid Tumors (RECIST) or

TNM staging is performed. These measures are performed to stratify patients into risk-groups and to measure the efficacy of treatment, allowing clinicians to make decisions for treatment, and to assess survival or chance of recurrence of disease.

1.6 Personalizing medicine via medical imaging

The current clinical practice is informed by clinical trials, which are performed on fairly homogenous patient populations. Patients, and cancer itself, have proven to be very diverse, which may benefit from stratification. The method of stratifying patients into groups to tailor medical practice is called stratified or personalized medicine [60]. Patients with cancer are currently stratified using TNM staging, which can be more or less accurate depending on the type of tumor, the histology, and other factors [61-63]. TNM can broadly differentiate patients with large differences, i.e., TNM stage I vs TNM stage III/IV. Within stage, and between patients with advanced cancer (stage III/IV), it becomes more difficult to differentiate for survival chance and locoregional control (LRC) [64, 65]. Stratifying patients more accurately would allow for more options in clinical decision-making, for example by sparing those patients from invasive therapy when survival prognosis is expected to be low regardless of treatment. To improve stratification, methods using more of the available information on the patient and the disease need to be employed. Factors such as comorbidities can affect treatment of a patient immensely. However, the onset of genomics and the amount of information that this provides clinical decision makers has revealed that tumors are more individual than previously believed, but also that the amount of information which can be considered by clinicians is currently too limited [66].

AI could be used to analyze this data instead, or be added to the clinical analysis. While AI is broadly defined as the ability of a computer to do tasks that normally require human intelligence [67], in the context of this area of research and also this thesis it is meant to describe the process of drawing statistical inferences from large datasets using machine learning techniques. An example of AI in a clinical setting is ECG monitoring of a patient, where a computer is able to detect early signs of atrial fibrillation, based on a model which was trained on large amounts of ECG data. However, a more complex problem such as predicting how a patient with cancer will react to treatment or how their disease will develop will require more information.

Machine learning (ML) is a type of AI that can improve performing tasks through experience, through feeding it more and diverse data. Easily understandable examples of this are regression models, which are models that tune the variables of a function to describe the relationship between input variables and an output response. The data input into the regression model consists of one or more variables, and a certain outcome to

classify or predict for each sample. By providing many samples, the regression model can make an estimation to link the input variables to a certain output, e.g. a patient's survival, or the risk of side-effects due to a certain type of treatment.

For clinical prognostic purposes, these ML methods can be applied to create predictive models, which are designed to predict an individual patient's outcome, such as OS, loco-regional failure (LRF), distant metastasis (DM), progression-free survival (PFS), or risk of toxicity. The models use data of patients for which the outcome has already been determined, either because the patient experienced an event or because the patient exceeded a certain follow-up (FU) time. By taking this past data, the chance of a patient having an event by a certain time can be predicted for future patients.

To build these models, prognostic markers of the outcome need to be determined. Previously mentioned markers such as tumor size, the presence and number of lymph nodes, the presence and number of DM, patient age, patient sex, and more factors can all potentially be used for predictive and prognostic purposes [68]. For most cancer types, an older age and/or male sex are linked to a poor prognosis [69-71]. For a more specific case, adenocarcinoma histology in NSCLC has been linked to an increased likelihood of metastasis to the brain [72]. For HNSCC, extranodal extension of the lymph node metastases is an indicator for a poor prognosis [73, 74]. For oropharynx specifically, while Human papilloma virus (HPV) infection is an important risk factor for incidence, HPV positive oropharyngeal cancers have a higher average OS than HPV negative [75]. For BM, tumor volume, the volume of healthy brain tissue that has been irradiated, prescription dose, and previous radiation of the same ROI have been found to be predictive of ARE [57].

The information used for these predictive models can come from many sources, such as clinical patient characteristics, biological characteristics, and radiological semantic characteristics. An often-overlooked part is the quantitative information present in radiological images. For example, abdominal muscle-mass can be tied to a worse prognosis, which is easy to quantify using CT imaging of the chest [76]. The images also contain the tumor mass, which contains quantitative information of the cancer itself that can be extracted. Besides general descriptions such as tumor size and location, more intricate shape features such as the sphericity of the tumor are available, but also the textures present inside a tumor, which may indicate heterogeneity can be extracted. Radiomics is the quantification of radiological images in a number of features describing tumor phenotype, which can be analyzed using ML algorithms.

1.7 Radiomics

Radiomics as a term was first used in 2012 [77], but the concept of using a computer to perform quantitative analysis of a patients' health is much older. In 1960, automated diagnosis of patients was attempted using radiological imaging [78-80], which failed ultimately due to a lack of computational power and advanced machine learning methods. However, attempts later to instead augment a clinician's diagnosis through computer-aided diagnosis (CAD) proved to be more successful [81-83]. Radiomics similarly seeks to augment, not replace, a clinician's judgement. As it has been proven that certain protein expression patterns can be linked to radiological tumor phenotypes [84-86], quantitative analysis of medical image data could provide information that a clinical is unable to perceive qualitatively.

The radiomics pipeline can be divided into 4 steps: i) imaging and segmentation, ii) image pre-processing, iii) feature extraction and iv) data analysis. An overview of these steps can be found in Figure 4.

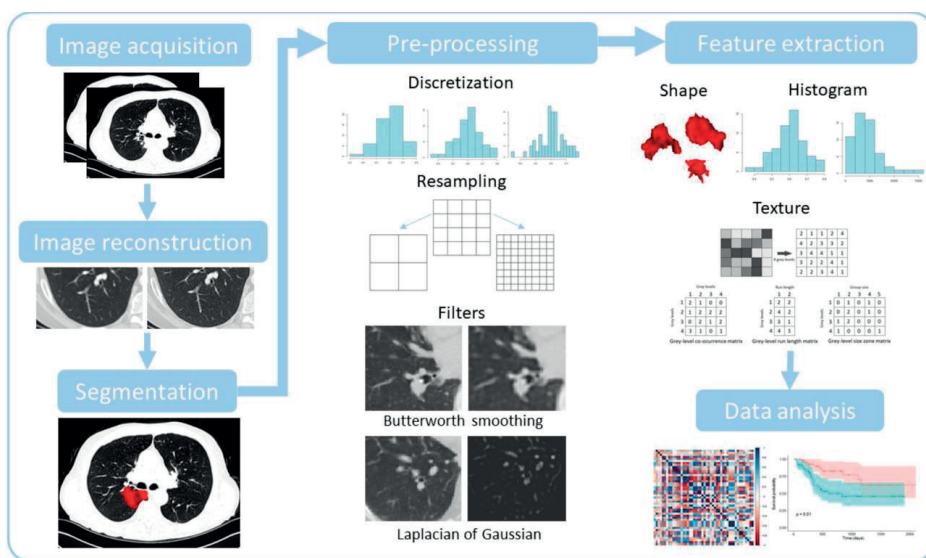


Figure 4. Schematic overview of a typical radiomics pipeline [87].

1.7.1 Imaging and segmentation

Images are acquired using certain imaging and reconstruction parameters. Which settings are used depend on the imaging task, the machine used, and the preference of technicians/clinicians. The different settings contribute to image quality, and cause a large variation in imaging to exist between and within hospitals. Even small differences

between parameters can cause large discrepancies between extracted features, negatively affecting the generalizability of the created models [88-90].

Radiomics requires the segmentation of a certain ROI to extract features. In most studies, this will be the primary tumour volume, but other regions such as the peritumoral regions or lymph nodes can also be investigated for predictive value. 3D segmentation of a (tumour) volume is usually a time-consuming task, and needs to be performed by a radiologist, radiotherapist, or similar expert. In fields such as RT, segmentation of the tumour and other ROIs is routine clinical practice. For other fields, segmentations would need to be made specifically for the radiomics based model, which may affect the feasibility of implementing radiomics in a clinical setting. However, automated methods through deep learning (DL) algorithms may allow this process to be (semi-)automatized [91-93].

1.7.2 Pre-processing

Pre-processing refers to, among others, the removal of unwanted data, imputing missing values within data, and the removal of outliers that negatively affect the data. For radiomics, image pre-processing and feature pre-processing take place. Differences in imaging and acquisition parameters cause noise and systematic differences between images. These factors make models made on certain datasets not generalizable on other datasets, and need to be harmonized through image pre-processing, which reduces noise in the image and enhances features of the image that are of interest. Radiomics includes three standard methods of pre-processing in the pipeline: intensity discretization, pixel size resampling, and spatial domain filtering.

Intensity discretization either resamples the intensity values in the image to a fixed number ('bin number discretization') or to a number of fixed intensity ranges ('binwidth discretization'). Reducing the number of intensities reduces the noise within the image, and the required computational power to extract texture features.

Pixel size resampling does two things: it changes the number of pixels in the x- and y-dimensions (the resolution) in each 2D slice, and changes the number of slices present in each 3D image, to set dimensions. This is done by either simply removing pixels and/or slices when down sampling images from higher resolutions to lower resolutions, or by interpolating when up sampling images. This causes radiomics features to be rotationally invariant [94, 95].

Lastly, image filtering refers to passing an image filter over every pixel in an image to (possibly) change the intensities, enhancing certain features and removing others. The filters applied are neighborhood operations, which means that for every pixel in an image

a kernel is applied on the pixel and neighboring pixels to determine its new intensity value. Two commonly applied image filters for a radiomics pipeline are wavelet and Laplacian of Gaussian (LoG) filters. Wavelet filtering applies 3D coif wavelet transforms along the three axes of the original images, performed at two spatial frequencies (high and low). LoG filtering applies a Laplacian filter that highlights edges in an image, after the application of a Gaussian-smoothing filter that reduces noise.

1.7.3 Feature extraction

A handcrafted radiomics feature is a pre-defined quantifiable element that can be extracted from a certain ROI within an image. An example of a simple radiomics feature is the mean pixel value, which describes the average intensity found within the ROI. On CT-images, this feature describes the density of the tumor, as a higher mean pixel value means relatively more radiation was absorbed by the tumor. Any formula that describes these pixel values, or the relation between different pixel values, can be used in a radiomics analysis. The most common radiomics features can be roughly divided in three groups. First order and histogram statistics features describe the total distribution of voxel intensities in the ROI. Shape and size features describe the 2D and 3D spatial characteristics of the ROI. Texture features describe the relative spatial distribution of the intensity values within the ROI. Texture features, and are derived from six texture matrices that are imposed over the images: gray-level co-occurrence [96], gray-level run length [97], gray-level size-zone [98], gray-level distance-zone [99], gray-level dependence [100], and neighborhood gray-tone difference matrix [101]. A description of each matrix can be found in supplementary materials 1. Depending on the feature extraction settings, hundreds of radiomics features can be extracted. However, the number of first order statistics and texture features are multiplied by the number of different filters applied to the base image, which can increase the total number of features to thousands. Supplementary materials 2 provides an overview of the most commonly used radiomics features.

1.7.4 Data analysis

The data analysis consists of two main steps: dimensionality reduction and modelling. Dimensionality reduction can be performed in a number of ways, such as eliminating redundant features, selecting predictive features, or by grouping features through methods such as principal component analysis (PCA) or least absolute shrinkage and selection operator (LASSO). Regardless of the method, dimensionality reduction shrinks the number of features to a smaller subset of (correlated) features, or a number of derivative features, which are predictive of the targeted outcome. Dimensionality reduction is necessary, as the number of features extracted from an image is generally much larger compared to the number of patients used to train the model. This may result in overfitting, which means the model was tailored too much to the data it was trained on, resulting in poor performance on other data. [102, 103].

There are a number of different feature selection methods. Features that produce not a number (NaN) can be removed, as well as features that have (close to) no variation. Furthermore, features that highly correlate can be removed, as these will have redundant information on the outcome. Features that through test-retest studies are proven not reproducible can be removed, as features should not be sensitive to small variations between scans. Lastly, there exist supervised feature selections methods that test, either univariate or multivariate, the correlation of the features to the outcome to select relevant features [104], and methods that reduce features to a set of predictive derivatives, such as PCA and LASSO [105, 106].

Using the selected features, a ML algorithm needs to be selected to train a model on. The models included in this thesis are generalized linear model (GLM) [107], Cox proportional-hazards [108], Random Forest (RF) [109], and extreme gradient boosting (XGBoost) [110]. Most models produce a prediction score from 0 to 1 whether the likelihood a binary event is true or not. This prediction score is used to create a receiving operator characteristic (ROC) curve, which summarizes the sensitivity and specificity of a binary prediction over all possible cut-offs to classify the data as true or false [111]. The area under the curve (AUC) of this ROC-curve is a measure of how well the model can discriminate, ranging from zero to one. A higher AUC means the classifier is able to better separate the positive classes from negative classes. For example, an AUC of 0.8 means there is an 80% chance the classifier is able to accurately classify a positive or negative class. The optimal AUC of 1 means that the classifier is able to perfectly separate the positive cases from negative cases, while in a worst case scenario a classifier has an AUC of 0.5, which means it has no discriminative capacity, as the chance any positive case gets successfully labelled as a positive class is 50%. An AUC below 0.5 means the classifier labels more negative classes as positive, and vice versa. This most likely indicates a problem occurred with the model building, such as samples being mislabeled. A classifier with an AUC of 0 is perfectly reciprocating the classes.

A Cox proportional-hazards model is a regression model used for survival data. These models use right-censored outcome data that include a binary outcome and a time to either an event (e.g. death or tumor recurrence) or losing a patient to follow-up. The outcome of this model is a prognostic index (PI), which is a measure from 0 to 1, which is tied to relative survival in the investigated patient population, with a longer survival time having a higher PI. This PI can be used to stratify patients into risk-groups, which shown in Kaplan-Meier curves which plot the survival chance of a population against time. Harrell's C-index (CI) is also calculated from the PI, which is a measure of how well the order of PI and actual survival times align, ranging from 0 to 1. Similar to AUC, a CI of 1 means that the classifier is able to perfectly order patients according to their survival time based on the PI. Conversely, a CI of 0.5 means that the ordering of patient survival times is completely

random, and a CI below 0.5 means the model is classifying patients reversely, indicating an error in model building.

Aerts et al. (2014) used radiomics to significantly predict OS of patients with NSCLC and HNSCC, using a cox proportional-hazards regression model trained on stage I to IV NSCLC data [112]. Radiomics has further been successfully applied for NSCLC on CT-imaging to predict OS [113-119], PFS [120-122], LRC [113, 114, 119, 122-124], DM [114, 119, 125, 126], and pathological response [127, 128]. Specifically to predict BM development for patients with NSCLC, several studies have shown radiomics has predictive value [129-131]. However, these studies usually suffer from low patient numbers, lack of external validation, and a lack of proper staging, meaning BM may have been present at baseline but not been detected. For HNSCC, radiomics on CT images has been successfully applied to predict HPV status non-invasively [132-136], OS [135, 137-140], PFS [137, 139, 141-143], LRC [132, 137, 138, 143, 144], local failure [145], detection of extra-nodal extension [146], and predicting treatment-related toxicity [147, 148]. Most of these studies are on smaller cohorts, sometimes without external validation, and a large study focussed on advanced stages of HNSCC is currently lacking. Lastly, radiomics on T1-MR for BM has proven to be able to predict LRC [149]. It has also proven to be able to differentiate RN from tumour progression on post-treatment MRI [150, 151]. However, studies on risk stratification of RN pre-treatment to our knowledge do not exist.

1.8 Deep radiomics

An alternative to using (handcrafted) radiomics is to use DL, or deep radiomics [152]. DL can directly input images into convolutional neural networks (CNN) [153]. Neural networks (NN) are models that consist of an input layer, a number of hidden layers, and an output layer. Each layer consists of nodes that connects to all the nodes of the previous and next layer. Each node has a certain weight, and if the output of a node passes a certain threshold it activates and passes information to the next layer, eventually leading to the output layer that produces a certain prediction.

CNNs are a type of NN that use convolutional filters, which are placed systematically over a 2D image input. The newly resulting 2D activation layer is used as further input in the next layers. CNNs do two things: transform the input image into feature maps, and then internally trains a model to produce a prediction. The (C)NN model is trained iteratively through backpropagation by minimizing a loss function, which is a measure of the discrepancy between predicted values and actual values. By back propagating through the model and changing nodal weights according to this loss function, the loss function is lowered. This process is then repeated, until the model converges to a minimum loss. To

train a model, a metric such as the AUC for prediction models is monitored on a separate validation dataset. DL has been used for prognosis prediction in HNSCC [154, 155], NSCLC [156-159], and BM [160].

1.9 Thesis outline: predictive modelling using radiomics in conjunction with clinical features and deep radiomics on advanced stages of cancer.

An important part of clinical decision-making is the estimation how a patient will respond to a certain treatment. Examples include an estimation of the survival chance, the chance of recurrence, or the chance of toxicity. One method that may allow better stratification patients is through radiomics, where tumor subtypes which can enhance the current risk stratification may be determined using the large amounts of information contained in radiological imaging [161]. In the introduction, several studies applying radiomics for HNSCC, NSCLC, and BM were listed. However, we identified several unsolved problems in the treatment and follow-up of stage III-IVB HNSCC, stage III NSCLC as well as BM.

We hypothesize that **quantitative information from tumor regions of advanced NSCLC, HNSCC, and brain metastases on medical imaging acquired prior to treatment is able to be predictive for survival, tumor recurrence, and toxicity related outcomes.**

This thesis is divided in the following chapters.

1.9.1 Chapter 2: current state of handcrafted radiomics

Chapter 2 provides an overview of the current state of handcrafted radiomics and future prospects of precision medicine.

1.9.2 Chapter 3 -4: Predictive radiomics for HNSCC

Current risk-stratification of advanced HNSCC patients is lacking. Identifying patients with high- and low-chances of survival pre-treatment could improve clinical decision-making. In chapter three, the feasibility of extracting radiomics features to improve prognostic prediction of patients with stage III-IVB HNSCC is investigated. The features were used to train a Cox proportional-hazards model to predict OS. This large-scale, multicenter study included retrospective and prospectively collected data, and is compared to risk-stratification using the gold standard of TNM 8th edition, and models trained on clinical and biological predictive variables. Furthermore, in Chapter four, the feasibility of including radiomics features extracted from the regions around the tumor, peritumoral regions, to improve prediction of OS, DM, and LRF, is investigated.

1.9.3 Chapter 5: Predictive radiomics for NSCLC

PCI can drastically reduce risk of BM development for patients with stage III NSCLC. However, PCI has neurological side effects, and current methods to stratify patients for risk of BM are lacking. A model that could identify patients at high risk of BM could therefore improve clinical decision-making as well as selection of patients for clinical trials evaluating BM prevention strategies. In Chapter five, the feasibility of extracting radiomics features to predict BM development in a multicentre cohort of patients with stage III NSCLC is investigated. Studies that have investigated the ability of radiomics to predict BM or DM risk in patients with NSCLC generally lack a rigorous method of staging and imaging to ensure that we know no patients had BM at baseline, and that the primary tumour is properly delineated. The data included in this study was staged using PET/CT for the lung, and dedicated brain imaging (either MRI or CECT), and was collected from multiple different centres to ensure the model is generalizable. The GLM trained on these radiomics features is compared to models trained on a list of known clinical risk factors of BM development to investigate the complementary value of radiomics.

1.9.4 Chapter 6: Predictive radiomics for BM

Current risk-stratification of patients treated with SRT for RN is lacking. Identifying patients at risk of RN could improve clinical decision-making, for example allowing clinicians and patients to opt systemic therapy (if available and having intracranial activity) instead. In Chapter six, the feasibility of using handcrafted and deep radiomics to predict ARE before treatment with SRT of BM is investigated. This study was performed using a very large training dataset (>1400 patients) and an external validation dataset (~240 patients), ensuring we have a high-quality model built on a large volume of patients which has been tested for generalizability. As the features were extracted from MR images, different pre-processing methods to harmonize the image intensity distribution within the dataset are tried out. Using the selected radiomics features, an XGBoost model is trained. This is compared to a CNN trained directly on the images, and on XGBoost models using DL extracted features and a list of clinical, biological, and treatment related predictors of ARE.

1.9.5 Chapter 7: Image quality

A major source of image quality, and subsequently for the quality of models created with radiomics and deep radiomics, is the slice spacing and slice thickness. These two parameters are usually linked to each other, with the thickness between slices matching the spacing so the image contains the entire scanned volume. A larger slice thickness means that information on a larger volume is included in each slice, leading to a lower resolution. Slice thickness has been shown to affect DL-based segmentation models [162] and predictive radiomics models [163]. Interpolation methods that increase the resolution by adding interpolated slices between existing slices could potentially increase

performance of these models. In Chapter seven, we conclude the thesis by investigating the feasibility of interpolation on chest CT images.

1.9.6 Chapter 8-12: Conclusion, impact statement, summaries, acknowledgements, Curriculum Vitae and List of publications

Chapter 8 provides a general discussion on the problems still facing the field of radiomics, future prospects, and recommendations, considering the studies and their included in this thesis. To close off the thesis, chapter 9 provides an impact statement, 10 and 11 provide summaries in English and Dutch. Finally, chapter 12 contains the acknowledgements to friends, family, and colleagues for this thesis, chapter 13 contains a curriculum vitae, and chapter 14 provides a list of publications which include me as (co-)author.









Radiomics for the management of cancer in advanced stages	1	Introduction and outline of thesis	
	2	Present state and future of radiomics	
	3	Radiomics for HNSCC survival	
	4	HNSCC peritumoral radiomics	
	5	Radiomics for prediction of NSCLC BM	
	6	Radiomics for prediction of BM radionecrosis	
	7	Slice interpolation for chest CT images	
	8	Discussion and conclusion	

Figure 5. Table of contents.

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin.* 2015;65(2):87-108.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin.* 2021;71(3):209-49.
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin.* 2020;70(1):7-30.
4. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin.* 2019;69(1):7-34.
5. Thompson-Harvey A, Yetukuri M, Hansen AR, Simpson MC, Adjei Boakye E, Varvares MA, et al. Rising incidence of late-stage head and neck cancer in the United States. *Cancer.* 2020;126(5):1090-101.
6. Houston KA, Mitchell KA, King J, White A, Ryan BM. Histologic Lung Cancer Incidence Rates and Trends Vary by Race/Ethnicity and Residential County. *Journal of Thoracic Oncology.* 2018;13(4):497-509.
7. Roughley A, Damonte E, Taylor-Stokes G, Rider A, Munk VC. Impact of Brain Metastases on Quality of Life and Estimated Life Expectancy in Patients with Advanced Non-Small Cell Lung Cancer. *Value Health.* 2014;17(7):A650.
8. Mroz EA, Rocco JW. Intra-tumor heterogeneity in head and neck cancer and its clinical implications. *World J Otorhinolaryngol Head Neck Surg.* 2016;2(2):60-7.
9. Berghoff AS, Schur S, Fureder LM, Gatterbauer B, Dieckmann K, Widhalm G, et al. Descriptive statistical analysis of a real life cohort of 2419 patients with brain metastases of solid cancers. *ESMO Open.* 2016;1(2):e000024.
10. Waqar SN, Samson PP, Robinson CG, Bradley J, Devarakonda S, Du L, et al. Non-small-cell Lung Cancer With Brain Metastasis at Presentation. *Clin Lung Cancer.* 2018;19(4):e373-e9.
11. O'Beirn M, Benghiat H, Meade S, Heyes G, Sawlani V, Kong A, et al. The Expanding Role of Radiosurgery for Brain Metastases. *Medicines (Basel).* 2018;5(3).
12. Brown PD, Ahluwalia MS, Khan OH, Asher AL, Wefel JS, Gondi V. Whole-Brain Radiotherapy for Brain Metastases: Evolution or Revolution? *J Clin Oncol.* 2018;36(5):483-91.
13. Collins LG, Haines C, Perkel R, Enck RE. Lung cancer: diagnosis and management. *Am Fam Physician.* 2007;75(1):56-63.
14. Zanoni DK, Patel SG, Shah JP. Changes in the 8th Edition of the American Joint Committee on Cancer (AJCC) Staging of Head and Neck Cancer: Rationale and Implications. *Curr Oncol Rep.* 2019;21(6):52.
15. Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med.* 2010;362(25):2380-8.
16. Solomon BJ, Mok T, Kim DW, Wu YL, Nakagawa K, Mekhail T, et al. First-line crizotinib versus chemotherapy in ALK-positive lung cancer. *N Engl J Med.* 2014;371(23):2167-77.

17. Soo RA, Lim SM, Syn NL, Teng R, Soong R, Mok TSK, et al. Immune checkpoint inhibitors in epidermal growth factor receptor mutant non-small cell lung cancer: Current controversies and future directions. *Lung cancer*. 2018;115:12-20.
18. Lee CK, Man J, Lord S, Cooper W, Links M, GebSKI V, et al. Clinical and Molecular Characteristics Associated With Survival Among Patients Treated With Checkpoint Inhibitors for Advanced Non-Small Cell Lung Carcinoma: A Systematic Review and Meta-analysis. *JAMA Oncol*. 2018;4(2):210-6.
19. Gandini S, Botteri E, Iodice S, Boniol M, Lowenfels AB, Maisonneuve P, et al. Tobacco smoking and cancer: A meta-analysis. *International Journal of Cancer*. 2008;122(1):155-64.
20. Connor J. Alcohol consumption as a cause of cancer. *Addiction*. 2017;112(2):222-8.
21. Tas F, Ciftci R, Kilic L, Karabulut S. Age is a prognostic factor affecting survival in lung cancer patients. *Oncology letters*. 2013;6(5):1507-13.
22. van der Schroeff MP, Derks W, Hordijk GJ, de Leeuw RJ. The effect of age on survival and quality of life in elderly head and neck cancer patients: a long-term prospective study. *European Archives of Oto-Rhino-Laryngology*. 2007;264(4):415-22.
23. Cook MB, McGlynn KA, Devesa SS, Freedman ND, Anderson WF. Sex Disparities in Cancer Mortality and Survival. *Cancer Epidemiology, Biomarkers & Prevention*. 2011;20(8):1629-37.
24. Littlewood TJ. The impact of hemoglobin levels on treatment outcomes in patients with cancer. *Seminars in Oncology*. 2001;28:49-53.
25. Bloodgood JC. Radiosensitive Tumors and Tumors that First should be Subjected to Operation. *Radiology*. 1930;14(3):254-62.
26. Moeller BJ, Richardson RA, Dewhirst MW. Hypoxia and radiotherapy: opportunities for improved outcomes in cancer treatment. *Cancer and Metastasis Reviews*. 2007;26(2):241-8.
27. Dubben H-H, Thames HD, Beck-Bornholdt H-P. Tumor volume: a basic and specific response predictor in radiotherapy. *Radiother Oncol*. 1998;47(2):167-74.
28. De Ruyscher D, Niedermann G, Burnet NG, Siva S, Lee AW, Hegi-Johnson F. Radiotherapy toxicity. *Nature Reviews Disease Primers*. 2019;5(1):1-20.
29. Langendijk JA, Doornaert P, Verdonck-de Leeuw IM, Leemans CR, Aaronson NK, Slotman BJ. Impact of late treatment-related toxicity on quality of life among patients with head and neck cancer treated with radiotherapy. *Journal of clinical oncology*. 2008;26(22):3770-6.
30. Goguen LA, Posner MR, Norris CM, Tishler RB, Wirth LJ, Annino DJ, et al. Dysphagia after sequential chemoradiation therapy for advanced head and neck cancer. *Otolaryngology—Head and Neck Surgery*. 2006;134(6):916-22.
31. Hope A, El Naqa I, Bradley J, Vivic M, Lindsay P, Bosch W, et al. Radiation pneumonitis/fibrosis risk based on dosimetric, clinical, and location-related factors. *International Journal of Radiation Oncology, Biology, Physics*. 2004;60(1):S204.
32. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424.
33. Chow LQM. Head and Neck Cancer. *N Engl J Med*. 2020;382(1):60-72.

34. Howren MB, Christensen AJ, Karnell LH, Funk GF. Psychological factors associated with head and neck cancer treatment and survivorship: evidence and opportunities for behavioral medicine. *J Consult Clin Psychol*. 2013;81(2):299-317.
35. Chen SC, Lai YH, Liao CT, Chang JT, Lin CC. Unmet information needs and preferences in newly diagnosed and surgically treated oral cavity cancer patients. *Oral Oncol*. 2009;45(11):946-52.
36. Murphy BA, Dietrich MS, Wells N, Dwyer K, Ridner SH, Silver HJ, et al. Reliability and validity of the Vanderbilt Head and Neck Symptom Survey: a tool to assess symptom burden in patients treated with chemoradiation. *Head Neck*. 2010;32(1):26-37.
37. Barta JA, Powell CA, Wisnivesky JP. Global Epidemiology of Lung Cancer. *Ann Glob Health*. 2019;85(1).
38. Travis WD. Pathology of lung cancer. *Clin Chest Med*. 2011;32(4):669-92.
39. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol*. 2016;11(1):39-51.
40. Reck M, Rodriguez-Abreu D, Robinson AG, Hui R, Czoszi T, Fulop A, et al. Five-Year Outcomes With Pembrolizumab Versus Chemotherapy for Metastatic Non-Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score \geq 50. *J Clin Oncol*. 2021;39(21):2339-49.
41. Mok T, Camidge DR, Gadgeel SM, Rosell R, Dziadziuszko R, Kim DW, et al. Updated overall survival and final progression-free survival data for patients with treatment-naive advanced ALK-positive non-small-cell lung cancer in the ALEX study. *Ann Oncol*. 2020;31(8):1056-64.
42. Hendriks LEL, Dingemans A-MC, De Ruysscher DKM, Aarts MJ, Barberio L, Cornelissen R, et al. Lung Cancer in the Netherlands. *Journal of Thoracic Oncology*. 2021;16(3):355-65.
43. de Jong EE, van Elmpt W, Leijenaar RT, Hoekstra OS, Groen HJ, Smit EF, et al. [18F]FDG PET/CT-based response assessment of stage IV non-small cell lung cancer treated with paclitaxel-carboplatin-bevacizumab with or without nitroglycerin patches. *Eur J Nucl Med Mol Imaging*. 2017;44(1):8-16.
44. Antonia SJ, Villegas A, Daniel D, Vicente D, Murakami S, Hui R, et al. Overall Survival with Durvalumab after Chemoradiotherapy in Stage III NSCLC. *New England Journal of Medicine*. 2018;379(24):2342-50.
45. Witlox WJA, Ramaekers BLT, Zindler JD, Eekers DBP, van Loon JGM, Hendriks LEL, et al. The Prevention of Brain Metastases in Non-Small Cell Lung Cancer by Prophylactic Cranial Irradiation. *Front Oncol*. 2018;8:241.
46. Ries L, Eisner M, Kosary C, Hankey B, Miller B, Clegg L, et al. SEER cancer statistics review, 1975–2000. Bethesda, MD: National Cancer Institute. 2003;2.
47. Walker AE, Robins M, Weinfeld FD. Epidemiology of brain tumors: the national survey of intracranial neoplasms. *Neurology*. 1985;35(2):219-26.
48. Percy AK, Elveback LR, Okazaki H, Kurland LT. Neoplasms of the central nervous system. Epidemiologic considerations. *Neurology*. 1972;22(1):40-8.
49. Soffietti R, Ruda R, Mutani R. Management of brain metastases. *J Neurol*. 2002;249(10):1357-69.

50. Gavrilovic IT, Posner JB. Brain metastases: epidemiology and pathophysiology. *J Neurooncol.* 2005;75(1):5-14.
51. Nayak L, Lee EQ, Wen PY. Epidemiology of brain metastases. *Curr Oncol Rep.* 2012;14(1):48-54.
52. Sperduto PW, Mesko S, Li J, Cagney D, Aizer A, Lin NU, et al. Survival in Patients With Brain Metastases: Summary Report on the Updated Diagnosis-Specific Graded Prognostic Assessment and Definition of the Eligibility Quotient. *J Clin Oncol.* 2020;38(32):3773-84.
53. Sperduto PW, Mesko S, Li J, Cagney D, Aizer A, Lin NU, et al. Survival in Patients With Brain Metastases: Summary Report on the Updated Diagnosis-Specific Graded Prognostic Assessment and Definition of the Eligibility Quotient. *Journal of Clinical Oncology.* 2020;38(32):3773-84.
54. Sperduto PW, De B, Li J, Carpenter D, Kirkpatrick J, Milligan M, et al. The Graded Prognostic Assessment (GPA) for Lung Cancer Patients with Brain Metastases: Initial Report of the Small Cell Lung Cancer GPA and Update of the Non-Small Cell Lung Cancer GPA including the Effect of Programmed Death Ligand-1 (PD-L1) and Other Prognostic Factors. *Int J Radiat Oncol Biol Phys.* 2022.
55. Gu L, Qing S, Zhu X, Ju X, Cao Y, Jia Z, et al. Stereotactic Radiation Therapy (SRT) for Brain Metastases of Multiple Primary Tumors: A Single Institution Retrospective Analysis. *Front Oncol.* 2019;9:1352.
56. Kraft J, Mayinger M, Willmann J, Brown M, Tanadini-Lang S, Wilke L, et al. Management of multiple brain metastases: a patterns of care survey within the German Society for Radiation Oncology. *J Neurooncol.* 2021;152(2):395-404.
57. Sneed PK, Mendez J, Vemer-van den Hoek JG, Seymour ZA, Ma L, Molinaro AM, et al. Adverse radiation effect after stereotactic radiosurgery for brain metastases: incidence, time course, and risk factors. *J Neurosurg.* 2015;123(2):373-86.
58. Le Rhun E, Dhermain F, Vogin G, Reynolds N, Metellus P. Radionecrosis after stereotactic radiotherapy for brain metastases. *Expert Rev Neurother.* 2016;16(8):903-14.
59. Loganadane G, Dhermain F, Louvel G, Kuv P, Deutsch E, Le Pécoux C, et al. Brain Radiation Necrosis: Current Management With a Focus on Non-small Cell Lung Cancer Patients. *Frontiers in Oncology.* 2018;8.
60. Krzyszczyk P, Acevedo A, Davidoff EJ, Timmins LM, Marrero-Berrios I, Patel M, et al. The growing role of precision and personalized medicine for cancer treatment. *Technology (Singap World Sci).* 2018;6(3-4):79-100.
61. Cascinu S, Staccioli MP, Gasparini G, Giordani P, Catalano V, Ghiselli R, et al. Expression of vascular endothelial growth factor can predict event-free survival in stage II colon cancer. *Clin Cancer Res.* 2000;6(7):2803-7.
62. Wang Z, Jiang W, Zheng L, Yan J, Dai J, Huang C, et al. Consideration of Age Is Necessary for Increasing the Accuracy of the AJCC TNM Staging System of Pancreatic Neuroendocrine Tumors. *Frontiers in Oncology.* 2019;9.
63. Navani N, Fisher DJ, Tierney JF, Stephens RJ, Burdett S, Burdett S, et al. The Accuracy of Clinical Staging of Stage I-IIa Non-Small Cell Lung Cancer: An Analysis Based on Individual Participant Data. *Chest.* 2019;155(3):502-9.

64. Lamont EB, Christakis NA. Complexities in prognostication in advanced cancer: “to help them live their lives the way they want to”. *JAMA*. 2003;290(1):98-104.
65. Christakis NA, Lamont EB. Extent and determinants of error in doctors’ prognoses in terminally ill patients: prospective cohort study. *BMJ*. 2000;320(7233):469-72.
66. Burney IA, Lakhtakia R. Precision Medicine: Where have we reached and where are we headed? *Sultan Qaboos Univ Med J*. 2017;17(3):e255-e8.
67. Briganti G, Le Moine O. Artificial Intelligence in Medicine: Today and Tomorrow. *Front Med (Lausanne)*. 2020;7:27.
68. Zandwijk Nv, Mooi WJ, Rodenhuis S. Prognostic factors in NSCLC. Recent experiences. *Lung cancer*. 1995;12:S27-S33.
69. 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*. 2010;68(3):398-402.
70. Tammemagi CM, Neslund-Dudas C, Simoff M, Kvale P. In lung cancer patients, age, race-ethnicity, gender and smoking predict adverse comorbidity, which in turn predicts treatment and survival. *Journal of Clinical Epidemiology*. 2004;57(6):597-609.
71. Baatenburg de Jong RJ, Hermans J, Molenaar J, Briare JJ, le Cessie S. Prediction of survival in patients with head and neck cancer. *Head & Neck*. 2001;23(9):718-24.
72. Hung JJ, Jeng WJ, Hsu WH, Chou TY, Huang BS, Wu YC. Predictors of death, local recurrence, and distant metastasis in completely resected pathological stage-I non-small-cell lung cancer. *J Thorac Oncol*. 2012;7(7):1115-23.
73. Wang Z, Zeng Q, Li Y, Lu T, Liu C, Hu G. Extranodal Extension as an Independent Prognostic factor in Laryngeal Squamous Cell Carcinoma Patients. *J Cancer*. 2020;11(24):7196-201.
74. Beltz A, Zimmer S, Michaelides I, Evert K, Psychogios G, Bohr C, et al. Significance of Extranodal Extension in Surgically Treated HPV-Positive Oropharyngeal Carcinomas. *Front Oncol*. 2020;10:1394.
75. Sabatini ME, Chiocca S. Human papillomavirus as a driver of head and neck cancers. *British Journal of Cancer*. 2020;122(3):306-14.
76. Rier HN, Jager A, Sleijfer S, Maier AB, Levin MD. The Prevalence and Prognostic Value of Low Muscle Mass in Cancer Patients: A Review of the Literature. *Oncologist*. 2016;21(11):1396-409.
77. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-6.
78. Meyers PH, Nice CM, Jr., Becker HC, Nettleton WJ, Jr., Sweeney JW, Meckstroth GR. Automated Computer Analysis of Radiographic Images. *Radiology*. 1964;83:1029-34.
79. Lodwick GS, Haun CL, Smith WE, Keller RF, Robertson ED. Computer diagnosis of primary bone tumors: A preliminary report. *Radiology*. 1963;80(2):273-5.
80. Winsberg F, Elkin M, Macy Jr J, Bordaz V, Weymouth W. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology*. 1967;89(2):211-5.

81. Chan HP, Doi K, Vyborny CJ, Schmidt RA, Metz CE, Lam KL, et al. Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis. *Invest Radiol.* 1990;25(10):1102-10.
82. Kobayashi T, Xu X-W, MacMahon H, Metz CE, Doi K. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. *Radiology.* 1996;199(3):843-8.
83. Li F, Arimura H, Suzuki K, Shiraishi J, Li Q, Abe H, et al. Computer-aided detection of peripheral lung cancers missed at CT: ROC analyses without and with localization. *Radiology.* 2005;237(2):684-90.
84. Hobbs SK, Shi G, Homer R, Harsh G, Atlas SW, Bednarski MD. Magnetic resonance image-guided proteomics of human glioblastoma multiforme. *J Magn Reson Imaging.* 2003;18(5):530-6.
85. 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(7):821-31.
86. O'Connor JP, Jayson GC, Jackson A, Ghiorghiu D, Carrington BM, Rose CJ, et al. Enhancing fraction predicts clinical outcome following first-line chemotherapy in patients with epithelial ovarian carcinoma. *Clin Cancer Res.* 2007;13(20):6130-5.
87. Fornacon-Wood I, Faivre-Finn C, O'Connor JPB, Price GJ. Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype. *Lung cancer.* 2020;146:197-208.
88. Neisius U, El-Rewaidy H, Nakamori S, Rodriguez J, Manning WJ, Nezafat R. Radiomic Analysis of Myocardial Native T1 Imaging Discriminates Between Hypertensive Heart Disease and Hypertrophic Cardiomyopathy. *JACC Cardiovasc Imaging.* 2019;12(10):1946-54.
89. Castellano G, Bonilha L, Li LM, Cendes F. Texture analysis of medical images. *Clin Radiol.* 2004;59(12):1061-9.
90. van Timmeren JE, Leijenaar RTH, van Elmpt W, Wang J, Zhang Z, Dekker A, et al. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography.* 2016;2(4):361-5.
91. Webb JM, Adusei SA, Wang Y, Samreen N, Adler K, Meixner DD, et al. Comparing deep learning-based automatic segmentation of breast masses to expert interobserver variability in ultrasound imaging. *Comput Biol Med.* 2021;139:104966.
92. Zhou X. Automatic Segmentation of Multiple Organs on 3D CT Images by Using Deep Learning Approaches. *Adv Exp Med Biol.* 2020;1213:135-47.
93. Zhu HT, Zhang XY, Shi YJ, Li XT, Sun YS. Automatic segmentation of rectal tumor on diffusion-weighted images by deep learning with U-Net. *J Appl Clin Med Phys.* 2021;22(9):324-31.
94. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys.* 2017;44(3):1050-62.
95. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology.* 2020;295(2):328-38.

96. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*. 1973;SMC-3(6):610-21.
97. Galloway MM. Texture analysis using grey level run lengths. 1974 July 01, 1974.
98. Thibault G, Fertil B, Navarro C, Pereira S, Lévy N, Sequeira J, et al. Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification 2009.
99. Thibault G, Angulo J, Meyer F, editors. Advanced statistical matrices for texture characterization: Application to DNA chromatin and microtubule network classification. 18th IEEE International Conference on Image Processing (ICIP); 2011 2011-09-11; Bruxelles, Belgium: IEEE.
100. Sun C, Wee WG. Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing*. 1983;23(3):341-52.
101. Amadasun M, King R. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics*. 1989;19(5):1264-74.
102. Ying X, editor An overview of overfitting and its solutions. *Journal of Physics: Conference Series*; 2019: IOP Publishing.
103. Dietterich T. Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*. 1995;27(3):326-7.
104. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: A data perspective. *ACM computing surveys (CSUR)*. 2017;50(6):1-45.
105. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996;58(1):267-88.
106. Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*. 1933;24(6):417-41.
107. Nelder JA, Wedderburn RWM. Generalized Linear Models. *Journal of the Royal Statistical Society Series A (General)*. 1972;135(3):370-84.
108. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B (Methodological)*. 1972;34(2):187-220.
109. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*: Routledge; 2017.
110. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*. 2000;28(2):337-407, 71.
111. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
112. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*. 2014;5(1):4006.
113. Oikonomou A, Khalvati F, Tyrrell PN, Haider MA, Tarique U, Jimenez-Juan L, et al. Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy. *Sci Rep*. 2018;8(1):4003.

114. Yu W, Tang C, Hobbs BP, Li X, Koay EJ, Wistuba, II, et al. Development and Validation of a Predictive Radiomics Model for Clinical Outcomes in Stage I Non-small Cell Lung Cancer. *Int J Radiat Oncol Biol Phys.* 2018;102(4):1090-7.
115. Starkov P, Aguilera TA, Golden DI, Shultz DB, Trakul N, Maxim PG, et al. The use of texture-based radiomics CT analysis to predict outcomes in early-stage non-small cell lung cancer treated with stereotactic ablative radiotherapy. *Br J Radiol.* 2019;92(1094):20180228.
116. Wang L, Dong T, Xin B, Xu C, Guo M, Zhang H, et al. Integrative nomogram of CT imaging, clinical, and hematological features for survival prediction of patients with locally advanced non-small cell lung cancer. *Eur Radiol.* 2019;29(6):2958-67.
117. Ramella S, Fiore M, Greco C, Cordelli E, Sicilia R, Merone M, et al. A radiomic approach for adaptive radiotherapy in non-small cell lung cancer patients. *Plos One.* 2018;13(11):e0207455.
118. Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. *Sci Rep.* 2017;7(1):588.
119. Akinci D'Antonoli T, Farchione A, Lenkowicz J, Chiappetta M, Cicchetti G, Martino A, et al. CT Radiomics Signature of Tumor and Peritumoral Lung Parenchyma to Predict Nonsmall Cell Lung Cancer Postsurgical Recurrence Risk. *Academic Radiology.* 2020;27(4):497-507.
120. Tunalı I, Gray JE, Qi J, Abdalah M, Jeong DK, Guvenis A, et al. Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: An early report. *Lung cancer.* 2019;129:75-9.
121. Song J, Dong D, Huang Y, Zang Y, Liu Z, Tian J, editors. Association between tumor heterogeneity and progression-free survival in non-small cell lung cancer patients with EGFR mutations undergoing tyrosine kinase inhibitors therapy. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2016 16-20 Aug. 2016.
122. Franceschini D, Cozzi L, De Rose F, Navarra P, Fogliata A, Franzese C, et al. A radiomic approach to predicting nodal relapse and disease-specific survival in patients treated with stereotactic body radiation therapy for early-stage non-small cell lung cancer. *Strahlenther Onkol.* 2020;196(10):922-31.
123. Dissaux G, Visvikis D, Da-Ano R, Pradier O, Chajon E, Barillot I, et al. Pretreatment (18)F-FDG PET/CT Radiomics Predict Local Recurrence in Patients Treated with Stereotactic Body Radiotherapy for Early-Stage Non-Small Cell Lung Cancer: A Multicentric Study. *J Nucl Med.* 2020;61(6):814-20.
124. Mattonen SA, Palma DA, Haasbeek CJ, Senan S, Ward AD. Early prediction of tumor recurrence based on CT texture changes after stereotactic ablative radiotherapy (SABR) for lung cancer. *Med Phys.* 2014;41(3):033502.
125. Fan L, Fang M, Tu W, Zhang D, Wang Y, Zhou X, et al. Radiomics Signature: A Biomarker for the Preoperative Distant Metastatic Prediction of Stage I Nonsmall Cell Lung Cancer. *Acad Radiol.* 2019;26(9):1253-61.
126. Huynh E, Coroller TP, Narayan V, Agrawal V, Hou Y, Romano J, et al. CT-based radiomic analysis of stereotactic body radiation therapy patients with lung cancer. *Radiother Oncol.* 2016;120(2):258-66.
127. Coroller TP, Agrawal V, Narayan V, Hou Y, Grossmann P, Lee SW, et al. Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiother Oncol.* 2016;119(3):480-6.

128. Coroller TP, Agrawal V, Huynh E, Narayan V, Lee SW, Mak RH, et al. Radiomic-Based Pathological Response Prediction from Primary Tumors and Lymph Nodes in NSCLC. *J Thorac Oncol.* 2017;12(3):467-76.
129. Chen A, Lu L, Pu X, Yu T, Yang H, Schwartz LH, et al. CT-Based Radiomics Model for Predicting Brain Metastasis in Category T1 Lung Adenocarcinoma. *AJR Am J Roentgenol.* 2019;213(1):134-9.
130. Xu X, Huang L, Chen J, Wen J, Liu D, Cao J, et al. Application of radiomics signature captured from pretreatment thoracic CT to predict brain metastases in stage III/IV ALK-positive non-small cell lung cancer patients. *J Thorac Dis.* 2019;11(11):4516-28.
131. Sun F, Chen Y, Chen X, Sun X, Xing L. CT-based radiomics for predicting brain metastases as the first failure in patients with curatively resected locally advanced non-small cell lung cancer. *Eur J Radiol.* 2021;134:109411.
132. Bogowicz M, Riesterer O, Ikenberg K, Stieb S, Moch H, Studer G, et al. Computed Tomography Radiomics Predicts HPV Status and Local Tumor Control After Definitive Radiochemotherapy in Head and Neck Squamous Cell Carcinoma. *Int J Radiat Oncol Biol Phys.* 2017;99(4):921-8.
133. Huang C, Cintra M, Brennan K, Zhou M, Colevas AD, Fischbein N, et al. Development and validation of radiomic signatures of head and neck squamous cell carcinoma molecular features and subtypes. *EBioMedicine.* 2019;45:70-80.
134. Leijenaar RT, Bogowicz M, Jochems A, Hoebbers FJ, Wesseling FW, Huang SH, et al. Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: a multicenter study. *Br J Radiol.* 2018;91(1086):20170498.
135. Parmar C, Leijenaar RT, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. *Sci Rep.* 2015;5:11044.
136. Yu K, Zhang Y, Yu Y, Huang C, Liu R, Li T, et al. Radiomic analysis in prediction of Human Papilloma Virus status. *Clin Transl Radiat Oncol.* 2017;7:49-54.
137. Cozzi L, Franzese C, Fogliata A, Franceschini D, Navarria P, Tomatis S, et al. Predicting survival and local control after radiochemotherapy in locally advanced head and neck cancer by means of computed tomography based radiomics. *Strahlenther Onkol.* 2019;195(9):805-18.
138. Leger S, Zwanenburg A, Pilz K, Lohaus F, Linge A, Zophel K, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Sci Rep.* 2017;7(1):13206.
139. Ou D, Blanchard P, Rosellini S, Levy A, Nguyen F, Leijenaar RTH, et al. Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status. *Oral Oncol.* 2017;71:150-5.
140. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJ. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. *Front Oncol.* 2015;5:272.
141. Lv W, Yuan Q, Wang Q, Ma J, Feng Q, Chen W, et al. Radiomics Analysis of PET and CT Components of PET/CT Imaging Integrated with Clinical Parameters: Application to Prognosis for Nasopharyngeal Carcinoma. *Mol Imaging Biol.* 2019;21(5):954-64.

142. Mo X, Wu X, Dong D, Guo B, Liang C, Luo X, et al. Prognostic value of the radiomics-based model in progression-free survival of hypopharyngeal cancer treated with chemoradiation. *Eur Radiol.* 2020;30(2):833-43.
143. Zhai TT, Langendijk JA, van Dijk LV, Halmos GB, Witjes MJH, Oosting SF, et al. The prognostic value of CT-based image-biomarkers for head and neck cancer patients treated with definitive (chemo-)radiation. *Oral Oncol.* 2019;95:178-86.
144. Bogowicz M, Tanadini-Lang S, Guckenberger M, Riesterer O. Combined CT radiomics of primary tumor and metastatic lymph nodes improves prediction of loco-regional control in head and neck cancer. *Sci Rep.* 2019;9(1):15198.
145. Kuno H, Qureshi MM, Chapman MN, Li B, Andreu-Arasa VC, Onoue K, et al. CT Texture Analysis Potentially Predicts Local Failure in Head and Neck Squamous Cell Carcinoma Treated with Chemoradiotherapy. *AJNR Am J Neuroradiol.* 2017;38(12):2334-40.
146. Kann BH, Aneja S, Loganadane GV, Kelly JR, Smith SM, Decker RH, et al. Pretreatment Identification of Head and Neck Cancer Nodal Metastasis and Extranodal Extension Using Deep Learning Neural Networks. *Sci Rep.* 2018;8(1):14036.
147. Sheikh K, Lee SH, Cheng Z, Lakshminarayanan P, Peng L, Han P, et al. Predicting acute radiation induced xerostomia in head and neck Cancer using MR and CT Radiomics of parotid and submandibular glands. *Radiat Oncol.* 2019;14(1):131.
148. Abdollahi H, Mostafaei S, Cheraghi S, Shiri I, Rabi Mahdavi S, Kazemnejad A. Cochlea CT radiomics predicts chemoradiotherapy induced sensorineural hearing loss in head and neck cancer patients: A machine learning and multi-variable modelling study. *Phys Med.* 2018;45:192-7.
149. Mouraviev A, Detsky J, Sahgal A, Ruschin M, Lee YK, Karam I, et al. Use of radiomics for the prediction of local control of brain metastases after stereotactic radiosurgery. *Neuro Oncol.* 2020;22(6):797-805.
150. Peng L, Parekh V, Huang P, Lin DD, Sheikh K, Baker B, et al. Distinguishing True Progression From Radionecrosis After Stereotactic Radiation Therapy for Brain Metastases With Machine Learning and Radiomics. *Int J Radiat Oncol Biol Phys.* 2018;102(4):1236-43.
151. Zhang Z, Yang J, Ho A, Jiang W, Logan J, Wang X, et al. A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from MR images. *Eur Radiol.* 2018;28(6):2255-63.
152. Schmidhuber J. Deep learning in neural networks: An overview. *Neural Networks.* 2015;61:85-117.
153. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics.* 1980;36(4):193-202.
154. Wahid KA, He R, Dede C, Mohamed ASR, Abdelaal MA, van Dijk LV, et al. Combining Tumor Segmentation Masks with PET/CT Images and Clinical Data in a Deep Learning Framework for Improved Prognostic Prediction in Head and Neck Squamous Cell Carcinoma. *medRxiv.* 2021.
155. Naser MA, Wahid KA, Mohamed AS, Abdelaal MA, He R, Dede C, et al. Progression Free Survival Prediction for Head and Neck Cancer using Deep Learning based on Clinical and PET-CT Imaging Data. *medRxiv.* 2021.

156. Jiao Z, Li H, Xiao Y, Dorsey J, Simone CB, Feigenberg S, et al. Integration of Deep Learning Radiomics and Counts of Circulating Tumor Cells Improves Prediction of Outcomes of Early Stage NSCLC Patients Treated With Stereotactic Body Radiation Therapy. *International Journal of Radiation Oncology* Biology* Physics*. 2022;112(4):1045-54.
157. Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin Cancer Res*. 2019;25(11):3266-75.
158. Cherukuri N, Bethapudi NR, Thotakura VSK, Chitturi P, Basha CZ, Mummidi RM, editors. Deep Learning for Lung Cancer Prediction using NSCLS patients CT Information. 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS); 2021: IEEE.
159. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, et al. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. *PLoS medicine*. 2018;15(11):e1002711.
160. Cha YJ, Jang WI, Kim M-S, Yoo HJ, Paik EK, Jeong HK, et al. Prediction of response to stereotactic radiosurgery for brain metastases using convolutional neural networks. *Anticancer Research*. 2018;38(9):5437-45.
161. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *European Journal of Cancer*. 2012;48(4):441-6.
162. Huang K, Rhee DJ, Ger R, Layman R, Yang J, Cardenas CE, et al. Impact of slice thickness, pixel size, and CT dose on the performance of automatic contouring algorithms. *Journal of Applied Clinical Medical Physics*. 2021;22(5):168-74.
163. Li Y, Lu L, Xiao M, Derclé L, Huang Y, Zhang Z, et al. CT Slice Thickness and Convolution Kernel Affect Performance of a Radiomic Model for Predicting EGFR Status in Non-Small Cell Lung Cancer: A Preliminary Study. *Sci Rep-Uk*. 2018;8(1):17913.

1.10 Supplementary materials

1.10.1 Gray Level Co-occurrence Matrix (GLCM) Features

A Gray Level Co-occurrence Matrix (GLCM) of size $N_g \times N_g$ describes the second-order joint probability function of an image region constrained by the mask and is defined as $P(i, j|\delta, \theta)$. The $(i, j)^{\text{th}}$ element of this matrix represents the number of times the combination of levels i and j occur in two pixels in the image, that are separated by a distance of δ pixels along angle θ . The distance δ from the center voxel is defined as the distance according to the infinity norm. For $\delta = 1$, this results in 2 neighbors for each of 13 angles in 3D (26-connectivity) and for $\delta = 2$ a 98-connectivity (49 unique angles).

As a two dimensional example, let the following matrix I represent a 5x5 image, having 5 discrete grey levels:

$$I = \begin{bmatrix} 1 & 2 & 5 & 2 & 3 \\ 3 & 2 & 1 & 3 & 1 \\ 1 & 3 & 5 & 5 & 2 \\ 1 & 1 & 1 & 1 & 2 \\ 1 & 2 & 4 & 3 & 5 \end{bmatrix}$$

For distance $\delta = 1$ (considering pixels with a distance of 1 pixel from each other) and angle $\theta = 0^\circ$ (horizontal plane, i.e. voxels to the left and right of the center voxel), the following symmetrical GLCM is obtained:

$$P = \begin{bmatrix} 6 & 4 & 3 & 0 & 0 \\ 4 & 0 & 2 & 1 & 3 \\ 3 & 2 & 0 & 1 & 2 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 3 & 2 & 0 & 2 \end{bmatrix}$$

1.10.2 Gray Level Run Length Matrix (GLRLM) Features

A Gray Level Run Length Matrix (GLRLM) quantifies gray level runs, which are defined as the length in number of pixels, of consecutive pixels that have the same gray level value. In a gray level run length matrix $P(i, j|\theta)$, the (i, j) th element describes the number of runs with gray level i and length j occur in the image (ROI) along angle θ .

As a two dimensional example, consider the following 5x5 image, with 5 discrete gray levels:

$$I = \begin{bmatrix} 5 & 2 & 5 & 4 & 4 \\ 3 & 3 & 3 & 1 & 3 \\ 2 & 1 & 1 & 1 & 3 \\ 4 & 2 & 2 & 2 & 3 \\ 3 & 5 & 3 & 3 & 2 \end{bmatrix}$$

The GLRLM for $\theta = 0$, where 0 degrees is the horizontal direction, then becomes:

$$P = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 \\ 4 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 \end{bmatrix}$$

1.10.3 Gray Level Size Zone Matrix (GLSZM) Features

A Gray Level Size Zone (GLSZM) quantifies gray level zones in an image. A gray level zone is defined as the number of connected voxels that share the same gray level intensity. A voxel is considered connected if the distance is 1 according to the infinity norm (26-connected region in a 3D, 8-connected region in 2D). In a gray level size zone matrix (i, j) the (i, j) th element equals the number of zones with gray level i and size j appear in image. Contrary to GLCM and GLRLM, the GLSZM is rotation independent, with only one matrix calculated for all directions in the ROI.

As a two dimensional example, consider the following 5x5 image, with 5 discrete gray levels:

$$I = \begin{bmatrix} 5 & 2 & 5 & 4 & 4 \\ 3 & 3 & 3 & 1 & 3 \\ 2 & 1 & 1 & 1 & 3 \\ 4 & 2 & 2 & 2 & 3 \\ 3 & 5 & 3 & 3 & 2 \end{bmatrix}$$

The GLSZM then becomes:

$$P = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 \end{bmatrix}$$

1.10.4 Neighbouring Gray Tone Difference Matrix (NGTDM) Features

A Neighbouring Gray Tone Difference Matrix quantifies the difference between a gray value and the average gray value of its neighbours within distance δ . The sum of absolute differences for gray level i is stored in the matrix. Let Xgl be a set of segmented voxels and $xgl(jx, jy, jz) \in Xgl$ be the gray level of a voxel at position (jx, jy, jz) , then the average gray level of the neighbourhood is:

$$A^- i = A^-(jx, jy, jz) = \frac{1}{W} \sum_{\delta kx=-\delta}^{\delta} \sum_{\delta ky=-\delta}^{\delta} \sum_{\delta kz=-\delta}^{\delta} xgl(jx + kx, jy + ky, jz + kz), \text{ where } (kx, ky, kz) \neq (0, 0, 0) \text{ and } xgl(jx + kx, jy + ky, jz + kz) \in Xgl$$

Here, W is the number of voxels in the neighbourhood that are also in Xgl .

As a two dimensional example, let the following matrix I represent a 4x4 image, having 5 discrete grey levels, but no voxels with gray level 4:

$$I = \begin{bmatrix} 1 & 2 & 5 & 2 \\ 3 & 5 & 1 & 3 \\ 1 & 3 & 5 & 5 \\ 3 & 1 & 1 & 1 \end{bmatrix}$$

The following NGTDM can be obtained:

i	n_i	pi	s_i
1	6	0.375	13.35
2	2	0.125	2.00
3	4	0.25	2.63
4	0	0.00	0.00
5	4	0.25	10.075

1.10.5 Gray Level Dependence Matrix (GLDM)

A Gray Level Dependence Matrix (GLDM) quantifies gray level dependencies in an image. A gray level dependency is defined as a the number of connected voxels within distance δ that are dependent on the center voxel. A neighbouring voxel with gray level j is considered dependent on center voxel with gray level i if $|i - j| \leq \alpha$. In a gray level dependence matrix $P(i, j)$ the (i, j) th element describes the number of times a voxel with gray level i with j dependent voxels in its neighbourhood appears in image.

As a two dimensional example, consider the following 5x5 image, with 5 discrete gray levels:

$$I = \begin{bmatrix} 5 & 2 & 5 & 4 & 4 \\ 3 & 3 & 3 & 1 & 3 \\ 2 & 1 & 1 & 1 & 3 \\ 4 & 2 & 2 & 2 & 3 \\ 3 & 5 & 3 & 3 & 2 \end{bmatrix}$$

For $\alpha = 0$ and $\delta = 1$, the GLDM then becomes:

$$P = \begin{bmatrix} 0 & 1 & 2 & 1 \\ 1 & 2 & 3 & 0 \\ 1 & 4 & 4 & 0 \\ 1 & 2 & 0 & 0 \\ 3 & 0 & 0 & 0 \end{bmatrix}$$

CHAPTER 2

2

Acknowledgements

We thank Jean Coenen, from MAASTRO clinic, for providing us with the figures included in this paper.

Keywords

Radiomics, machine learning, oncology, precision medicine

A review on radiomics and the future of theragnostics for patient selection in precision medicine

Simon A. Keek, MSc¹, Ralph T.H. Leijenaar, MSc¹, Arthur Jochems*, PhD¹, Henry C. Woodruff*, PhD^{1,2}

* Equal contribution as last author.

1 The D-Lab: Decision Support for Precision Medicine GROW - School for Oncology and Developmental Biology & MCCC Maastricht University Medical Centre+ Maastricht, The Netherlands

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

2.1 Abstract

The growing complexity and volume of clinical data and the associated decision-making processes in oncology promote the advent of precision medicine. Precision (or personalized) medicine describes preventive and/or treatment procedures that take individual patient variability into account when proscribing treatment, and has been hindered in the past by the strict requirements of accurate, robust, repeatable, and preferably non-invasive biomarkers to stratify both the patient and the disease. In oncology, tumour subtypes are traditionally measured through repeated invasive biopsies, which are taxing for the patient and are cost and labour intensive. Quantitative analysis of routine clinical imaging provides an opportunity to capture tumour heterogeneity non-invasively, cost-effectively, and on large scale. In current clinical practice radiological images are qualitatively analysed by expert radiologists whose interpretation is known to suffer from inter- and intra-operator variability. Radiomics, the high-throughput mining of image features from medical images [1], provides a quantitative and robust method to assess tumour heterogeneity, and radiomics-based signatures provide a powerful tool for precision medicine in cancer treatment. This study aims to provide an overview of the current state of radiomics as a precision medicine decision support tool. We first provide an overview of the requirements and challenges radiomics currently faces in being incorporated as a tool for precision medicine, followed by an outline of radiomics' current applications in the treatment of various types of cancer. We finish with a discussion of possible future advances that can further develop radiomics as a precision medicine tool.

2.2 Introduction

2.2.1 Background

Technological advances have led to an abundance of novel diagnostic techniques and imaging modalities available to oncology. [2] Additional complexity is added by genetic [3] and micro-environmental [4] heterogeneity of tumours and between patients. [5] Due to the large volumes and complexity of modern data [6], new methods to facilitate clinical decision-making are required.

Precision (or personalized) medicine describes preventive and treatment procedures that take into account an individual patient's characteristics together with their specific disease(s) [7]. A common approach to precision medicine is data-mining, i.e. discovering patterns in large databases of diversified cohorts using powerful computational tools such as machine learning. Patterns can be discovered within the variability of patient populations that allow for the stratification of patient groups and the identification of the ideal treatment for the individual patient [8], thus improving patient outcome. [9-11] However, this requires large databases of patients in order to cover as much of the variations within a population as possible.

An important source of large-scale data that could be used are radiological images derived during routine oncological examinations. Tumours exhibit phenotypical differences which can be visualized through routine medical imaging [12], which in turn allows for visualization of the entire tumour volume or sub-regions on a macroscopic level, at baseline and longitudinally. However, imaging in a clinical setting is primarily used qualitatively, and clinical decision-making is based on visual assessments of the tumour by radiologists. Radiomics offers a quantitative alternative to assess tumour heterogeneity quantitatively. Radiomics is an advanced image feature analysis methodology, which formats standard clinical images from computed tomography (CT), medical resonance imaging (MRI), and/or positron emission tomography (PET) into a multidimensional source for data mining. [1] A large number of image features are extracted from imaging data using various mathematical algorithms. These features, together with gold standard information, are used by machine learning algorithms, computational methods that "learn" correlations from data, creating models that automate and improve classification of tumour phenotype and genomic profile [13-15] as imaging biomarkers.

Radiomics-based imaging biomarkers have shown to outperform common prognostic models based on clinical parameters such as TNM [13]. However, radiomics does not intend to replace current clinical decision-making, but rather aims to provide a supplement to current measures such as clinical, treatment and genomic data, all incorporated into a

decision support system. [16] To do so, a robust, repeatable, and cost-effective method to clinically implement radiomics is required.

2.2.2 Radiomics workflow

A typical radiomics analysis starts with data selection: choosing the image to analyse, the imaging protocol to use, and the correlated outcome. The image typically contains the primary tumour volume, which is analysed and linked a certain outcome, such as tumour type, overall survival, or tumour recurrence. Proper data selection is important to create useful models, as it needs to be reproducible and applicable across sizeable cohorts. Large heterogeneous datasets are required to provide enough data to validate the model on different sub-samplings of the data (cross-validation). [17] In addition, the quality of the data is dependent on the image acquisition protocols used in clinical centres, which can often vary extensively, as well as the imaged site, scanner properties, reconstruction methods and motion artefacts. [18, 19] Guidelines for image acquisition and standardized protocols are therefore beneficial for producing large, high-quality datasets. [20] In the case of non-standardized imaging protocols, sharing of imaging protocols should be encouraged.

After image acquisition and volume reconstruction, a region of interest (ROI) is defined, usually, but not necessarily, through slice-by-slice delineation of the tumour in the case of 3D images. This is a labour- intensive process, and the variance caused by inter- and intra-operator variability is an issue. [21, 22] A (semi-) automatic segmentation method to reduce workload and uncertainty caused by human error is therefore preferred. Besides operator variability, image segmentation, protocol standardization, slice interval, reconstruction method, time-point and respiratory motion have all been found to have effects on feature reproducibility. [23-35] Methods to improve reproducibility include multiple segmentations by different clinicians and phantom studies to determine the effects caused by different scanners.

Since the values of extracted features (mostly mathematical formulas using pixel/voxel intensities as input) depend on image reconstruction and pre-processing methods, proper reporting of methods such as filtering techniques, intensity discretisation and voxel resampling is critical for interoperability of the radiomics features. Many of the extracted features are noise driven and need to be removed to improve model performance. The same applies to features that are highly correlated with other features or existing clinical parameters, as they do not provide any meaningful addition to the model. Test-retest studies which repeat the imaging processes after a short period of time are indispensable, as they measure the amount of variation inherent in the measurements. Stability and correlation tests can be used to make a selection based on the most robust, repeatable and non-redundant features. [36, 37]

The extracted features are fed into machine learning methods together with clinical outcomes or pathology results to construct classification, predictive, or prognostic models. Prognostic models aim to predict a certain outcome regardless of therapy, while predictive models provide information about the effects of a certain therapeutic intervention. However, the number of extracted features is often larger than the number of patients included in a cohort, which risks overfitting the model. The best solution to prevent overfitting is to increase the number of samples used to train the model. While clinical data is abundant compared to research trial data, sharing between institutes has proven to be difficult due to various ethical, political and administrative issues. [38] An alternative to large datasets is to reduce the number of features to a subset of the most relevant features. Various filtering-based techniques for feature selection can be used, such as the univariate Fisher score and Gini index tests, or multivariate algorithms such as mutual information or Conditional infomax feature extraction [39], which identify and select a sub-set of features based on predictive power. Valid predictive modelling requires separate independent datasets for training and validation.

Various different machine learning models are available, such as neural networks, decision trees, support vector machines and multiple regression techniques. The modelling procedure has been shown to affect performance of prediction models based on radiomics features. [39] Common measures of predictive performance of models are discrimination and calibration. [40] Discrimination is a measure of the model to assign a higher risk-prediction to patients positive to a certain outcome, compared to patients without the outcome, which can be quantified using the sensitivity, specificity, or through the area under the curve (AUC) of the receiver operating characteristic. The AUC is equal to the probability that a positive event is correctly labelled as a positive event, and is given in the range of 0 to 1. Alternatively, the Concordance Index (CI), a measure of goodness of fit for classification models with binary outcome ranging from 0 to 1, can be used. Both AUC and CI show a perfect predictive performance at 1, while at 0.5 the predictive performance is completely random. Calibration is an internal measure of the model's agreement between observed outcomes and predicted outcomes. The calibration is usually assessed through a calibration slope, where different resamplings of observed outcomes are plotted against predicted outcomes. If 100% agreement between these two is found at multiple samplings, then the calibration slope will be 1. Finally, a log-rank test is usually used to test the significance of the difference between survival curves of two patient groups. This is used when separating patients in low- and high-risk groups based on radiomics features.

These measures of predictive performance are used to internally and externally validate the model. Internal validation is necessary to estimate and reduce the optimism in model performance, which is the degree a trained model performs worse when making predictions on new data. Internal validation uses the data used to train the model, and

can be performed through methods such as bootstrap analysis or cross-validation. [41] External validation uses an independent, external dataset to validate the accuracy of the predictive model, and to assess the generalizability of a predictive model. [41] Figure 1 shows an overview of the steps involved to train and validate a predictive model.

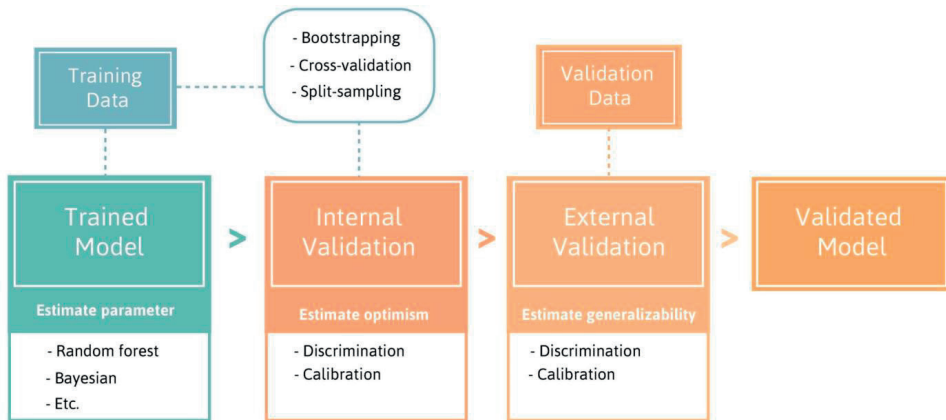


Figure 1. Overview of the steps involved to train and validate a predictive model.

Effective and transparent radiomics studies require rigorous compliance with several guidelines, including effective validation. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Initiative is a set of guidelines made for studies creating and/or validating prediction models. [42] There are guidelines for the source and specific information of data, the type of predictive model, procedures for building the model and the method for internal validation, and measurements of model performance. Whereas the TRIPOD initiative covers prediction models in general, the Radiomics Quality Score (RQS) [43] is being developed specifically for radiomics studies. The RQS assesses the quality of a study using a checklist and reports compliance as a percentage. Some of the guidelines include robust segmentations, test-retest stability of the determined features, the standardization or thorough description of imaging protocols used, valid feature selection and internal/external validation. [44] An overview of the different steps and the RQS criteria is shown in figure 2.

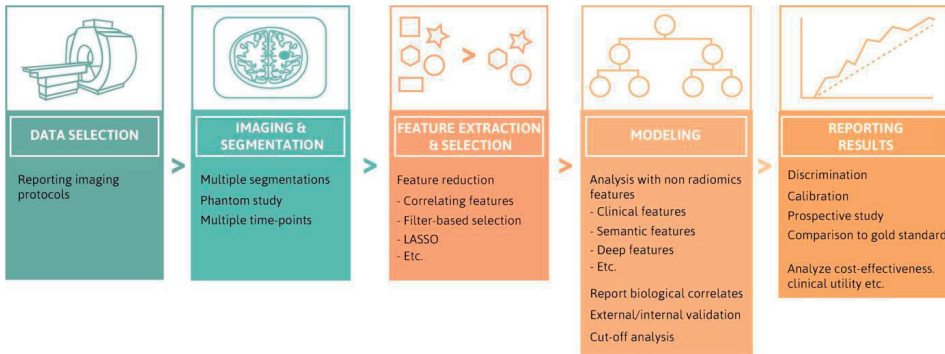


Figure 2. Overview of steps of a Radiomics analysis (top) with corresponding RQS score criteria for each step (bottom).

The aim of these guidelines is to provide key details of model development and validation, which in turn allows for better reproducibility and critical appraisal of predictive models. For future and past studies, authors should check the RQS score and TRIPOD initiative to determine the quality of their methodology and allow the field of radiomics to mature. The ultimate objective of precision medicine is to link the tumour phenotype to a certain clinical endpoint, with the goal of improving clinical decision making. Therefore, the next section will describe the use of radiomics in various studies and their efficacy in determining clinical endpoints.

2.2.3 Role in precision medicine

Aerts et al. (2014) performed a radiomics analysis on a large CT dataset (N=1019) of lung- and H&N- cancer patients. Using a feature selection algorithm to reduce the number of features from 440 to a prognostic signature of 4 features, they found that a model built using this signature was significantly more prognostic of overall survival (OS) than a measure of tumour volume, and combining the radiomics signature with tumour volume also provided a better prognostic ability. The model was validated on different patient groups and cancer types. [13] The radiomics signature showed slightly higher prognostic performance when validated in an external lung dataset than TNM or tumour volume (CI of 0.65 vs. 0.63 and 0.60 respectively). For two external H&N cohorts, the signature showed higher performance compared to volume or TNM in one case (CI of 0.69 vs. 0.65 and 0.66 respectively), and similar performance in the other (CI = 0.69 vs. 0.68 and 0.69 respectively). This radiomics signature was also externally validated in a study by Leijenaar et al. (2015) on a large set of oropharyngeal squamous cell carcinoma patients (N = 542). [45] The signature showed good discrimination and calibration (CI = 0.628 and calibration slope of 0.855), and after the population were split in two groups using the median value

of the signature score, significant differences in OS (long rank p-value = $2e-5$) could be observed.

Furthermore, CT radiomics features have been shown to be prognostic of distant metastasis and 12-month survival in glioblastoma [46], and pathological response to treatment [47], local recurrence [48, 49], histology subtype [50, 51], OS [36, 50] and prediction of radiation induced pneumonitis [52, 53] for lung cancer. In head- and neck squamous cell carcinomas, radiomics has proven to improve the prediction overall and progression free survival, and determining HPV status. [54]

Delta-radiomics is an alternative analysis which measures the change of radiomics features longitudinally. Certain features have been proven to change during treatment, indicating that this may be an additional source of information [55]. Delta-radiomics on CT has shown to be prognostic in non-small cell lung cancer (NSCLC) for OS, local recurrence and distant metastasis. [56] For H&N cancer patients, delta-radiomics features have proven to be a predictive and prognostic biomarker, as well provide additional information of the presence of HPV for patient stratification. [36, 54, 57, 58]

An additional source of routine medical images for radiomics analysis are cone-beam CT (CBCT) images, often used in radiotherapy for daily positioning before treatment. Van Timmeren et al (2017) have used CBCT data of NSCLC patients to validate a previously constructed CT radiomics signature. The signature was found to be predictive of OS in three different independent CBCT datasets (CI = 0.59-0.66), indicating CBCT could potentially be a useful source of information for radiomics analysis. [59]

FDG-PET-based quantitative image analysis shows promise in improving prognosis in pancreatic cancer. A study by Cui et al. used quantitative imaging features to predict OS, and showed better prognostic compared to the use of conventional prognostic variables of tumour volume and maximum SUV (CI of 0.66 vs 0.48-0.64). [60] FDG-PET-based radiomics features correlate to mortality, local failure and distant metastasis for pancreatic cancer [61], and have also shown to be predictive in oesophageal cancer [62], tumour response in cervical cancer [63, 64] and local control [65] and OS [63] in H&N cancer.

MRI-based radiomics has shown promise in prostate cancer: a study by Shoshana et al. (2016) used T2-weighted MRI radiomics features to differentiate between peripheral and transition zone prostate tumours (AUC=0.61-0.71), in a patient dataset from three different institutions. [66] Furthermore, a study by Vallières et al (2015) use a combination of FDG-PET and MRI texture features to predict the lung metastasis in soft-tissue sarcomas. They found that a multivariable model was highly predictive of lung metastasis in soft-tissue sarcomas (AUC=0.98), validated through bootstrapping procedures. However, the study lacked external validation for a valid conclusion. [67] In the context of glioblastoma,

several studies using MRI data have shown that a radiomics model may accurately detail the molecular subtype of the tumour [68-70], OS [69-71] and predicting short versus long-term survival. [72] Finally, for an imaging method outside of radiology, Zhang et al. (2016) proposed a radiomics approach to ultrasound elastography, to use the density of tumour tissue for classification as benign or malignant. A signature of seven features, out of a total of 364 extracted features, was able to accurately (AUC= 0.92) discriminate between benign and malignant tumour tissue. [73]

To reduce inter- and intra-observer delineation variability and to workload, a number of (semi-) automatic segmentation methods have been proposed and tested in radiomics studies in recent years. Several studies have shown that (semi-) automatic segmentation methods reduce inter-observer delineation variability compared to manual segmentation of lung lesions. [74-77] For example, a study by Parmar et al. (2014) compared the robustness of 56 radiomics features derived with manual segmentation of tumour volume by five experts to a semi-automatic method performed two times by three experts, and showed that semi-automatically derived features have significantly higher reproducibility compared to manually derived features. [77] Full automatic segmentation of tumours is also a possibility, as shown by Li et al (2017). This study used radiomics features in a random forest model to classify tumour tissue on a voxel level. The algorithm was trained and tested on publically available datasets, and showed promising accuracy in classifying tumour tissue, necrosis, normal tissue and oedema. [78]

Semantic features, unique qualitative characteristics that provide information about the prognosis and (sub) type of lesions, are an alternative method to describe tumour (sub) type, and could be useful in improving prediction of certain endpoints. Some examples of semantic features are the presence of cavitation or calcification in the tumour, or features describing the roundness or spiculation of the tumours. In a study on NSCLC, Yip et al. (2017) studied 9 semantic features, consisting of 3 binary features and 6 categorical classifiers, and 57 radiomics features describing NSCLC cancer phenotypes. To study the correlation between features they used Spearman's Rank-Order Correlation, which is a measure of the strength and direction of association between two variables. Spearman's rank ranges from -1 to 1, with both extremes signifying perfect correlation between two variables. The study found significant association between radiomics features and binary semantic features (AUC = 0.56-0.76), but no or weak correlation was found between classification semantic and radiomics features (Spearman's correlation = 0.002-0.65). This indicates that radiomics and semantic features have complementary but distinct roles in outcome prediction, as they have both been proven to be able to significantly improve prediction outcomes. [79]

Lastly, deep learning tools, such as convolutional neural networks (CNNs), could be a method to augment radiomics analysis. Deep learning algorithms are able to learn features from imaging data without much manual input, provided that a large amount of data is available. Deep learning has been successfully implemented in a number of different studies using medical imaging data. [80, 81] Orlando et al. (2017) used a combination CNN learned and hand-crafted discriminative features to detect red lesions (a collective term for micro aneurysms and haemorrhages), one of the earliest signs in diabetic retinopathy. The combination of features was used in a random forest classifier to discriminate between true and false red lesion candidates, and compared against either set of features separately. The combination achieved higher AUC values compared to the separate feature prediction models (AUC of 0.89 vs 0.79/0.73 for CNN and handcrafted features respectively). Recently published articles have already shown that radiomics analysis may benefit from incorporating deep learning methods. [82-85] For example, Lao et al. (2017) used a combination of hand-crafted and deep radiomics features to predict OS for Glioblastoma Multiforme patients on MRI images. After feature selection, a radiomics signature was created, using exclusively deep learned features, that was able to accurately predict OS in an independent validation dataset (AUC = 0.71). Deep learning augmented radiomics analysis has also been reported to be effective in assessing treatment response in bladder cancer [86], where conversely a signature built solely on hand-crafted features was found to have better prognostic performance. These results indicate that deep learning will have an increasingly important role in predictive modelling [87], and have a complementary role with hand-crafted features in a radiomics analysis framework.

2.3 Discussion

Radiomics has been shown to be suitable for classification, prediction and prognosis of various clinical endpoints and tumour types. Many studies show a clear improvement over conventional measures predicting clinical endpoints, although variation in feature stability due to different scanners, imaging protocols and tumour motion still leaves a lot of room for improvement. [13, 60] The segmentation of tumours also proves to be a small but persistent obstacle, as it is a labour- and time-intensive process and is heavily influenced by inter- and intra-segmentation variation [21, 22]. However, numerous studies have reported methods to allow for a more automatic approach to segmentation [74-78], which in turn could lead to a more robust radiomics analysis.

Combining radiomics features with deep learning features or semantic features may be able to further improve prognostic performance. Several studies have proven the effectiveness of using these features independently in predictive modelling. [80-87] In studies comparing the prognostic performance of these features to hand-crafted

radiomics features, results were found to be mixed, indicating these methods may have distinct and complementary roles in improving prognosis.

A larger hurdle for radiomics is the transition to clinical implementation. While routine delineation is already in place in radiotherapy settings, a clinical platform to easily perform radiomics analysis during routine check-up/treatment is not. The main challenge of precision treatment is to correctly integrate various sources of data quantitatively and subsequently use this data to provide specific clinical predictions that accurately and robustly estimate outcomes as a function of the possible decisions.

Numerous methods, besides radiomics, are currently in use that make use of novel biomarkers, as well as conventional clinical factors. However, many of these methods lack external validation of their legitimacy, reproducibility, or clinical validity. [88] Radiomics offers a solution that integrates multiple measures into one prediction of outcome, with the added benefit of automation, which could save time and money in a clinical environment.

While many radiomics studies include external validation steps, sharing of clinical data is still an issue [89]. The difficulty in sharing data may be overcome through a centralised database, or conversely through decentralised distributed learning platforms. [90] To facilitate a centralised database, data has to be made available in accordance with the FAIR principles: Findability, Accessibility, Interoperability and Reusability. [89] An example of an effort to increase data shareability is through the development of ontologies to describe radiomics features. [29]

The distributed learning method instead aims to solve the problem of sparse data by avoiding the numerous ethical, legal and administrative issues involved in sharing data between institutes. Instead of the images being collected from numerous institutes in one central location, the model is sent and trained on site without any data leaving a particular institute. The trained models are then collected, analysed and integrated into a single model. Several proof-of-concept studies have proven that a distributed learning approach is feasible using clinical parameters [90-92], and the next step would be to integrate radiomics features, by sending a platform to extract radiomics features on-site in conjunction with the untrained predictive model. This way, a distributed learning method could provide the necessary volume and variety in data to achieve a machine driven approaches to medicine.

2.4 Conclusion

In conclusion, radiomics provides a novel non-invasive method of assessing tumour subtype, using the mostly untapped source of data of routine clinical images. The technique is often hampered by studies with small sample sizes and lack of external validation. In addition, variability in features caused by differences in imaging modality, protocols and respiratory motion, and a lack of interoperability, may decrease the generalizability of the created radiomics models. In the future, research should be informed by guidelines such as RQS and TRIPOD, which improve the validity of radiomics as a clinically accepted field. The clinical value of the technique has already been described for a wide range of tumours and a number of different clinical outcomes. The added fact that the analysis can be performed in an automated fashion makes the technique attractive for clinical implementation to reduce workload. Performing studies on different tumour sites/types in future research may prove the generalizability of the method, and consequently lead to radiomics becoming a standard method clinically. In the future, larger volumes of data will be available for use in Radiomics studies by means of centralised, publically accessible datasets and distributed learning. Combining radiomics with other parameters will lead to high quality decision support systems, and deep learning and semantic feature approaches may be combined with radiomics analyses to increase predictive accuracies of these models even further. Radiomics has a way ahead before full implementation in clinic is a reality, but may prove to be invaluable in realizing precision medicine in cancer treatment.

2.5 References

1. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-6.
2. Burstein HJ, Krilov L, Aragon-Ching JB, Baxter NN, Chiorean EG, Chow WA, et al. Clinical Cancer Advances 2017: Annual Report on Progress Against Cancer From the American Society of Clinical Oncology. *J Clin Oncol*. 2017;35(12):1341-67.
3. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*. 2012;366(10):883-92.
4. Milosevic MF, Fyles AW, Wong R, Pintilie M, Kavanagh MC, Levin W, et al. Interstitial fluid pressure in cervical carcinoma: within tumor heterogeneity, and relation to oxygen tension. *Cancer*. 1998;82(12):2418-26.
5. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346-52.
6. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, et al. Rapid-learning system for cancer care. *J Clin Oncol*. 2010;28(27):4268-74.
7. Garraway LA, Verweij J, Ballman KV. Precision oncology: an overview. *J Clin Oncol*. 2013;31(15):1803-5.
8. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015;372(9):793-5.
9. Aerts HJ, Bussink J, Oyen WJ, van Elmpt W, Folgering AM, Emans D, et al. Identification of residual metabolic-active areas within NSCLC tumours using a pre-radiotherapy FDG-PET-CT scan: a prospective validation. *Lung Cancer*. 2012;75(1):73-6.
10. 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(3):386-92.
11. Suit H, Skates S, Taghian A, Okunieff P, Efrid JT. Clinical implications of heterogeneity of tumor response to radiation therapy. *Radiother Oncol*. 1992;25(4):251-60.
12. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*. 2016;278(2):563-77.
13. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
14. Grossmann P, Stringfield O, El-Hachem N, Bui MM, Rios Velazquez E, Parmar C, et al. Defining the biological basis of radiomic phenotypes in lung cancer. *Elife*. 2017;6.
15. Panth KM, Leijenaar RTH, Carvalho S, Lieuwes NG, Yaromina A, Dubois L, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo

- preclinical experiment with doxycycline inducible GADD34 tumor cells. *Radiotherapy and Oncology*. 2015;116(3):462-6.
16. Lambin P, Zindler J, Vanneste BG, De Voorde LV, Eekers D, Compter I, et al. Decision support systems for personalized and participative radiation oncology. *Adv Drug Deliv Rev*. 2017;109:131-53.
 17. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30(9):1234-48.
 18. Nehmeh SA, Erdi YE. Respiratory motion in positron emission tomography/computed tomography: a review. *Semin Nucl Med*. 2008;38(3):167-76.
 19. Sonke JJ, Belderbos J. Adaptive radiotherapy for lung cancer. *Semin Radiat Oncol*. 2010;20(2):94-106.
 20. de Jong EEC, van Elmpt W, Hoekstra OS, Groen HJM, Smit EF, Boellaard R, et al. Quality assessment of positron emission tomography scans: recommendations for future multicentre trials. *Acta Oncol*. 2017;56(11):1459-64.
 21. Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. *J Clin Oncol*. 2003;21(13):2574-82.
 22. Schwartz LH, Mazumdar M, Brown W, Smith A, Panicek DM. Variability in response assessment in solid tumors: effect of number of lesions chosen for measurement. *Clin Cancer Res*. 2003;9(12):4318-23.
 23. Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, et al. Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. *Int J Radiat Oncol Biol Phys*. 2014;90(4):834-42.
 24. Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol*. 2013;52(7):1391-7.
 25. Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. *J Nucl Med*. 2012;53(5):693-700.
 26. Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep*. 2016;6:23428.
 27. Fave X, Mackin D, Yang J, Zhang J, Fried D, Balter P, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? *Med Phys*. 2015;42(12):6784- 97.
 28. Oliver JA, Budzevich M, Zhang GG, Dilling TJ, Latifi K, Moros EG. Variability of Image Features Computed from Conventional and Respiratory-Gated PET/CT Images of Lung Cancer. *Transl Oncol*. 2015;8(6):524-34.
 29. Kalpathy-Cramer J, Mamomov A, Zhao B, Lu L, Cherezov D, Napel S, et al. Radiomics of Lung Nodules: A Multi-Institutional Study of Robustness and Agreement of Quantitative Imaging Features. *Tomography*. 2016;2(4):430-7.

30. Beichel RR, Smith BJ, Bauer C, Ulrich EJ, Ahmadvand P, Budzevich MM, et al. Multi-site quality and variability analysis of 3D FDG PET segmentations based on phantom and clinical image data. *Med Phys*. 2017;44(2):479-96.
31. Tan Y, Guo P, Mann H, Marley SE, Juanita Scott ML, Schwartz LH, et al. Assessing the effect of CT slice interval on unidimensional, bidimensional and volumetric measurements of solid tumours. *Cancer Imaging*. 2012;12:497-505.
32. Larue R, van Timmeren JE, de Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncol*. 2017;56(11):1544-53.
33. Van Timmeren J, Leijenaar R, van Elmpt W, Wang J, Zhang Z, Dekker A, et al. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? 2016.
34. Larue RT, Defraene G, De Ruyscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br J Radiol*. 2017;90(1070):20160665.
35. Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. *Sci Rep*. 2015;5:11075.
36. Parmar C, Leijenaar RT, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. *Sci Rep*. 2015;5:11044.
37. Larue R, Van De Voorde L, van Timmeren JE, Leijenaar RTH, Berbee M, Sosef MN, et al. 4DCT imaging to assess radiomics feature stability: An investigation for thoracic cancers. *Radiother Oncol*. 2017;125(1):147-53.
38. Doshi P, Jefferson T, Del Mar C. The imperative to share clinical study reports: recommendations from the Tamiflu experience. *PLoS Med*. 2012;9(4):e1001201.
39. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci Rep*. 2015;5:13087.
40. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-38.
41. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol*. 2003;56(5):441-7.
42. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med*. 2015;13:1.
43. Radiomics.world [Webpage]. Maastriclinic; c2014-2017; [updated 2017; cited 2017 25 August]. Available from: <http://www.radiomics.world/>.

44. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, Jong EECd, Timmeren Jv, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* (In Press). 2017.
45. Leijenaar RT, Carvalho S, Hoebbers FJ, Aerts HJ, van Elmpt WJ, Huang SH, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncol*. 2015;54(9):1423-9.
46. Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RT, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol*. 2015;114(3):345-50.
47. Coroller TP, Agrawal V, Narayan V, Hou Y, Grossmann P, Lee SW, et al. Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiother Oncol*. 2016;119(3):480-6.
48. Mattonen SA, Palma DA, Johnson C, Louie AV, Landis M, Rodrigues G, et al. Detection of Local Cancer Recurrence After Stereotactic Ablative Radiation Therapy for Lung Cancer: Physician Performance Versus Radiomic Assessment. *Int J Radiat Oncol Biol Phys*. 2016;94(5):1121-8.
49. Huynh E, Coroller TP, Narayan V, Agrawal V, Romano J, Franco I, et al. Associations of Radiomic Data Extracted from Static and Respiratory-Gated CT Scans with Disease Recurrence in Lung Cancer Patients Treated with SBRT. *PLoS One*. 2017;12(1):e0169172.
50. Song J, Liu Z, Zhong W, Huang Y, Ma Z, Dong D, et al. Non-small cell lung cancer: quantitative phenotypic analysis of CT images as a potential marker of prognosis. *Sci Rep*. 2016;6:38282.
51. Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology. *Front Oncol*. 2016;6:71.
52. Cunliffe A, Armato SG, 3rd, Castillo R, Pham N, Guerrero T, Al-Hallaq HA. Lung texture in serial thoracic computed tomography scans: correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development. *Int J Radiat Oncol Biol Phys*. 2015;91(5):1048-56.
53. Anthony GJ, Cunliffe A, Castillo R, Pham N, Guerrero T, Armato SG, 3rd, et al. Incorporation of pre-therapy 18 F-FDG uptake data with CT texture features into a radiomics model for radiation pneumonitis diagnosis. *Med Phys*. 2017;44(7):3686-94.
54. Ou D, Blanchard P, Rosellini S, Levy A, Nguyen F, Leijenaar RTH, et al. Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status. *Oral Oncol*. 2017;71:150-5.
55. van Timmeren JE, Leijenaar RTH, van Elmpt W, Reymen B, Lambin P. Feature selection methodology for longitudinal cone-beam CT radiomics. *Acta Oncol*. 2017;56(11):1537-43.
56. Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. *Sci Rep*. 2017;7(1):588.
57. Leijenaar R, Nesteruk M, Feliciani G, Hoebbers F, Van Timmeren J, Van Elmpt W, et al. EP-1608: Deriving HPV status from standard CT imaging: a radiomic approach with independent validation. *Radiotherapy and Oncology*. 2017;123:S868-S9.

58. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJ. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. *Front Oncol.* 2015;5:272.
59. van Timmeren JE, Leijenaar RTH, van Elmpt W, Reymen B, Oberije C, Monshouwer R, et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. *Radiotherapy and Oncology.* 2017;123(3):363-9.
60. Cui Y, Song J, Pollom E, Alagappan M, Shirato H, Chang DT, et al. Quantitative Analysis of (18) F- Fluorodeoxyglucose Positron Emission Tomography Identifies Novel Prognostic Imaging Biomarkers in Locally Advanced Pancreatic Cancer Patients Treated With Stereotactic Body Radiation Therapy. *Int J Radiat Oncol Biol Phys.* 2016;96(1):102-9.
61. Folkert MR, Setton J, Apte AP, Grkovski M, Young RJ, Schoder H, et al. Predictive modeling of outcomes following definitive chemoradiotherapy for oropharyngeal cancer based on FDG-PET image characteristics. *Phys Med Biol.* 2017;62(13):5327-43.
62. Desbordes P, Ruan S, Modzelewski R, Pineau P, Vauclin S, Gouel P, et al. Predictive value of initial FDG-PET features for treatment response and survival in esophageal cancer patients treated with chemo-radiation therapy using a random forest classifier. *PLoS One.* 2017;12(3):e0173208.
63. El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.* 2009;42(6):1162-71.
64. Yang F, Thomas MA, Dehdashti F, Grigsby PW. Temporal analysis of intratumoral metabolic heterogeneity characterized by textural features in cervical cancer. *Eur J Nucl Med Mol Imaging.* 2013;40(5):716-27.
65. Bogowicz M, Leijenaar RTH, Tanadini-Lang S, Riesterer O, Pruschy M, Studer G, et al. Post-radiochemotherapy PET radiomics in head and neck cancer - The influence of radiomics implementation on the reproducibility of local control tumor models. *Radiother Oncol.* 2017.
66. Ginsburg SB, Algohary A, Pahwa S, Gulani V, Ponsky L, Aronen HJ, et al. Radiomic features for prostate cancer detection on MRI differ between the transition and peripheral zones: Preliminary findings from a multi-institutional study. *J Magn Reson Imaging.* 2017;46(1):184-93.
67. Vallieres M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol.* 2015;60(14):5471-96.
68. Gevaert O, Mitchell LA, Achrol AS, Xu J, Echegaray S, Steinberg GK, et al. Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. *Radiology.* 2014;273(1):168-74.
69. Zhou H, Vallieres M, Bai HX, Su C, Tang H, Oldridge D, et al. MRI features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro Oncol.* 2017;19(6):862-70.
70. Yang D, Rao G, Martinez J, Veeraraghavan A, Rao A. Evaluation of tumor-derived MRI-texture features for discrimination of molecular subtypes and prediction of 12-month survival status in glioblastoma. *Med Phys.* 2015;42(11):6725-35.

71. McGarry SD, Hurrell SL, Kaczmarowski AL, Cochran EJ, Connelly J, Rand SD, et al. Magnetic Resonance Imaging-Based Radiomic Profiles Predict Patient Prognosis in Newly Diagnosed Glioblastoma Before Therapy. *Tomography*. 2016;2(3):223-8.
72. Prasanna P, Patel J, Partovi S, Madabhushi A, Tiwari P. Radiomic features from the peritumoral brain parenchyma on treatment-naive multi-parametric MR imaging predict long versus short-term survival in glioblastoma multiforme: Preliminary findings. *Eur Radiol*. 2016.
73. Zhang Q, Xiao Y, Suo J, Shi J, Yu J, Guo Y, et al. Sonoelastomics for Breast Tumor Classification: A Radiomics Approach with Clustering-Based Feature Selection on Sonoelastography. *Ultrasound Med Biol*. 2017;43(5):1058-69.
74. Rios Velazquez E, Aerts HJ, Gu Y, Goldgof DB, De Ruyscher D, Dekker A, et al. A semiautomatic CT-based ensemble segmentation of lung tumors: comparison with oncologists' delineations and with the surgical specimen. *Radiother Oncol*. 2012;105(2):167-73.
75. Heye T, Merkle EM, Reiner CS, Davenport MS, Horvath JJ, Feuerlein S, et al. 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*. 2013;266(3):812-21.
76. Velazquez ER, Parmar C, Jermoumi M, Mak RH, van Baardwijk A, Fennessy FM, et al. Volumetric CT-based segmentation of NSCLC using 3D-Slicer. *Sci Rep*. 2013;3:3529.
77. Parmar C, Rios Velazquez E, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One*. 2014;9(7):e102107.
78. Li Q, Bai H, Chen Y, Sun Q, Liu L, Zhou S, et al. A Fully-Automatic Multiparametric Radiomics Model: Towards Reproducible and Prognostic Imaging Signature for Prediction of Overall Survival in Glioblastoma Multiforme. *Sci Rep*. 2017;7(1):14331.
79. Yip SSF, Liu Y, Parmar C, Li Q, Liu S, Qu F, et al. Associations between radiologist-defined semantic and automatically computed radiomic features in non-small cell lung cancer. *Sci Rep*. 2017;7(1):3519.
80. Zheng Y, Liu D, Georgescu B, Nguyen H, Comaniciu D. 3D Deep Learning for Efficient and Robust Landmark Detection in Volumetric Data. In: Navab N, Hornegger J, Wells WM, Frangi A, editors. *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I*. Cham: Springer International Publishing; 2015. p. 565-72.
81. Ravishankar H, Sudhakar P, Venkataramani R, Thiruvankadam S, Annangi P, Babu N, et al. Understanding the Mechanisms of Deep Transfer Learning for Medical Images. In: Carneiro G, Mateus D, Peter L, Bradley A, Tavares JMRS, Belagiannis V, et al., editors. *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings*. Cham: Springer International Publishing; 2016. p. 188-96.
82. Li Z, Wang Y, Yu J, Guo Y, Cao W. Deep Learning based Radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Sci Rep*. 2017;7(1):5467.

83. Kumar D, Chung AG, Shaifee MJ, Khalvati F, Haider MA, Wong A. Discovery Radiomics for Pathologically-Proven Computed Tomography Lung Cancer Prediction. In: Karray F, Campilho A, Cheriet F, editors. *Image Analysis and Recognition: 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5–7, 2017, Proceedings*. Cham: Springer International Publishing; 2017. p. 54-62.
84. Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, et al. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Sci Rep*. 2017;7(1):10353.
85. Jochems A, Hoebbers F, De Ruyscher D, Leijenaar R, Walsh S, O'Sullivan B, et al. EP-1605: Deep learning of radiomics features for survival prediction in NSCLC and Head and Neck carcinoma. *Radiotherapy and Oncology*.123:S866.
86. Cha KH, Hadjiiski L, Chan HP, Weizer AZ, Alva A, Cohan RH, et al. Bladder Cancer Treatment Response Assessment in CT using Radiomics with Deep-Learning. *Sci Rep*. 2017;7(1):8738.
87. Orlando JC, Prokofyeva E, del Fresno M, Blaschko MB. Learning to Detect Red Lesions in Fundus Photographs: An Ensemble Approach based on Deep Learning. *Medical Image Analysis*. 2017.
88. Vickers AJ. Prediction models: revolutionary in principle, but do they do more good than harm? *J Clin Oncol*. 2011;29(22):2951-2.
89. Lustberg T, van Soest J, Jochems A, Deist T, van Wijk Y, Walsh S, et al. Big Data in radiation therapy: challenges and opportunities. *Br J Radiol*. 2017;90(1069):20160689.
90. Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital - A real life proof of concept. *Radiother Oncol*. 2016;121(3):459-67.
91. Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and validating a survival prediction model for NSCLC patients through distributed learning across three countries. *International Journal of Radiation Oncology*Biophysics*. 2017.
92. Deist TM, Jochems A, van Soest J, Nalbantov G, Oberije C, Walsh S, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clinical and Translational Radiation Oncology*. 2017;4:24-31.

CHAPTER 3

3

Keywords

radiomics, head- and neck cancer, precision medicine, machine learning, survival study

A prospectively validated prognostic model for patients with locally advanced squamous cell carcinoma of the head and neck based on radiomics of computed tomography images

S. A. Keek¹, F. W. R. Wesseling², H. C. Woodruff^{1,3}, J. E. van Timmeren⁴, I. H. Nauta⁵, T. K. Hoffmann⁶, S. Cavalieri⁷, G. Calareso⁸, S. P. Primakov¹, R. T. H. Leijenaar¹⁴, L. Licitra^{7,9}, M. Ravanelli¹⁰, K. Scheckenbach¹¹, T. Poli¹², D. Lanfranco¹², M. R. Vergeer¹³, C. R. Leemans⁵, R. H. Brakenhoff⁵, F. J. P. Hoebers², P. Lambin^{1,3}

1 The D-Lab, Department of Precision Medicine, GROW- School for Oncology, Maastricht University, Maastricht, Universiteitssingel 40, 6229 ER Maastricht, The Netherlands

2 Department of Radiation Oncology (MAASTRO), GROW- School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Postbus 3035, 6202 NA Maastricht, The Netherlands

3 Department of Radiology and Nuclear Medicine, GROW – School for Oncology, Maastricht University Medical Centre+, PO Box 5800, 6202 AZ Maastricht, The Netherlands

4 Department of Radiation Oncology, University Hospital Zürich, University of Zürich, Rämistrasse 100, 8091 Zürich, Switzerland

5 Amsterdam UMC, Vrije Universiteit Amsterdam, Otolaryngology/Head and Neck Surgery, Cancer Center Amsterdam, Postbus 7057, 1007 MB Amsterdam, The Netherlands

6 Dept. of Otorhinolaryngology, Head Neck Surgery, i2SOUL consortium, University of Ulm, Frauensteige 14a (Haus 18), 89075 Ulm, Germany

7 Head and Neck Medical Oncology Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, via Giacomo Venezian, 1 20133 Milano, and University of Milan, Italy

8 Radiology Unit, Fondazione IRCCS Istituto Nazionale dei Tumori via Giacomo Venezian, 1 20133 Milano, Italy

9 Department of Oncology and Hemato-Oncology, University of Milan, Via S. Sofia 9/1, 20122 Milano, Italy

10 Department of Medicine and Surgery, University of Brescia, Viale Europa, 11 - 25123 Brescia, Italy

11 Dept. of Otorhinolaryngology- Head and Neck Surgery, University Hospital Düsseldorf, Moorenstr. 5, 40225 Düsseldorf, Germany

12 Maxillofacial Surgery Unit, Department of Medicine and Surgery, University of Parma – University Hospital of Parma, via Università, 12 – I, 43121 Parma, Italy

13 Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Radiation Oncology, Cancer Center Amsterdam, Postbus 7057, 1007 MB Amsterdam, The Netherlands

14 Oncoradiomics SA, Liège, Clos Chanmurly 13, 4000 Liège, Belgium

3.1 Abstract

3.1.1 Background

Patients with locoregionally advanced head-and-neck squamous cell carcinoma (HNSCC) have high relapse and mortality rates. Imaging-based decision support could improve outcome by optimizing personalized treatment, and support stratification in clinical trials. We hypothesize a multifactorial prognostic model including radiomics features improves risk-stratification for advanced HNSCC compared to TNM 8th edition, the gold standard.

3.1.2 Patient and methods

Data of 666 retrospective (training) and 143 prospective (validation) stage III-IVA/B HNSCC patients were collected. A multivariable Cox regression model was trained to predict overall survival (OS) based on radiomics features derived from the primary tumor on diagnostic CTs. Separate analyses were performed using TNM8, tumor volume, clinical and biological variables, and combinations thereof with radiomics features. Patient stratification was assessed through Kaplan-Meier (KM) curves and log-rank test for significance (P -value <0.05). The prognostic accuracy was reported through the concordance-index (CI).

1.3 Results

A model combining an 11-feature radiomics signature, clinical and biological variables, TNM8, and volume could significantly stratify the validation cohort into three risk groups ($P<0.01$), with a CI of 0.79 in validation.

1.4 Conclusion

A combination of radiomics features with other predictors can predict OS very accurately for advanced HNSCC patients and improves on the current gold standard of TNM8.

Simple Summary

Advanced head and neck cancer patients generally have a high mortality rate. Improving prognosis could help with this survival rate as it may improve clinical decision making. Radiomics features calculated from images of the tumor describe tumor size, shape and pattern. These characteristics may be linked to patient survival, which is investigated in this paper. We combined radiomics features with other biomarkers of survival of 809 patients to make a prognosis before treatment. We then compared the predicted prognosis with the actual outcome to see how well our model performs. Our model was able to make three distinct risk groups of low-, medium-, and high-survival patients. With these findings doctors may make a better judgement of treatment and follow-up per patient, which might improve clinical outcome.

3.2 Introduction

Head and neck squamous-cell carcinomas (HNSCC) are malignant tumors typically occurring in the oral cavity (OC), larynx, and pharynx. In Europe, 140,000 new cases are diagnosed yearly leading to approximately 70,000 deaths. [1] Despite advances in treatment, 3-years survival for locoregionally-advanced HNSCC remained 40%-50%. [2, 3] Management of HNSCC patients starts with a diagnostic workup of the tumor, lymph node metastases, and distant metastases (TNM) to stage the tumor. [4] Furthermore, p16 protein expression determined by immunostaining, a surrogate marker of HPV infection, has been included as a relevant factor in the American Joint Committee on Cancer (AJCC) 8th edition for staging of oropharyngeal cancer, in which different staging systems for p16-positive and p16-negative oropharyngeal carcinomas were introduced. [5] Besides TNM stage, prognosis depends on clinical (e.g. patients' comorbidities, performance status) and biological (e.g. invasive growth or gene expression) factors, and for patients treated with surgery on microscopic examination of the resection specimen. [4] RNA and DNA profiling have identified molecular subtypes of HNSCC with different prognosis. [6] Some of these subtypes may include primary tumors with high heterogeneity which may react differently to treatment. [7] Defining a robust and clinically viable method to determine these subtypes is therefore essential for effective treatment of HNSCC patients.

Routine pre-treatment radiological imaging provides a source of non-invasively acquired information of the primary tumor that could be investigated for the ability to determine clinically relevant subtypes. Advanced image analysis methods such as radiomics allow for the analysis of radiographic medical images by extracting large amounts of so-called features using mathematical algorithms and finding correlations with biological and/or clinical outcomes using machine learning techniques. Previous studies have shown that radiomics in computed tomography (CT) imaging could play a role in improving prediction of prognosis of HNSCC. [8-14]

We hypothesize that the multicentric "Big Data and Models for Personalized Head and Neck Cancer Decision Support" project (BD2Decide) [15, 16] dataset provides the necessary breadth to create statistically significant high-quality models that can add complementary information to other well-known but under-utilized clinical and biological factors. [17-19] Similar to the inclusion of HPV-status to TNM8, we believe that combining these factors may improve prediction of patient prognosis instead of using them independently. We also hypothesize that a multifactorial machine learning model, including radiomics features derived from the primary tumor, can outperform the current gold standard (TNM8) in stratifying locally advanced HNSCC patients into overall survival (OS) risk groups. This new signature of radiomics features was compared against an existing signature. Furthermore,

mixed models containing TNM, tumor volume, radiomics features, clinical variables, and biological variables were developed and validated.

3.3 Materials and Methods

3.3.1 Patient characteristics

Protocol details were registered on Open Science Framework (DOI: 10.17605/OSF.IO/H4DFB). The study population was composed of locoregionally advanced HNSCC patients (stage III-IVA/B (M0) according to TNM7) receiving treatment with curative intent between 2008 and 2017, collected within the framework of the BD2Decide project (ClinicalTrials.gov Identifier NCT02832102, <http://www.bd2decide.eu/>). [15, 16] The collected patient population was originally staged at diagnosis the TNM7 staging system. During the BD2Decide project these patients were re-staged to I-IVA/B (M0) using the newly developed TNM8 staging system. The ethical approval statement and an overview of the inclusion criteria can be found in supplementary materials B. Patients' data were collected both retrospectively (diagnosis between 2008 and 2014) and prospectively (diagnosis between 2015 and 2017). The retrospective and the prospective datasets were assigned as the training and validation datasets, respectively. OS was defined as the time between the primary diagnosis and death or censored at the date of last follow-up while follow-up was consequently performed for at least three years. Patients alive with follow-up less than 2 years were excluded and defined as "lost to follow-up". Median follow-up times for training and validation datasets were determined separately through the reverse Kaplan-Meier (KM) estimate. [20] Similarity in patient characteristics between cohorts was assessed through two-proportion z-tests to test whether there is a difference in a categorical variable, or unpaired two-sample t-tests to test whether there is a difference in a numerical variable. For the latter, the assumptions of the data having a normal distribution and possessing the same variance in both cohorts were tested through Shapiro-Wilk's test and f-test, respectively. The significance level was set to 5%.

3.3.2 CT acquisition parameters

CT images were acquired at each center with scanners, acquisition protocols, and reconstruction protocols according to standard operating procedures (SOPs) at the respective centers for diagnostic imaging. All CT images were either diagnostic or radiotherapy treatment planning images of comparable diagnostic quality, all with an intravenous contrast injection protocol. All CT scans within the framework of the BD2Decide project had a 3 mm slice thickness or less. Any CT scan with artifacts in more than 50% of the slices containing the primary tumor mass was excluded. For patients who received radiation therapy, the gross tumor volume (GTV) of the primary tumor was delineated at each center according to local delineation protocols by experienced radiation oncologists.

The GTV was defined as the visual extent of tumor as described in the radiology report and if needed adapted based on the report of the physical examination. Figure 1 gives an example of a CT with the primary tumor delineated. For patients who did not receive radiation treatment, the primary tumor volume was delineated locally by or supervised by expert radiologists according to local delineation protocols. For all patients treated with radiotherapy, all contours were on CT in conjunction with PET/MRI, which has been proven to greatly decrease inter observer variation. [21, 22] All contours were additionally peer-reviewed by radiation oncologists based on diagnostic information. Lastly, all delineations were visually judged by a single observer in the BD2Decide consortium for deficiencies.



Figure 1. Computed tomography image of patient with stage 3 oral cavity in transverse plane. The tumor is shown outlined in red.

3.3.3 Feature extraction

Features were extracted from the delineated primary tumor volume of the pre-processed images. A full list of software packages used in the present study is shown in table I supplementary materials B. Feature extraction was performed in python 3.6.10, with the package PyRadiomics version 2.2.0. [23] To lessen the impact of heterogeneity in the imaging data caused by differences in scanners and imaging protocols, pre-processing of the images and post-processing of the extracted features was performed. An overview of pre- and post-processing techniques applied to the data has been described in supplementary materials B. Both International Biomarker Standardization Initiative (IBSI)-compliant [24, 25] and a non-IBSI compliant feature were extracted. Features extracted through PyRadiomics contain a single first order feature, first order kurtosis, which differs from the IBSI definition. A description of the features can be found in supplementary materials B, and the full list in the PyRadiomics documentation. [26]

3.3.4 Feature selection

Univariate feature selection was performed by fitting a univariate cox model for each individual feature. Afterwards, we select features based on the individual feature's association with survival. This is done by choosing features with a testing association *P*-value (Wald-test) lower than the threshold of 0.05. [27] *P*-values were adjusted for multiple testing through false discovery rate (FDR) adjustment. [28] The function used 100-repeat 10-fold cross-validation to determine the best performing features on average.

5 Radiomics model

The selected features were used to train a multivariable Cox-model on the training dataset. Afterwards, the prognostic model performance was assessed through external validation on the validation dataset. This was done according to the principles and methods described by Royston and Altman (2013) [29], described in supplementary materials B. Model discrimination performance was determined through Harrell's C-index (CI). A CI near 0.5 indicates that the predictions are no better than chance while values near 1 indicate almost perfect discriminative performance. Risk-stratified KM curves were generated for each model, which allowed for visual comparison between models, and provided the opportunity to determine how well the cohort could be stratified into risk groups. Three risk-groups were determined using threshold values at the 33rd and 66th percentile of the calculated prognostic index (PI). A log-rank test was performed to determine the significance of the split of the low-risk vs. the medium-risk group, and the medium-risk vs. the high-risk group. In addition, predicted survival curves for each risk group are determined. The PI was used to estimate the survival curve, which was then averaged over the entire risk-group. These curves were plotted alongside the observed KM-curves. The observed survival curves and predicted survival curves aligning indicates that the model fits correctly to the data.

3.3.6 Staging, volume, and clinical models

The performance of the radiomics model was compared to risk stratification based on TNM8, primary tumor volume, and a model built from clinical and biological features. The radiomics feature “original_shape_VoxelVolume” was used as a surrogate for tumor volume. This feature was added to the list of selected features and used to create a separate model. [30] The clinical and biological model was built from a list of known predictors of survival in HNSCC, which can be found in Supplementary materials B. All features had less than 10% of values missing. For any missing values imputation was performed using the ‘missForest’ package in R. [31] This imputation method trains a Random Forest (RF) model on the existing data to predict the missing values. Separate imputation was performed for the training dataset and the validation dataset. Feature selection on the clinical and biological covariates was performed through univariate Cox modelling, selecting univariate significant covariates through chi-square test *P*-values after correcting for multiple testing (FDR). [28] The significant features were added to the list of radiomics features and used to create separate models. In addition, a combined model using radiomics, tumor volume, and clinical/biological variables was created and validated.

3.3.7 Validation of existing radiomics signatures

Aerts et al. reported on a radiomics signature to predict survival in lung cancer patients which they validated on HNSCC cohorts. This signature was evaluated both on our validation and the full cohort (training and validation) [32] and its performance was compared to the radiomics signature created in this study. The four features used to create the signature by Aerts were extracted from the primary tumor volume after the appropriate pre-processing steps. The feature values were multiplied with the β coefficients reported in the article to calculate the linear predictor. The article used a single cut-off value based on the median of the linear predictor to stratify the patients into low- and high-risk groups. We apply these cut-offs in order to determine two risk-groups and compare these to risk-stratification using the median of the linear predictor estimated by our novel models.

3.3.8 Radiomics quality score and TRIPOD

For quality assurance the radiomics quality score (RQS) [33, 34] was calculated and transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) [35] recommendations were followed. A description of these statements and the results can be found in supplementary materials B.

3.4 Results

3.4.1 Clinical, biological, and imaging characteristics

In total, 666 retrospective and 143 prospective patients were collected and analyzed in this study. An overview of patient characteristics for both cohorts is presented in table 1.

Table 1. Patient characteristics overview for retrospective and prospective patient cohorts.

Study		Retrospective	Prospective	P-value
		(N=666)	(N=143)	
Sex (% male/N)		72/482	65/93	P = 0.10
Age (Median / range)		63 / 29-89	64 / 38-93	P = 0.17
HN tumor site (%/N)	- Hypopharynx	15/96	15/21	P = 0.93
	- Oropharynx	43/289	36/51	P = 0.11
	- Oral cavity	15/100	29/42	P < 0.01
	- Larynx	27/181	20/29	P = 0.11
p16+ Oropharynx (%/N)		22/146	26/37	P = 0.36
Stage TNM7 th edition (%/N)	- III	31/206	28/40	P = 0.55
	- IVa	59/390	67/96	P = 0.07
	- IVb	10/70	5/7	P = 0.06
Stage TNM8 th edition (%/N)	p16+ oropharynx - I	11/74	12/17	P = 0.90
	- II	6/42	9/13	P = 0.31
	- III	5/30	5/7	P = 1
	Non-oropharynx/ p16- oropharynx - III	25/169	28/40	P = 0.59
	- IVa	37/248	38/54	P = 0.98
	- IVb	16/103	8/12	P = 0.04
Treatment (% of patients received type of treatment/N)	- RT only	29/191	15/22	P < 0.01
	- Surgery only	5/34	4/5	P < 0.01
	- CRT	37/245	36/51	P = 0.55
	- Surgery + RT	15 (102)	24/34	P = 0.16
	- Surgery + CH + RT	14/93	12/17	P = 0.60
Order of CH (% of CH patients/N)	- Adjuvant	15/51	12/8	P = 0.61
	- Concomitant	81/273	84/57	P = 0.64
	- Induction	4/15	4/3	P = 1

ACE-27 Co-morbidity (%/N)	= 0	30/204	38/52	<i>P</i> = 0.20
	= 1	41/272	38/52	<i>P</i> = 0.37
	= 2	20/133	16/21	<i>P</i> = 0.18
	= 3	9/57	8/11	<i>P</i> = 0.86
Smoking (%/N)	- Current	52/350	40/55	<i>P</i> = 0.01
	- Former	36/237	33/45	<i>P</i> = 0.44
	- Never	12/79	27/37	<i>P</i> < 0.01
Pack years (Median / range)		35 / 0 - 174	30 / 0-220	<i>P</i> = 1
Alcohol consumption (%/N)	- Current	66/445	48/67	<i>P</i> < 0.01
	- Former	13/84	12/17	<i>P</i> = 1
	- Never	21/137	40/55	<i>P</i> < 0.01
ECOG PS (%/N)	= 0	39/262	49/68	<i>P</i> < 0.01
	= 1	16/106	43/59	<i>P</i> < 0.01
	= 2	3/21	8/11	<i>P</i> = 0.22
	= 3	1/4	-	<i>P</i> = -
	= NA	41/273	4/5	
Hb level (Median / range)		8.8 / 5.0 -15.1	8.7 / 4.8-14.0	<i>P</i> = 0.27

HN = Head and neck, RT = Radiotherapy, CH = Chemotherapy, CRT = Chemoradiotherapy, ECOG PS = Eastern cooperative oncology group performance status

The median follow-up of patients in the training and validation cohort was 63 (49-79 95% CI) and 32 (26-37 95% CI) months, respectively. Two-year survival in the training and validation cohort was 78% and 75%, respectively. A log-rank test between survival curves shows that the difference between cohorts is not significant ($p=0.29$). KM plots of the cohorts are shown in supplementary materials A figure 1. As oropharyngeal carcinoma constituted a significant portion of the dataset (43%/N=294 for training, 36% N=51 for validation) we decided to build separate models for this group of patients (including both p16+ and p16-). A description of this model along with the results can be found in supplementary materials A. Supplementary materials B figure 1 shows an overview of the different parameters used for image acquisition and reconstruction in the training and validation datasets.

3.4.2 Model results

We extracted 1198 radiomics features from the primary tumor volume on all CT images. After unsupervised feature selection, 204 features remained. Eleven features were selected by supervised selection as being the most predictive of OS in the training cohort. The first two features were kurtosis, a first-order statistics feature, and sphericity,

a shape feature. The next four features are all LoG-filtered texture features consisting of GLSZM Gray level non-uniformity, GLDM entropy, GLRLM run entropy, and GLDM low gray level emphasis. Finally, five wavelet-filtered texture features were included: four differently wavelet-filtered GLSZM zone entropy features, and GLRLM low gray level run emphasis. All selected features were IBSI-compliant, except for the first-order statistics feature. Supplementary materials B table 2 shows an overview of the feature names. The slope of the PI in validation was 1.35. A log-rank test to see if the slope was significantly different from 1 resulted in a P -value of 0.38. This indicates the model calibrates well in the validation cohort, meaning the predicted and the expected outcome proportions for a certain testing population match. The joint test of all predictors with the offset of the PI gives a P -value of 0.86, indicating there is no evidence of a lack of fit on the validation cohort. The CI in training and validation were 0.65 and 0.67, respectively.

For the validation of the Aerts et al. (2014) signature [32], supplementary materials A figure 5 depicts KM survival curves for the combined training and validation cohort after stratification in two risk groups ($p < 0.01$), with a CI of 0.66. For some patients, one or more of the required features failed to extract due to the small size of the volume. Therefore, the calculation of the signature was not possible in all available patients, resulting in 633 patients of the training cohort and 139 patients in validation cohort. The performance of the validation in this study is similar to the reported performance of the validation on the lung dataset (CI of 0.65), but slightly lower than the performance on the two H&N datasets (both CI of 0.69).

Figure 2 shows KM survival graphs of the validation cohort split using the previously created signature [32] and the radiomics-only signature from this study. While the CI of the model performances are similar (0.66 and 0.67, respectively), the split and hazard ratio are significantly better using the newly created signature ($p = .22$ vs $p < 0.01$).

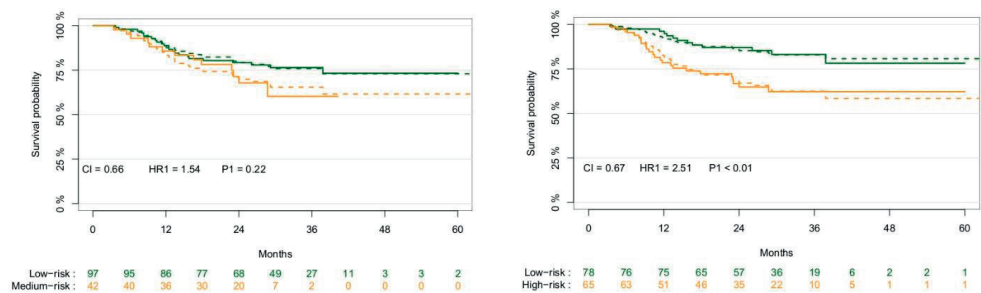


Figure 2. Kaplan-Meier survival plots of the validation cohort (N=139) stratified based on the previously created signature (left) and the newly created signature (right), showing the P -value of the split between risk-groups, model performance through the CI and the HR between the risk groups. The solid lines represent the observed survival curves, the dashed line the corresponding predicted survival curves.

Figure 3 shows the KM survival graphs of the training and validation cohorts. The P -values of the log-rank test of the low and medium and medium and high split were <0.01 for both in training, and 0.163 and 0.01 in validation, respectively.

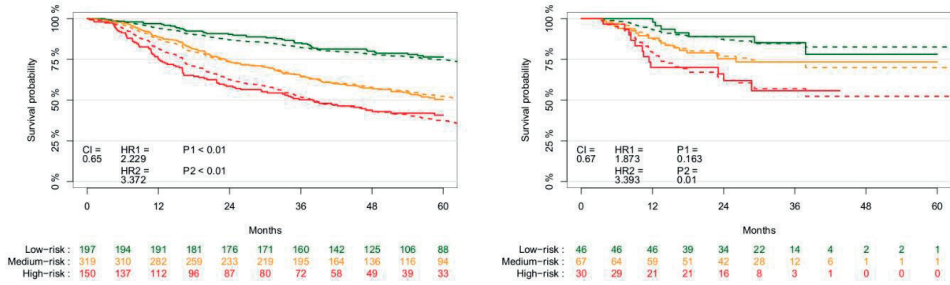


Figure 3. Kaplan-Meier survival curves of the training (left, N=666) validation (right, N=143) patient cohorts stratified into low-, medium-, and-high risk groups, showing log-rank test P -value of the split between risk groups and the CI of the radiomics features-based model-performance. The solid lines represent the observed survival curves, the dashed line the corresponding predicted survival curves.

Table 2. Selected clinical and biological features in the clinical, biological, and combined models, with univariate model coefficient, hazard ratio and significance to outcome shown.

Feature name	Model coefficient	Hazard ratio	P-value
TNM8	0.76	2.14	<0.01
Age	0.034	1.035	<0.01
ACE-27 comorbidity score	0.28	1.33	<0.01
Pack years	0.005	1.005	0.02
Alcohol at diagnosis	0.47	1.61	<0.01
P16-status	-1.3	0.27	<0.01
Hemoglobin level	-0.3	0.74	<0.01

Figure 4 shows KM survival curves of the validation cohort after stratification based on tumor volume, the selected clinical and biological parameters, and the selected radiomics features, with a CI of 0.71 and 0.79 in training and validation, respectively. The clinical features selected through univariate feature selection were TNM8 (higher stage has worse prognosis), age at diagnosis (higher age has worse prognosis), ACE-27 comorbidity score (higher score has worse prognosis), smoking pack years (higher pack years has worse prognosis), and alcohol consumption at time of diagnosis (current has worst prognosis), and the biological features were p16-status (p16-negative has worse prognosis) and clinical Hb level at baseline (lower Hb level has worse prognosis). The P -value of the log-rank test of the low and medium and medium and high split were both <0.01 in both training and validation.

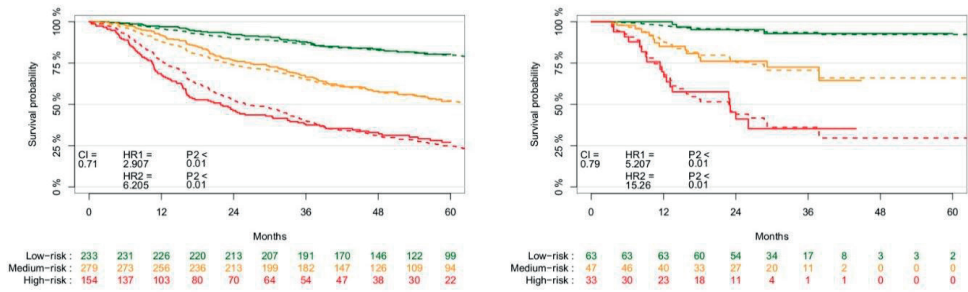


Figure 4. Kaplan-Meier survival cohorts of the training (left, N=666) validation (right, N=143) patient cohorts stratified into low-, medium-, and-high risk groups based on radiomics, tumor volume, clinical, and biological parameters, showing the P-value of the split between risk-groups and CI of the model performance. The solid lines represent the observed survival curves, the dashed line the corresponding predicted survival curves.

A full overview of the different combinations of models, with discrimination performance and hazard ratios for each model in training and validation, is provided in Table 3. In addition, Figure 4 provides an overview of the CI-indices of the validation results.

Table 3. Performance overview of all trained and/or validated models, showing Harrell's CI and HR values for each model.

Model	Full patient cohort						Oropharynx patient cohort					
	Training			Validation			Training			Validation		
	CI (95% CI)	HR 1vs.2 (95% CI)	HR 1vs.3 (95% CI)	CI (95% CI)	HR 1 vs. 2 (95% CI)	HR 1vs.3 (95% CI)	CI (95% CI)	HR 1vs.2 (95% CI)	HR 1vs.3 (95% CI)	CI (95% CI)	HR 1 vs. 2 (95% CI)	HR 1 vs. 3 (95% CI)
Staging TMM8	0.65 (0.64-0.65)	1.82 (1.40-2.35)	3.12 (2.32-4.21)	0.74 (0.73-0.75)	5.01 (2.11-11.85)	14.03 (5.16-38.17)	0.71 (0.69-0.72)	2.50 (1.62-3.87)	5.16 (3.24-8.23)	0.86 (0.81-0.87)	9.12 (1.28-64.90)	30.15 (4.97-182.90)
Radiomics	0.65 (0.64-0.65)	2.22 (1.64-3.03)	3.37 (2.41-4.72)	0.67 (0.66-0.69)	1.87 (0.78-4.52)	3.39 (1.33-8.64)	0.68 (0.67-0.69)	2.36 (1.45-3.86)	3.80 (2.21-6.52)	0.82 (0.78-0.85)	-*	-*
Radiomics + Staging	0.68 (0.68-0.69)	2.49 (1.77-3.44)	4.60 (3.24-6.53)	0.77 (0.75-0.78)	8.54 (1.97-37.98)	29.35 (6.73-127.94)	0.73 (0.73-0.74)	3.97 (2.20-7.18)	7.87 (4.39.27)	0.90 (0.88-0.92)	-*	-*
Radiomics (Volume)	0.62 (0.62-0.62)	1.48 (1.08-2.03)	3.17 (2.16-4.66)	0.68 (0.66-0.69)	1.23 (0.54-2.78)	7.98 (2.85-22.31)	0.64 (0.63-0.64)	1.81 (1.10-2.99)	3.29 (1.82-5.92)	0.87 (0.84-0.90)	-*	-*
Clinical	0.66 (0.66-0.67)	2.37 (1.76-3.19)	3.25 (2.40-4.40)	0.70 (0.69-0.72)	3.66 (1.40-9.54)	5.37 (2.10-13.72)	0.73 (0.72-0.74)	3.80 (2.18-6.63)	8.27 (4.82-14.18)	0.84 (0.81-0.87)		
Biological	0.63 (0.63-0.63)	2.83 (1.95-4.09)	3.94 (2.71-5.74)	0.70 (0.68-0.71)	13.03 (1.74-97.73)	23.19 (3.08-174.46)	0.68 (0.68-0.69)	4.28 (2.79-6.56)	6.74 (0.91-49.82)	0.84 (0.80-0.89)	*	*
Clinical + Biological	0.67 (0.66-0.67)	2.71 (1.95-3.75)	4.17 (3.00-5.78)	0.73 (0.72-0.74)	8.21 (2.37-28.39)	10.10 (2.97-34.36)	0.74 (0.74-0.75)	3.82 (2.16-6.76)	8.66 (5.08-14.76)	0.88 (0.85-0.90)	-*	-*

Table 3. Continued

Radiomics (includes volume)	0.65 (0.65-0.66)	1.78 (1.32-2.42)	3.64 (2.61-5.08)	0.68 (0.67-0.69)	2.19 (0.92-5.26)	3.84 (1.48-9.95)	0.68 (0.67-0.69)	2.47 (1.50-4.06)	3.94 (2.28-6.82)	0.82 (0.78-0.86)	_*
Radiomics + Clinical	0.69 (0.69-0.70)	2.94 (2.15-4.03)	4.79 (3.45-6.67)	0.74 (0.74-0.76)	4.65 (1.86-17.16)	11.38 (3.84-33.74)	0.73 (0.72-0.74)	3.80 (2.18-6.64)	8.27 (4.82-14.18)	0.84 (0.81-0.87)	_*
Radiomics + Biological	0.68 (0.68-0.68)	2.89 (2.04-4.08)	5.03 (3.52-7.17)	0.76 (0.74-0.77)	6.49 (1.91-22.06)	13.74 (3.96-47.66)	0.74 (0.74-0.75)	3.61 (2.13-6.12)	6.85 (4.12-11.39)	0.91 (0.90-0.93)	_*
Radiomics + Clinical + Biological	0.70 (0.70-0.70)	3.04 (2.17-4.27)	5.82 (4.10-8.28)	0.77 (0.77-0.78)	8.17 (2.36-28.24)	13.17 (3.86-44.85)	0.77 (0.77-0.78)	4.77 (2.65-8.60)	12.53 (7.03-22.31)	0.88 (0.85-0.90)	_*
Radiomics (includes volume) + Clinical + Biological	0.71 (0.71-0.71)	2.91 (2.11-4.01)	6.21 (4.44-8.68)	0.79 (0.78-0.80)	5.21 (1.70-15.98)	15.26 (5.14-45.32)	0.77 (0.77-0.77)	6.11 (3.23-11.53)	15.40 (8.17-29.03)	0.87 (0.84-0.89)	_*
p16-status	-	-	-	-	-	-	0.67 (0.67-0.68)	4.3 (2.81-6.59)	-	0.82 (0.78-0.85)	19.8 (2.38-165)
Aerts. 2014[32]	0.61 (0.61-0.61)	1.65 (1.30-2.09)	-	0.66	1.54 (0.77-3.06)	-	0.65 (0.64-0.66)	1.90 (1.3-2.77)	-	0.68 (0.63-0.73)	_*

The left side shows the models for the full patient cohort, both training (N=666) and validation (N=143), the right the oropharynx patient cohort, both training (N=294) and validation (N=51). * indicates no HR could be calculated, as the low-risk group did not have any events recorded.

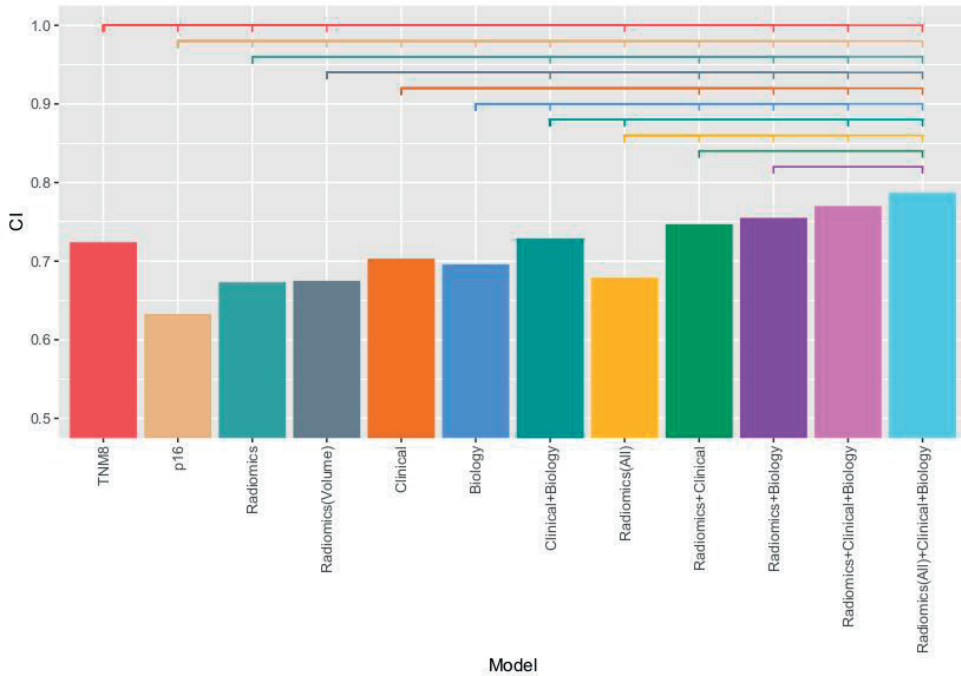


Figure 5. Bar-plot of the various models validated on the validation cohort (N=143). The y-axis indicates CI value, while the coloured bars above the bar show significant differences between models, with an indent meaning the model is significantly different, and no indent meaning no significant difference was found

From Table 3 and Figure 5, it can be observed that in the prospective cohort radiomics alone does not perform better than TNM8 (CI of 0.67 and 0.74 respectively, $p < 0.01$). Combining TNM8 and radiomics results in a higher performance than both separately, with a CI of 0.77. In combination with both clinical parameters and tumor volume, the highest discrimination performance was found (CI of 0.79). Similarly, oropharynx radiomics does not perform better than TNM8 (CI of 0.82 vs. 0.86, $p < 0.01$), but when combining both radiomics and TNM8 the highest performance in validation was achieved (CI of 0.90).

3.5 Discussion

For advanced tumors like those investigated in this study, being able to discern groups of poor versus good performing patients is key for personalized decision making. In this international, multicenter study, we created a multifactorial prediction model, including radiomics features extracted from the primary tumor volume, that can significantly stratify advanced HNSCC patients in good, average and poor prognostic groups, with a

CI of 0.79 in validation on a prospective cohort. These groups could be used in clinical decision making and for selecting patients for (de-)escalation trials and/or adjuvant treatment. While radiomics alone was not able to improve on TNM8, adding radiomics features to a model including TNM8, clinical, and biological variables improved the prognostic performance, significantly increasing CI from 0.73 to 0.79. We can therefore recommend adding these variables to the current clinical implementation of TNM8. These results coincide with other works reporting on the complementary value of radiomics in predictive modelling in head and neck cancer. [9, 36]

Eleven radiomics features were selected for the prediction of OS. The first two selected features were kurtosis, a first-order statistics feature that measures the 'peakness' of the distribution of pixel intensity values, and sphericity, a shape feature that measures the likeness of the ROI to a sphere. Sphericity being selected implies less spherical tumors may have a worse prognosis. The next four features are all LoG-filtered texture features consisting of GLSZM Gray level non-uniformity, a feature which measures the variability of gray-level intensity values, GLDM entropy and GLRLM run entropy, which both measure the heterogeneity in texture patterns, and GLDM low gray level emphasis, which measures the concentration of low intensity values. Finally, five wavelet-filtered texture features were included: four differently wavelet-filtered GLSZM zone entropy features, which measure the heterogeneity in texture patterns, and GLRLM low gray level run emphasis, which measures the concentration of low intensity values. Most of these features are linked to heterogeneity, reinforcing the theory that tumor heterogeneity correlates with a worse prognosis. [37, 38]

For most tested models we found a higher validation accuracy than training accuracy. The main reason for this is the smaller size of the validation dataset, which means the result is more prone to variance, which is reflected in the larger confidence intervals, especially for the smaller oropharyngeal analysis. Another contributing factor could be a larger number of 'hard' cases in the training dataset compared to the validation dataset. In this paper, we chose to validate on a prospectively collected dataset, which is for data splitting purposes an arbitrary reason. In a more balanced dataset with more similar patient datasets the discrepancy between training and validation may be lower.

Instead of using radiotherapy planning images only, which is conventional for radiomics studies, this study used diagnostic CT images as well, which are made routinely for any patient showing a locally advanced HNSCC. From these images radiomics features can be extracted in a semi-automatic fashion, making clinical application easy. In addition, the combined model was made using simple variables that are routinely determined in a clinical setting for every patient (TNM8 stage, ACE-27 comorbidity status, smoking and alcohol habits). This makes the potential application of the presented models in a clinical

setting relatively easy. For the next step, the created model could be tested in a clinical trial. However, as differences in scanners, scan settings, and acquisition settings have proven to have significant effects on feature reproducibility, further external validation of the models in a prospective study where these variables are controlled may be required.

Radiomics performs an estimation of the tumor volume using a 3D segmentation, as opposed to conventional methods of measuring tumor volume to predict survival. This single feature was found to be significantly predictive of OS, albeit with lower performance compared to TNM8 or the model based on radiomics features, but was not chosen in the multivariable model. The main reason for this is the interaction with other features in the correlation dimensionality reduction step. Volume has high correlation with other features, mostly shape features, and is therefore removed from the feature dataset before univariate selection is performed, revealing a shortcoming of this feature reduction step. However, the information provided by this feature should be retained in the remaining uncorrelated features.

The radiomics model in this study shows better performance in stratifying patients in risk-groups in the validation dataset when compared to the previously created and validated signature. [32] One large discrepancy between these models is the risk stratification: the previously developed signature was created with two risk-groups, instead of three. Most importantly, it was built on lung cancer. The difference in performance on different tumor sites demonstrates that prognostic models should be developed on specific tumor sites and stages, and with relevant clinical risk groups in mind.

While the amount of data used in this study was higher than most published radiomics studies, this was partially achieved by pooling data from different HNSCC sites. Separating these regions resulted in very small datasets in either or both, training and validation sets. While we had sufficient data to train an oropharynx model and found a relatively high performance of the model using radiomics features of 0.82 CI in validation, the validation dataset was relatively limited with respect to the number of patients, and particularly in number of events. Collecting more data from an individual tumor site would most likely result in more representative models. In addition, the patients in this study received different treatments. This can have a major impact on survival chance, and is therefore a major limitation. Similar to tumor region, separate models according to treatment would be preferred. However, treatment is heavily linked to region of the tumor, as for example the majority of surgeries were performed for oral cavity patients.

Compared to extracting radiomics features from just the primary tumor volume, TNM8 staging takes information from the primary tumor (T-stage), nodal involvement (N-stage), and the presence of distant metastases (M-stage) into consideration. In addition,

depending on the tumor region, additional information such as p16-status as surrogate for HPV-involvement, depth of invasion in surrounding tissues, and presence of extranodal extension are important. An improvement could be made by including radiomics features extracted from the lymph node metastases. This would require a multifactorial model with a binary condition for lymph node stage and would only incorporate features of those patients who have lymph node metastases.

Imaging artefacts caused by dental implants may have affected performance of the radiomics model. The artefacts make segmentation difficult, but also affect the radiomics features extracted from these images. While there was a limit on the number of artefacts allowed on images during patient selection, methods to reduce the artefacts may be considered for future studies. In addition, variability caused by the manual segmentation of tumors by different experts at each institute may have also affected model performance. Previous research has shown that inter- and intra-observer variability can possibly cause large differences in delineated volumes.[39] For shape and size radiomics features, this can cause a large decrease of their utility, and may affect other features to a lesser degree. The repeatability of deep learning based automatic segmentation methods will be able to negate inter-observer variabilities in the future.[40]

To compensate for inter-observer variability in the current project, each center performed delineations either directly by, or under supervision of, expert radiologists or radiation oncologists. And although delineations were performed according to local protocols, European guidelines are largely aligned, limiting the inter-observer effects on the delineated structures. Conversely, in a clinical application of the proposed model at different institutes, inter-observer variabilities will be an inevitability. The discriminative performance the model has shown despite these issues strengthens the potential of application in a clinical setting.

3.6 Conclusion

A multifactorial prognostic model for stage III and IV HNSCC (TNM7th edition) based on simple variables available for every patient and including CT-radiomics features is able to very accurately predict OS and to significantly discern different risk-groups. The multifactorial model was found to have higher predictive performance than the current gold standard of TNM8. This could be useful in treatment (de-)escalation trials and clinical decision-support.

3.7 Acknowledgements

Irene H. Nauta, Frederik W.R. Wesseling, Stefano Cavalieri, Kathrin Scheckenbach, Marco Ravanelli, Davide Lanfranco, Tito Poli, Thomas K. Hoffmann, Kathrin Scheckenbach, Giuseppina Calareso, Marije R. Vergeer, C René Leemans, Frank J.P. Hoebbers have recruited the patients used in the current study. Irene H. Nauta, Frederik W.R. Wesseling, Stefano Cavalieri, Kathrin Scheckenbach, Marco Ravanelli, Davide Lanfranco, and Thomas K. Hoffmann have collected and curated the clinical and biological covariates. Marije Vergeer, Frederik Wesseling, Giuseppina Calareso, Kathrin Scheckenbach, Marco Ravanelli, Davide Lanfranco, and Thomas K. Hoffmann have collected, segmented and curated the radiological images. Janita E. van Timmeren, Henry C. Woodruff, Philippe Lambin, and Frank J.P. Hoebbers have provided supervision. Ruud H. Brakenhoff, C René Leemans, Frank J.P. Hoebbers, Lisa Licitra, Kathrin Scheckenbach, Marco Ravanelli, Tito Poli, and Thomas K Hoffmann have acquired funding for and conceptualized the current study. Ralph T.H. Leijenaar and Sergey P. Primakov have provided technical support. Simon A. Keek and Henry C. Woodruff have verified the underlying data. Simon A. Keek has performed the analysis. Simon A. Keek, Frederik W.R. Wesseling, and Henry C. Woodruff have written the original draft. All other authors have read and provided comments on the writing/ contents of the current draft.

The authors and the investigators are grateful to Dr. Elena Martinelli, project manager of the BD2Decide project, who lead the Coordination work.

3.8 Funding

This work was supported by the European Union Horizon 2020 Framework Programme [grant number 689715].

3.9 Disclosure

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno), ERC-2018-PoC: 813200-CL-IO, ERC-2020-PoC: 957565-AUTO. DISTINCT. Authors also acknowledge financial support from SME Phase 2 (RAIL n°673780), EUROSTARS (DART, DECIDE, COMPACT-12053), the European Union's Horizon 2020 research and innovation programme under grant agreement: BD2Decide - PHC30- 689715, ImmunoSABR n° 733008, MSCA-ITN-PREDICT n° 766276, FETOPEN- SCANnTREAT n° 899549, CHAIMELEON n° 952172, EuCanImage n° 952103, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY n° UM 2017-8295), Interreg V-A Euregio Meuse-Rhine

(EURADIOMICS n° EMR4), and Genmab (n° 1044). This work was supported by the Dutch Cancer Society (KWF Kankerbestrijding), project number 12085/2018–2, KWF-A6C7072 (DESIGN), and KWF project number 12079/2018-2.

Dr. Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Varian medical, Oncoradiomics, ptTheragnostic, Health Innovation Ventures and DualTpharma. He received an advisor/presenter fee and/or reimbursement of travel costs/external grant writing fee and/or in-kind manpower contribution from Oncoradiomics, BHV, Merck and Convert pharmaceuticals. Dr. Lambin has shares in the company Oncoradiomics, Convert pharmaceuticals, MedC2 and LivingMed Biotech and is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Oncoradiomics and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAMito, three non-patentable invention (software) licensed to ptTheragnostic/ DNAMito, Oncoradiomics and Health Innovation Ventures.

Dr. Lisa Licitra further acknowledge grant/research support from Astrazeneca, BMS, Boehringer Ingelheim, Celgene International, Debiopharm International SA, Eisai, Exelixis inc, Hoffmann-La Roche Ltd, IRX Therapeutics inc, Medpace inc, Merck–Serono, MSD, Novartis, Pfizer, Roche, honoraria/consultation fees from Astrazeneca, Bayer, BMS, Eisai, MSD, Merck–Serono, Boehringer Ingelheim, Novartis, Roche, Debiopharm International SA, Sobi, Ipsen, Incyte Biosciences Italy srl, Doxa Pharma, Amgen, Nanobiotics Sa and GSK, and fees for public speaking/teaching from AccMed, Medical Science Foundation G. Lorenzini, Associazione Sinapsi, Think 2 IT, Aiom Servizi, Prime Oncology, WMA Congress Education, Fasi, DueCi promotion Srl, MI&T, Net Congress & Education, PRMA Consulting, Kura Oncology, Health & Life srl, Immuno-Oncology Hub.

Dr. Henry C. Woodruff has (minority) shares in the company Oncoradiomics.

Dr. Ralph T.H. Leijenaar is a salaried employee of the company Oncoradiomics, has shares in the company Oncoradiomics and is co-inventor of an issued patent with royalties on radiomics (PCT/NL2014/050728) licensed to Oncoradiomics.

Dr. C. René Leemans is an advisory board member at Merk & Co., Inc. and Rakuten Medical, and has received a research grant from Bristol Myers-Squibb.

RH Brakenhoff PhD, received research grants from GenMab, Bristol Myers-Squibb and InteRNA BV, and has collaboration with MSD.

3.10 References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394-424.
2. Mehra R, Ang KK, Burtness B. Management of human papillomavirus-positive and human papillomavirus-negative head and neck cancer. *Semin Radiat Oncol.* 2012;22(3):194-7.
3. Lubin JH, Purdue M, Kelsey K, Zhang ZF, Winn D, Wei Q, et al. Total exposure and exposure rate effects for alcohol and smoking and risk of head and neck cancer: a pooled analysis of case-control studies. *Am J Epidemiol.* 2009;170(8):937-47.
4. Lydiatt W, O'Sullivan B, Patel S. Major Changes in Head and Neck Staging for 2018. *American Society of Clinical Oncology Educational Book.* 2018(38):505-14.
5. Lydiatt WM, Patel SG, O'Sullivan B, Brandwein MS, Ridge JA, Migliacci JC, et al. Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. *CA Cancer J Clin.* 2017;67(2):122-37.
6. Qi Z, Barrett T, Parikh AS, Tirosh I, Puram SV. Single-cell sequencing and its applications in head and neck cancer. *Oral Oncol.* 2019;99:104441.
7. Mroz EA, Rocco JW. Intra-tumor heterogeneity in head and neck cancer and its clinical implications. *World J Otorhinolaryngol Head Neck Surg.* 2016;2(2):60-7.
8. Bogowicz M, Riesterer O, Ikenberg K, Stieb S, Moch H, Studer G, et al. Computed Tomography Radiomics Predicts HPV Status and Local Tumor Control After Definitive Radiochemotherapy in Head and Neck Squamous Cell Carcinoma. *Int J Radiat Oncol Biol Phys.* 2017;99(4):921-8.
9. Ou D, Blanchard P, Rosellini S, Levy A, Nguyen F, Leijenaar RTH, et al. Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status. *Oral Oncol.* 2017;71:150-5.
10. Xie C, Yang P, Zhang X, Xu L, Wang X, Li X, et al. Sub-region based radiomics analysis for survival prediction in oesophageal tumours treated by definitive concurrent chemoradiotherapy. *EBioMedicine.* 2019;44:289-97.
11. Cozzi L, Franzese C, Fogliata A, Franceschini D, Navarria P, Tomatis S, et al. Predicting survival and local control after radiochemotherapy in locally advanced head and neck cancer by means of computed tomography based radiomics. *Strahlenther Onkol.* 2019;195(9):805-18.
12. Wu W, Ye J, Wang Q, Luo J, Xu S. CT-Based Radiomics Signature for the Preoperative Discrimination Between Head and Neck Squamous Cell Carcinoma Grades. *Front Oncol.* 2019;9:821.
13. Liu Z, Cao Y, Diao W, Cheng Y, Jia Z, Peng X. Radiomics-based prediction of survival in patients with head and neck squamous cell carcinoma based on pre- and post-treatment (18)F-PET/CT. *Aging (Albany NY).* 2020;12(14):14593-619.
14. Head MDACC, Neck Quantitative Imaging Working G. Investigation of radiomic signatures for local recurrence using primary tumor texture analysis in oropharyngeal head and neck cancer patients. *Sci Rep.* 2018;8(1):1524.

15. Cavalieri S, De Cecco L, Brakenhoff RH, Serafini MS, Canevari S, Rossi S, et al. Development of a multiomics database for personalized prognostic forecasting in head and neck cancer: The Big Data to Decide EU Project. *Head Neck*. 2020.
16. Lopez-Perez L, Hernández L, Ottaviano M, Martinelli E, Poli T, Licitra L, et al., editors. *BD2Decide: Big Data and Models for Personalized Head and Neck Cancer Decision Support*. 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS); 2019 5-7 June 2019.
17. Ramroth H, Schoeps A, Rudolph E, Dyckhoff G, Plinkert P, Lippert B, et al. Factors predicting survival after diagnosis of laryngeal cancer. *Oral Oncol*. 2011;47(12):1154-8.
18. Faye-Lund H, Abdelnoor M. Prognostic factors of survival in a cohort of head and neck cancer patients in Oslo. *Eur J Cancer B Oral Oncol*. 1996;32B(2):83-90.
19. Smith EM, Rubenstein LM, Haugen TH, Pawlita M, Turek LP. Complex etiology underlies risk and survival in head and neck cancer human papillomavirus, tobacco, and alcohol: a case for multifactor disease. *J Oncol*. 2012;2012:571862.
20. Shuster JJ. Median follow-up in clinical trials. *J Clin Oncol*. 1991;9(1):191-2.
21. Steenbakkers RJ, Duppen JC, Fitton I, Deurloo KE, Zijp LJ, Comans EF, 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(2):435-48.
22. Rasch CR, Steenbakkers RJ, Fitton I, Duppen JC, Nowak PJ, Pameijer FA, et al. Decreased 3D observer variation with matched CT-MRI, for target delineation in Nasopharynx cancer. *Radiat Oncol*. 2010;5:21.
23. van Griethuysen JMM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. 2017;77(21):e104-e7.
24. Hatt M, Vallieres M, Visvikis D, Zwanenburg A. IBSI: an international community radiomics standardization initiative. *J Nucl Med*. 2018;59.
25. Zwanenburg A, Vallieres M, Abdalah MA, Aerts H, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology*. 2020:191145.
26. community P. <https://pyradiomics.readthedocs.io/en/latest/features.html> 2016 [
27. Emura T, Matsui S, Chen HY. compound.Cox: Univariate feature selection and compound covariate for predicting survival. *Comput Methods Programs Biomed*. 2019;168:21-37.
28. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995;57(1):289-300.
29. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol*. 2013;13:33.
30. Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, et al. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiother Oncol*. 2019;130:2-9.
31. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2011;28(1):112-8.

32. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014;5:4006.
33. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology.* 2017;14(12):749-62.
34. Sanduleanu S, Woodruff HC, de Jong EEC, van Timmeren JE, Jochems A, Dubois L, et al. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. *Radiother Oncol.* 2018;127(3):349-60.
35. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med.* 2015;162(10):735-6.
36. Vallieres M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts H, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep.* 2017;7(1):10117.
37. Fidler IJ. Critical factors in the biology of human cancer metastasis: twenty-eighth G.H.A. Clowes memorial award lecture. *Cancer Res.* 1990;50(19):6130-8.
38. Yokota J. Tumor progression and metastasis. *Carcinogenesis.* 2000;21(3):497-503.
39. Granzier RWY, Verbakel NMH, Ibrahim A, van Timmeren JE, van Nijnatten TJA, Leijenaar RTH, et al. MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability. *Sci Rep.* 2020;10(1):14163.
40. Nikolov S, Blackwell S, Mendes R, De Fauw J, Meyer C, Hughes C, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:180904430.* 2018.

3.11 Supplementary materials A

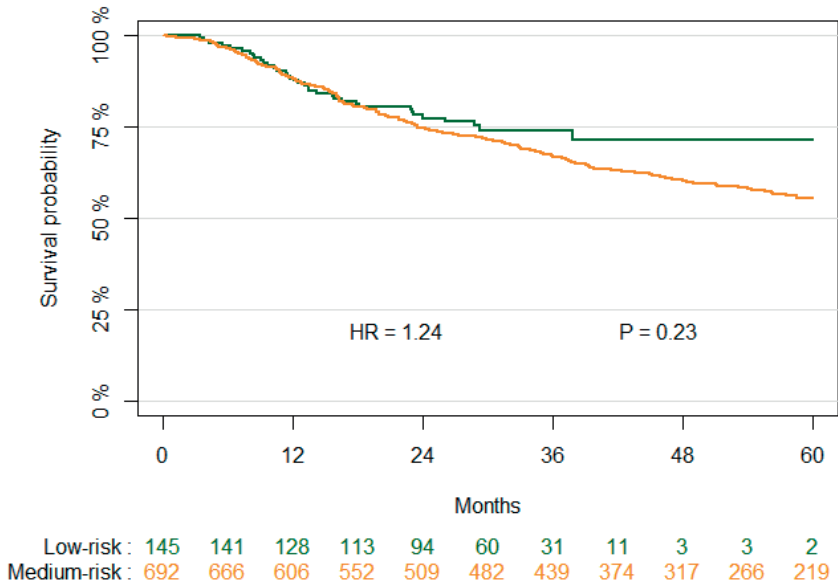


Figure 1. Kaplan-Meier survival curves of the retrospective and prospective cohorts, with p-value of the log-rank test, and the hazard ratio between the two groups.

Figure 2 shows Kaplan-Meier survival curves for the prospective cohort after stratification based on tumor volume, with a CI of 0.68. The p-values of the log-rank test of the low and medium and medium and high split were 0.62 and <0.01 , respectively.

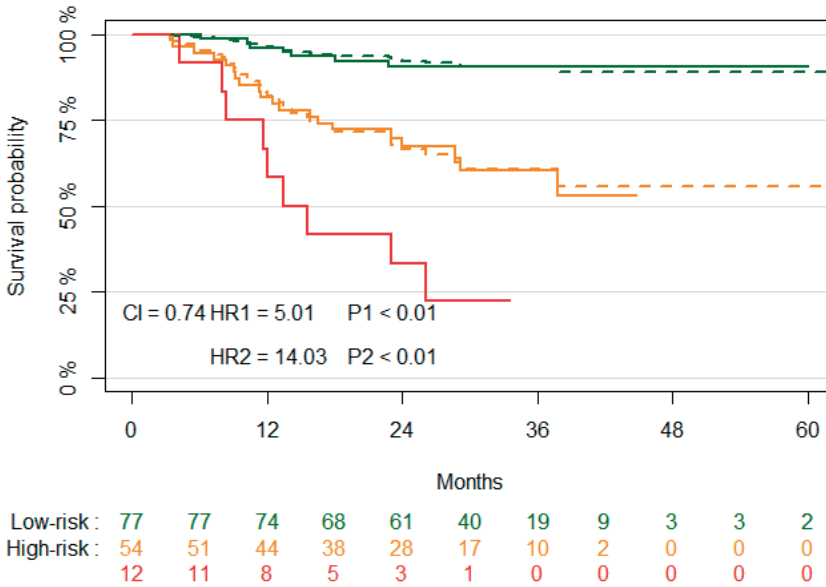


Figure 2. Kaplan-Meier survival curves prospective cohorts, with p-value of the log-rank test, and the hazard ratio between the three groups, stratified based on TNM8.

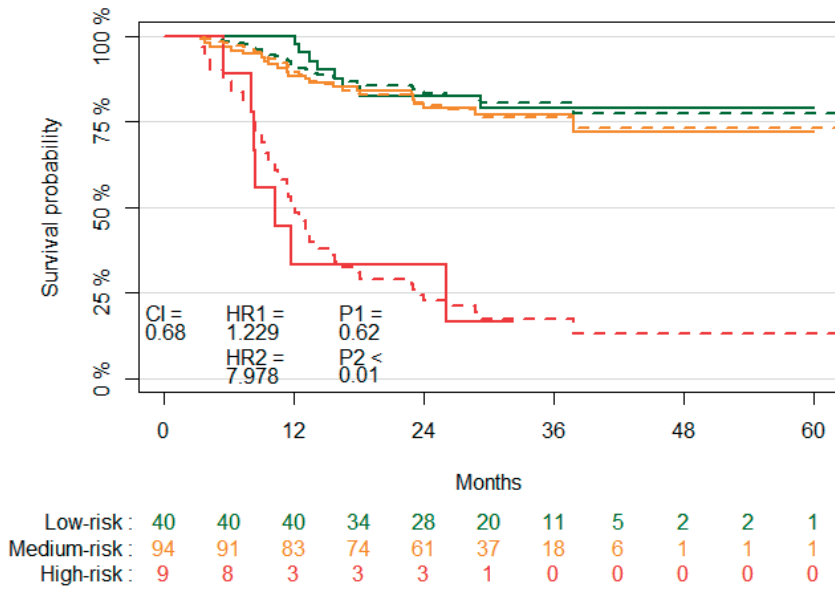


Figure 3. Kaplan-Meier survival cohorts of the prospective patient cohort (N=143) stratified based on tumor volume, showing the p-value of the split between risk-groups and CI of the model performance. The solid lines represent the observed survival curves, the dashed the corresponding predicted survival curves.

Figure 3 shows Kaplan-Meier survival curves of the prospective cohort after stratification based on clinical and biological features, with a CI of 0.73 in validation. The p-value of the log-rank test of the low and medium split was <0.01, but the p-value of the log-rank test of the medium and high split was not significant 0.57.

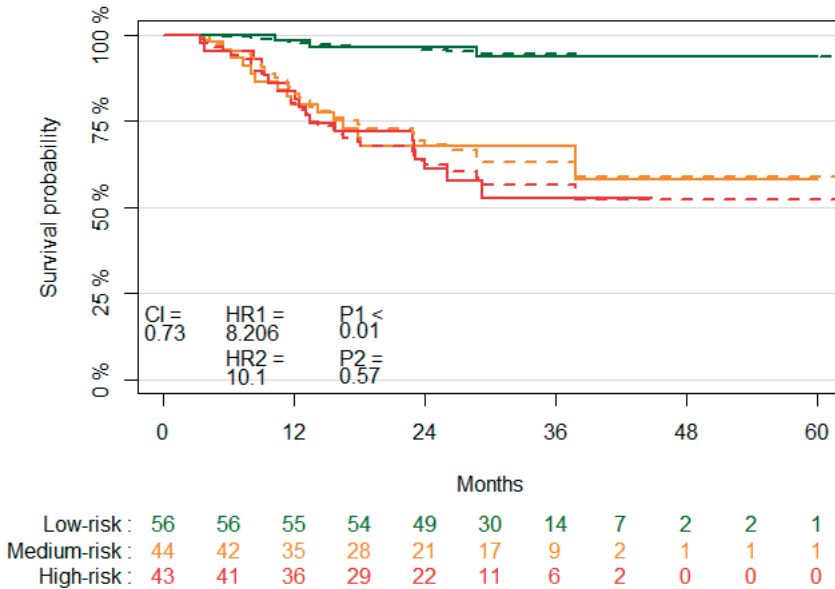


Figure 4. Kaplan-Meier survival cohorts of the prospective patient cohort (N=143) stratified based on clinical and biological parameters, showing the p-value of the split between risk-groups and CI of the model performance. The solid lines represent the observed survival curves, the dashed the corresponding predicted survival curves.

For the oropharynx patient cohort, eight features were selected as being the most predictive of OS, consisting of 1 first-order statistics feature, two shape features, 3 wavelet-filtered texture features, and 2 LoG-filtered texture features. All selected features were IBSI-compliant. Supplementary materials B table 3 shows an overview of the features. The slope of the PI in validation was 3.01, with a log-rank test p-value of 0.04, indicating certainty the slope in validation is larger than unity. The joint test of all predictors with the offset of the PI gives a p-value of 0.12, indicating there is no evidence of a lack of fit on the validation cohort. Kaplan-Meier survival curves of the prospective oropharynx cohort split based on radiomics features is shown in figure 4.

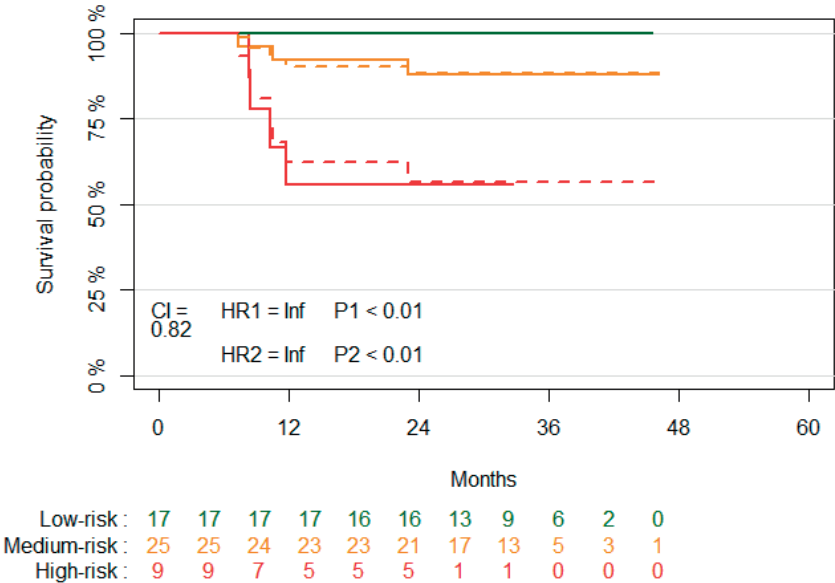


Figure 5. Kaplan-Meier survival curves of the oropharynx prospective patient (N=51) cohort using radiomics features without ComBat harmonization, showing log-rank test p-value of the split between risk groups and the CI of the model-performance in the prospective cohort. Risk group split based on median training prediction value. The solid lines represent the observed survival curves, the dashed the corresponding predicted survival curves.

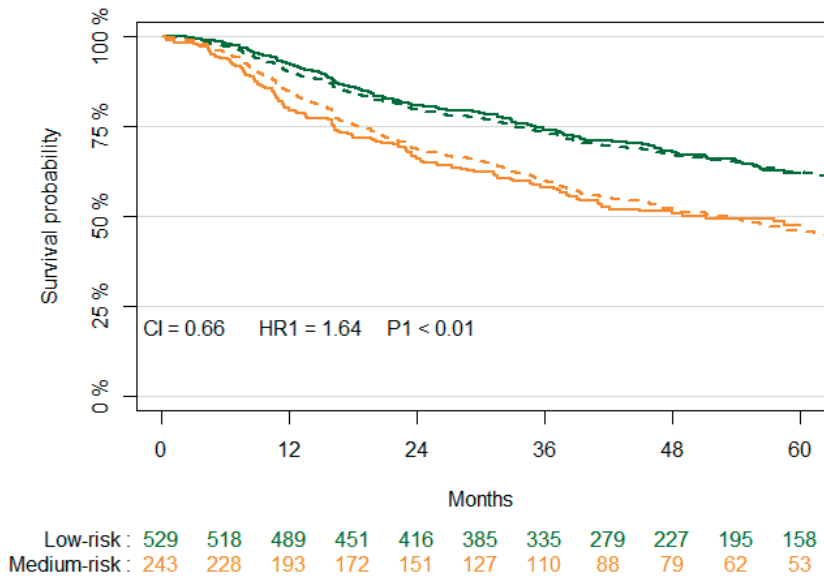


Figure 6. Kaplan-Meier survival cohorts of the full patient cohort (N=772) stratified based on the previously created signature, showing the p-value of the split between risk-groups, model performance through the CI and the HR between the risk groups. The solid lines represent the observed survival curves, the dashed the corresponding predicted survival curves.

3.12 Supplementary materials B

3.12.1 Ethical approval

The BD2Decide (H2020-PHC30-689715, IRB P-number P0125, ClinicalTrials.gov Identifier: NCT02832102) study procedures were approved by the Ethics Committees according to the Declaration of Helsinki, the European and local ethical conventions and legal aspects, as well as the European General Data Protection Regulation. The management and exchange of data, specimens, and imaging information were regulated between the partners through data and material transfer agreements and standard operating procedures. Central data, imaging, and material were anonymized by the centers prior to aggregation, and data were stored in a secured and locked information technology surrounding according to General Data Protection Regulation (GDPR). Protocol details were registered on Open Science Framework (DOI: 10.17605/OSF.IO/H4DFB).

3.12.2 Data description and inclusion criteria

Patient data were acquired from seven different centers: Fondazione IRCCS Istituto Nazionale dei Tumori Milano (INT), Azienda Ospedaliero Universitaria di Parma (AOP), Maastricht Radiation Oncology (MAASTRO), Amsterdam UMC, location VUmc, Heinrich-Heine-Universität Düsseldorf (UDUS), University of Brescia (UB), and University Ulm (UU). The data collected included clinical, biological, pathological, and radiological variables for each case. The inclusion criteria were: histological confirmation of HNSCC, age 18 years or above, clinical TNM stage III, IVA, or IVB based on AJCC 7th edition, administration of treatment with curative intent (any combination of surgery, radiotherapy, and chemotherapy), availability of pre-treatment tumor specimens, and availability of contrast-enhanced CT scan of the head and neck region.

3.12.3 Radiomic features description

Features can be divided into first-order HU intensity, histogram statistics, shape, and texture features. First order HU intensity and histogram statistics describe the total distribution of voxel intensities over the CT image. Shape features describe two- and three-dimensional size and shape of the GTV. Tumor volume measured through the voxel volume of the GTV is also a radiomics feature and can be seen as a more complex and complete feature than the size used for TNM staging. Texture features describe the relative spatial distribution of intensity values derived from 6 different matrices that are defined over the images: gray-level co-occurrence (GLCM)¹, gray-level run length (GLRLM)², grey-level size-zone (GLSZM)³, gray-level distance-zone (GLDZM)⁴, gray-level dependence (NGLDM)⁵, and neighborhood gray-tone difference matrix (NGTDM).⁶ In addition, two different image filters were applied to the original image, resulting in extra images to extract the earlier described first-order, histogram, and texture features. The first technique

is wavelet filtering, which involves 3D coif wavelet transforms along the three axes of the original images at 2 spatial frequencies (high and low) to decompose the images into 8 decomposed scans. The second filtering technique is Laplacian of Gaussian (LoG), which highlights regions of intensity change within an image. The LoG-filter was applied with 4 different standard deviation values (2-5 mm) of the Gaussian filter, resulting in 4 different LoG-filtered images.

3.12.4 Pre- and post-processing

As many radiomics features have been found to be dependent on voxel size and the number of gray levels⁷, all images were resampled to uniform 1x1x3 mm² voxels using the 'sitkBSpline' interpolator. The choice for voxel dimensions was made based on majority ruling, where we found that most patients had a slice spacing of 3mm and pixel spacing of ~1mm. Additionally, the intensity values of the images were resampled using a fixed bin-size of 25 Hounsfield Units (HU), resulting in images with ranges of 16-128 bins. This number of bins was chosen as a balance between reducing noise and limiting the size of the texture matrices on one hand and retaining a minimum contrast level in the lesions with less intensity ranges on the other. Disconnected voxels were removed to ensure only one fully connected structure was used for feature calculations. Z-score normalization metrics (mean and scale) for all radiomics features except for shape features were calculated in the training dataset and applied to the features in both datasets. To reduce the dimensionality of the data, unsupervised and supervised feature selection was performed on the training dataset. Any feature that failed to extract for any of the patients, for example because a filter was too large to apply to a smaller lesion, was removed. This strategy was adopted since all features selected for the signature need to be applicable to all (future) patients. Any feature with near-zero variance was also removed, as these features do not contain any useful information for a model. Highly correlated features were assumed to contain overlapping information about the outcome, so for each correlating feature pair one was selected and the other was removed. This was done because these features were considered redundant, and to reduce dimensionality and the chance of overfitting. Through absolute pairwise Spearman rank correlation highly correlating features (>0.85) were determined and the feature with the largest mean absolute correlation with the remaining features was removed from the dataset.

3.12.5 Model calibration

The prognostic indices (PI) of the training and validation dataset were determined. The PI, or linear predictor, is the sum of variables x in the model, multiplied by the corresponding regression coefficients β , defined as $\sum_i x_i \beta_i$. To determine the calibration slope, Cox regression was performed on the PI, and the unity value of the slope was tested through a log-rank test. Afterwards, a joint log-rank test on all the predictors plus the offset of the

PI was performed, and tested for non-significance, which would indicate a good fit for our model.

3.12.6 Clinical and biological covariates

The full list of clinical covariates was: age at diagnosis, sex, ACE-27 comorbidity score, smoking pack years, AJCC 8th edition TNM staging, smoking at time of diagnosis (yes/no/former, where former is defined as having stopped before enrolment), and alcohol consumption at time of diagnosis (yes/no/former, where former is defined as having stopped before enrolment). The list of biological covariates was: Hemoglobin (Hb) level, and HPV-status. HPV status was determined by p16 immunostaining followed by HPV DNA PCR for p16 positive cases. P16 positive, but HPV DNA negative cases, were considered HPV negative.

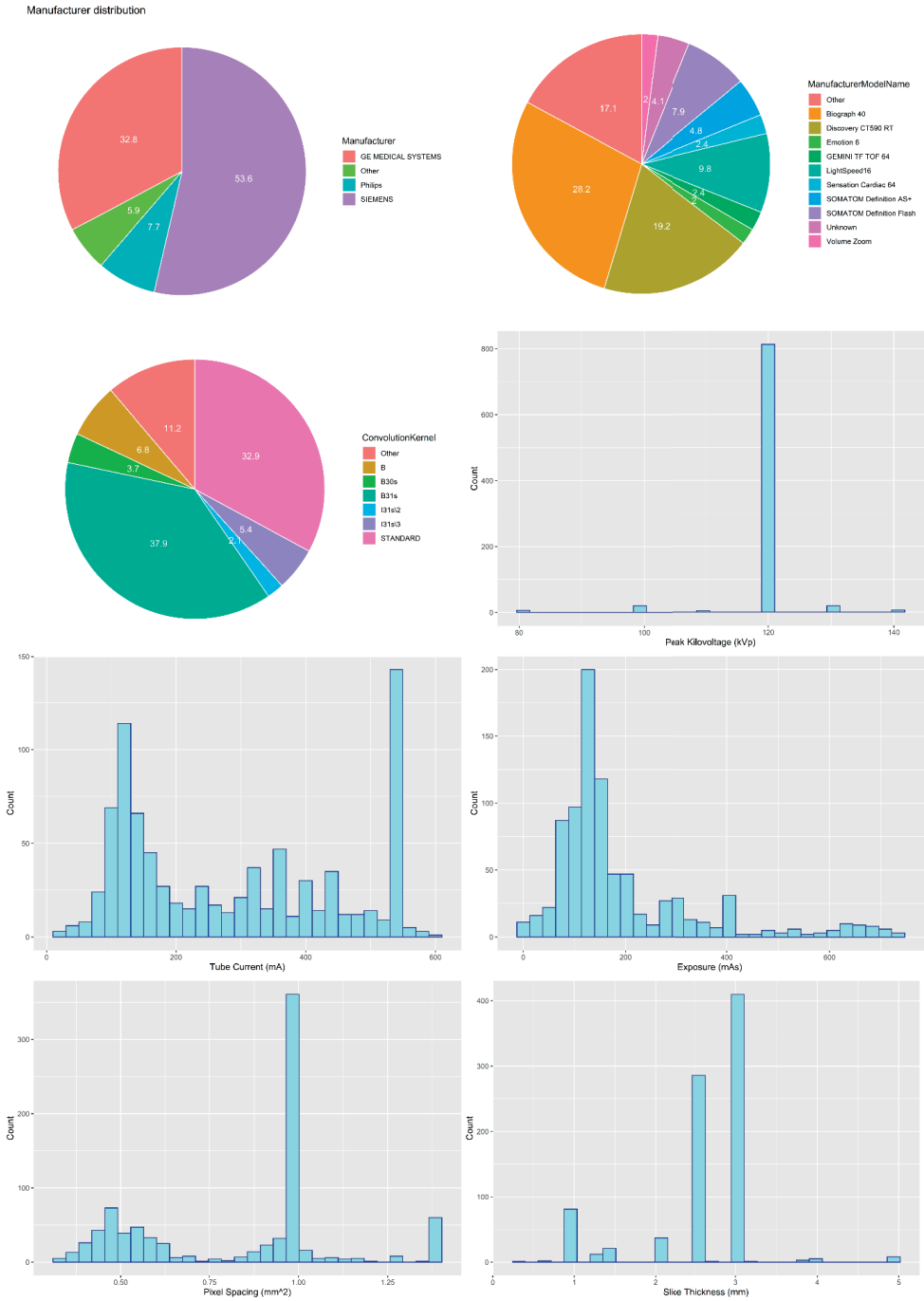


Figure 1. Distributions of imaging acquisition parameters for the full patient population (N=809).

Table 1. A Table of used R packages

Purposes	Functions	Packages	Versions
Spearman's rank correlation	'cor'	'stats'	3.6.3
ROC plots, AUC values, and test	'roc'	'pROC'	1.16.2
Feature selection	'nearZeroVar','uni.selection'	'caret','compound.cox'	6.0-86, 3.19
Cox proportional hazard modelling	'coxph','Surv'	'survival'	3.1.12
Harrel's C-index	'rcorr.cens'	'Hmisc'	4.4.0
Cox Survival Estimates	'survest'	'rms'	5.1.4
Create survival curves	'survfit'	'survest'	3.1.12
Drawing survival curves	'ggsurvplot'	'survminer'	0.4.7
Missing value imputation	'missForest'	'missForest'	1.4

ROC, receiver operating characteristic curve; AUC, area under the roc curve

Table 2. Selected radiomics features for the retrospective training cohort

#	Name feature
1	log.sigma.5.0.mm.3D_glszm_GrayLevelNonUniformity
2	wavelet.HLH_glszm_ZoneEntropy
3	wavelet.HLL_glszm_ZoneEntropy
4	wavelet.LLH_glszm_ZoneEntropy
5	original_shape_Sphericity
6	log.sigma.4.0.mm.3D_gldm_DependenceEntropy
7	wavelet.HHH_glrIm_LowGrayLevelRunEmphasis
8	wavelet.HHL_glszm_ZoneEntropy
9	log.sigma.5.0.mm.3D_gldm_LowGrayLevelEmphasis
10	original_firstorder_Kurtosis
11	log.sigma.2.0.mm.3D_glrIm_RunEntropy

Table 3. Selected radiomics features for the retrospective oropharynx training cohort

#	Name feature
1	original_shape_MajorAxisLength
2	wavelet.HHL_glszm_GrayLevelNonUniformity
3	log.sigma.5.0.mm.3D_glszm_GrayLevelNonUniformity
4	original_shape_Sphericity
5	wavelet.LLH_glszm_ZoneEntropy
6	original_firstorder_Maximum
7	log.sigma.4.0.mm.3D_glrIm_RunEntropy
8	wavelet.HLL_glszm_ZoneEntropy

3.12.7 RQS and TRIPOD

RQS measures metrics of the validity of a radiomics workflow, and the validity of the (external) validation. The RQS consists of 16 components, such as segmentation robustness, comparison to a gold standard, and cost-effectiveness of the clinical application, which together count up to a maximum of 36 points. Similarly, we followed the general procedure recommended in transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD).⁸ This guideline consists of a checklist of 22 points, which cover more general guidelines for articles such as proper reporting and structuring of the article. The RQS score calculated for this study was 75%. A significant portion of points were lost in criterion 11, as we did not apply for a clinical trial to test the signature created in this study. An overview of the point allocation is shown in Table 4. For the TRIPOD statement, and adherence of 76% was calculated. An overview of the point allocation is shown in Table 5.

Table 4. Radiomics quality score table. The table displays the different criteria, the maximum amount of points that can be acquired (or maximum points that can be deducted) and the points calculated in this study.

1	Image protocol quality - well-documented image protocols (for example, contrast, slice thickness, energy, etc.) and/or usage of public image protocols allow reproducibility/replicability	+ 1 (if protocols are well-documented) + 1 (if public protocol is used)	1
2	Multiple segmentations - possible actions are: segmentation by different physicians/algorithms/software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyse feature robustness to segmentation variabilities	1	1
3	Phantom study on all scanners - detect inter-scanner differences and vendor-dependent features. Analyse feature robustness to these sources of variability	1	0
4	Imaging at multiple time points - collect images of individuals at additional time points. Analyse feature robustness to temporal variabilities (for example, organ movement, organ expansion/shrinkage)	1	0
5	Feature reduction or adjustment for multiple testing - decreases the risk of overfitting. Overfitting is inevitable if the number of features exceeds the number of samples. Consider feature robustness when selecting features	- 3 (if neither measure is implemented) + 3 (if either measure is implemented)	3
6	Multivariable analysis with non radiomics features (for example, EGFR mutation) - is expected to provide a more holistic model. Permits correlating/inferencing between radiomics and non radiomics features	1	1
7	Detect and discuss biological correlates - demonstration of phenotypic differences (possibly associated with underlying gene-protein expression patterns) deepens understanding of radiomics and biology	1	1
8	Cut-off analyses - determine risk groups by either the median, a previously published cut-off or report a continuous risk variable. Reduces the risk of reporting overly optimistic results	1	1
9	Discrimination statistics - report discrimination statistics (for example, C-statistic, ROC curve, AUC) and their statistical significance (for example, p-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)	+ 1 (if a discrimination statistic and its statistical significance are reported) + 1 (if a resampling method technique is also applied)	2
10	Calibration statistics - report calibration statistics (for example, Calibration-in-the-large/slope, calibration plots) and their statistical significance (for example, P-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)	+ 1 (if a calibration statistic and its statistical significance are reported) + 1 (if a resampling method technique is also applied)	1

11	Prospective study registered in a trial database - provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker	+ 7 (for prospective validation of a radiomics signature in an appropriate trial)	0
12	Validation - the validation is performed without retraining and without adaptation of the cut-off value, provides crucial information with regard to credible clinical performance	- 5 (if validation is missing) + 2 (if validation is based on a dataset from the same institute) + 3 (if validation is based on a dataset from another institute) + 4 (if validation is based on two datasets from two distinct institutes) + 4 (if the study validates a previously published signature) + 5 (if validation is based on three or more datasets from distinct institutes)	9
13	Comparison to 'gold standard' - assess the extent to which the model agrees with/is superior to the current 'gold standard' method (for example, TNM-staging for survival prediction). This comparison shows the added value of radiomics	2	2
14	Potential clinical utility - report on the current and potential application of the model in a clinical setting (for example, decision curve analysis).	2	2
15	Cost-effectiveness analysis - report on the cost-effectiveness of the clinical application (for example, QALYs generated)	1	0
16	Open science and data - make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study	+ 1 (if scans are open source) + 1 (if region of interest segmentations are open source) + 1 (if code is open source) + 1 (if radiomics features are calculated on a set of representative ROIs and the calculated features and representative ROIs are open source)	3
Total score:		36	27

Table 5. TRIPOD statement checklist form, filled out for the present study.

		Development [D]	External validation [V]	Combined Development & External validation [D+V]
Y=yes; N=no; R=referenced; NA=not applicable				
Title and abstract				
1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.			0
i	The words developing/development, validation/validating, incremental/added value (or synonyms) are reported in the title	N	N	N
ii	The words prediction, risk prediction, prediction model, risk models, prognostic models, prognostic indices, risk scores (or synonyms) are reported in the title	Y	Y	Y
iii	The target population is reported in the title	Y	Y	Y
iv	The outcome to be predicted is reported in the title	Y	Y	Y
2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.			0
i	The objectives are reported in the abstract	Y	Y	Y
ii	Sources of data are reported in the abstract <i>E.g. Prospective cohort, registry data, RCT data.</i>	Y	Y	Y
iii	The setting is reported in the abstract <i>E.g. Primary care, secondary care, general population, adult care, or paediatric care. The setting should be reported for both the development and validation datasets, if applicable.</i>	Y	Y	Y
iv	A general definition of the study participants is reported in the abstract <i>E.g. patients with suspicion of certain disease, patients with a specific disease, or general eligibility criteria.</i>	Y	Y	Y
v	The overall sample size is reported in the abstract	Y	Y	Y
vi	The number of events (or % outcome together with overall sample size) is reported in the abstract <i>If a continuous outcome was studied, score Not applicable (NA).</i>	N	N	N
vii	Predictors included in the final model are reported in the abstract. For validation studies of well-known models, at least the name/acronym of the validated model is reported <i>Broad descriptions are sufficient, e.g. 'all information from patient history and physical examination'. Check in the main text whether all predictors of the final model are indeed reported in the abstract.</i>	Y	Y	Y
viii	The outcome is reported in the abstract	Y	Y	Y

ix	Statistical methods are described in the abstract <i>For model development, at least the type of statistical model should be reported. For validation studies a quote like "model's discrimination and calibration was assessed" is considered adequate. If done, methods of updating should be reported.</i>	Y	Y	Y
x	Results for model discrimination are reported in the abstract <i>This should be reported separately for development and validation if a study includes both development and validation.</i>	Y	Y	Y
xi	Results for model calibration are reported in the abstract <i>This should be reported separately for development and validation if a study includes both development and validation.</i>	N	N	N
xii	Conclusions are reported in the abstract <i>In publications addressing both model development and validation, there is no need for separate conclusions for both; one conclusion is sufficient.</i>	Y	Y	Y
3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.			1
i	The background and rationale are presented	Y	Y	Y
ii	Reference to existing models is included (or stated that there are no existing models)	Y	Y	Y
3b	Specify the objectives, including whether the study describes the development or validation of the model or both.			1
i	It is stated whether the study describes development and/or validation and/or incremental (added) value	Y	Y	Y
Methods				
4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.			1
i	The study design/source of data is described <i>E.g. Prospectively designed, existing cohort, existing RCT, registry/medical records, case control, case series. This needs to be explicitly reported; reference to this information in another article alone is insufficient.</i>	Y	Y	Y
4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.			1
i	The starting date of accrual is reported	Y	Y	Y
ii	The end date of accrual is reported	Y	Y	Y

Table 5. Continued

iii	The length of follow-up <u>and</u> prediction horizon/time frame are reported, if applicable <i>E.g. "Patients were followed from baseline for 10 years" and "10-year prediction of..."; notably for prognostic studies with long term follow-up.</i> <i>If this is not applicable for an article (i.e. diagnostic study or no follow-up), then score Not applicable (NA).</i>	Y	Y	Y
5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.			1
i	The study setting is reported (e.g. primary care, secondary care, general population) <i>E.g.: 'surgery for endometrial cancer patients' is considered to be enough information about the study setting.</i>	Y	Y	Y
ii	The number of centres involved is reported <i>If the number is not reported explicitly, but can be concluded from the name of the centre/centres, or if clearly a single centre study, score Yes.</i>	Y	Y	Y
iii	The geographical location (at least country) of centres involved is reported <i>If no geographical location is specified, but the location can be concluded from the name of the centre(s), score Yes.</i>	Y	Y	Y
5b	Describe eligibility criteria for participants.			1
i	In-/exclusion criteria are stated <i>These should explicitly be stated. Reasons for exclusion only described in a patient flow is not sufficient.</i>	Y	Y	Y
5c	Give details of treatments received, if relevant. <i>(i.e. notably for prognostic studies with long term follow-up)</i>			1
i	Details of any treatments received are described <i>This item is notably for prognostic modelling studies and is about treatment at baseline or during follow-up. The 'if relevant' judgment of treatment requires clinical knowledge and interpretation.</i> <i>If you are certain that treatment was not relevant, e.g. in some diagnostic model studies, score Not applicable.</i>	Y	Y	Y
6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.			1
i	The outcome definition is clearly presented <i>This should be reported separately for development and validation if a publication includes both.</i>	Y	Y	Y
ii	It is described how outcome was assessed (including all elements of any composite, for example CVD [e.g. MI, HF, stroke]).	Y	Y	Y
iii	It is described when the outcome was assessed (time point(s) since T0)	Y	Y	Y
6b	Report any actions to blind assessment of the outcome to be predicted.			0

i	Actions to blind assessment of outcome to be predicted are reported <i>If it is clearly a non-issue (e.g. all-cause mortality or an outcome not requiring interpretation), score Yes. In all other instances, an explicit mention is expected.</i>	N	N	N
7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.			1
i	All predictors are reported <i>For development, "all predictors" refers to all predictors that potentially could have been included in the 'final' model (including those considered in any univariable analyses). For validation, "all predictors" means the predictors in the model being evaluated.</i>	Y	Y	Y
ii	Predictor definitions are clearly presented	Y	Y	Y
iii	It is clearly described how the predictors were measured	Y	Y	Y
iv	It is clearly described when the predictors were measured	Y	Y	Y
7b	Report any actions to blind assessment of predictors for the outcome and other predictors.			0
i	It is clearly described whether predictor assessments were blinded for outcome <i>For predictors for which it is clearly a non-issue (e.g. automatic blood pressure measurement, age, sex) and for instances where the predictors were clearly assessed before outcome assessment, score Yes. For all other predictors an explicit mention is expected.</i>	N	N	N
ii	It is clearly described whether predictor assessments were blinded for the other predictors	N	N	N
8	Explain how the study size was arrived at.			1
i	It is explained how the study size was arrived at <i>Is there any mention of sample size, e.g. whether this was done on statistical grounds or practical/logistical grounds (e.g. an existing study cohort or data set of a RCT was used)?</i>	Y	Y	Y
9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.			1
i	The method for handling missing data (predictors and outcome) is mentioned <i>E.g. Complete case (explicit mention that individuals with missing values have been excluded), single imputation, multiple imputation, mean/median imputation. If there is no missing data, there should be an explicit mention that there is no missing data for all predictors and outcome. If so, score Yes. If it is unclear whether there is missing data (from e.g. the reported methods or results), score No. If it is clear there is missing data, but the method for handling missing data is unclear, score No.</i>	Y	Y	Y

Table 5: Continued

ii	If missing data were imputed, details of the software used are given <i>When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.</i>	Y	Y	Y
iii	If missing data were imputed, a description of which variables were included in the imputation procedure is given <i>When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.</i>	Y	Y	Y
iv	If multiple imputation was used, the number of imputations is reported <i>When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.</i>	Y	Y	Y
10a	Describe how predictors were handled in the analyses.			1
i	For continuous predictors it is described whether they were modelled as linear, nonlinear (type of transformation specified) or categorized <i>A general statement is sufficient, no need to describe this for each predictor separately. If no continuous predictors were reported, score Not applicable.</i>	NA	Not applicable	NA
ii	For categorical or categorized predictors, the cut-points were reported <i>If no categorical or categorized predictors were reported, score Not applicable.</i>	NA	Not applicable	NA
iii	For categorized predictors the method to choose the cut-points was clearly described <i>If no categorized predictors, score Not applicable.</i>	NA	Not applicable	NA
10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.			0
i	The type of statistical model is reported <i>E.g. Logistic, Cox, other regression model (e.g. Weibull, ordinal), other statistical modelling (e.g. neural network)</i>	Y	Not applicable	Y
ii	The approach used for predictor selection <u>before</u> modelling is described <i>'Before modelling' means before any univariable or multivariable analysis of predictor-outcome associations. If no predictor selection before modelling is done, score Not applicable. If it is unclear whether predictor selection before modelling is done, score No. If it is clear there was predictor selection before modelling but the method was not described, score No.</i>	Y	Not applicable	Y

iii	The approach used for predictor selection during modelling is described <i>E.g. Univariable analysis, stepwise selection, bootstrap, Lasso.</i> <i>'During modelling' includes both univariable or multivariable analysis of predictor-outcome associations. If no predictor selection during modelling is done (so-called full model approach), score Not applicable. If it is unclear whether predictor selection during modelling is done, score No. If it is clear there was predictor selection during modelling but the method was not described, score No.</i>	Y	Not applicable	Y
iv	Testing of interaction terms is described <i>If it is explicitly mentioned that interaction terms were not addressed in the prediction model, score Yes. If interaction terms were included in the prediction model, but the testing is not described, score No.</i>	N	Not applicable	N
v	Testing of the proportionality of hazards in survival models is described <i>If no proportional hazard model is used, score Not applicable.</i>	Y	Not applicable	Y
vi	Internal validation is reported <i>E.g. Bootstrapping, cross validation, split sample. If the use of internal validation is clearly a non-issue (e.g. in case of very large data sets), score Yes. For all other situations an explicit mention is expected.</i>	Y	Not applicable	Y
10c	For validation, describe how the predictions were calculated.			1
i.	It is described how predictions for individuals (in the validation set) were obtained from the model being validated <i>E.g. Using the original reported model coefficients with or without the intercept, and/or using updated or refitted model coefficients, or using a nomogram, spreadsheet or web calculator.</i>	Not applicable	Y	Y
10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models. <i>These should be described in methods section of the paper (item 16 addresses the reporting of the results for model performance).</i>			1
i	Measures for model discrimination are described <i>E.g. C-index / area under the ROC curve.</i>	Y	Y	Y
ii	Measures for model calibration are described <i>E.g. calibration plot, calibration slope or intercept, calibration table, Hosmer Lemeshow test, O/E ratio.</i>	Y	Y	Y
iii	Other performance measures are described <i>E.g. R2, Brier score, predictive values, sensitivity, specificity, AUC difference, decision curve analysis, net reclassification improvement, integrated discrimination improvement, AIC.</i>	Y	Y	Y
10e	Describe any model updating (e.g., recalibration) arising from the validation, if done.			Not applicable

Table 5. Continued

i	A description of model-updating is given <i>E.g. Intercept recalibration, regression coefficient recalibration, refitting the whole model, adding a new predictor</i> <i>If updating was done, it should be clear which updating method was applied to score Yes.</i> <i>If it is not explicitly mentioned that updating was applied in the study, score this item as 'Not applicable'.</i>	Not applicable	NA	NA
11	Provide details on how risk groups were created, if done. <i>If risk groups were not created, score this item as Yes.</i>			1
i	If risk groups were created, risk group boundaries (risk thresholds) are specified <i>Score this item separately for development and validation if a study includes both development and validation.</i> <i>If risk groups were not created, score this item as not applicable.</i>	Y	Y	Y
12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome and predictors.			0
i	Differences or similarities in <u>definitions</u> with the development study are described <i>Mentioning of any differences in all four (setting, eligibility criteria, predictors and outcome) is required to score Yes.</i> <i>If it is explicitly mentioned that there were no differences in setting, eligibility criteria, predictors and outcomes, score Yes.</i>	Not applicable	N	N
13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.			1
i	The flow of participants is reported	Y	Y	Y
ii	The number of participants with and without the outcome are reported <i>If outcomes are continuous, score Not applicable.</i>	Y	Y	Y
iii	A summary of follow-up time is presented <i>This notably applies to prognosis studies and diagnostic studies with follow-up as diagnostic outcome.</i> <i>If this is not applicable for an article (i.e. diagnostic study or no follow-up), then score Not applicable.</i>	Y	Y	Y
13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.			1
i	Basic demographics are reported	Y	Y	Y
ii	Summary information is provided for all predictors included in the final developed/validated model	Y	Y	Y
iii	The number of participants with missing data for predictors is reported	Y	Y	Y

iv	The number of participants with missing data for the outcome is reported	Y	Y	Y
13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).			1
i	Demographic characteristics (at least age and gender) of the validation study participants are reported along with those of the original development study	Not applicable	Y	Y
ii	Distributions of predictors in the model of the validation study participants are reported along with those of the original development study	Not applicable	Y	Y
iii	Outcomes of the validation study participants are reported along with those of the original development study	Not applicable	Y	Y
14a	Specify the number of participants and outcome events in each analysis.			1
i	The number of participants in each analysis (e.g. in the analysis of each model if more than one model is developed) is specified	Y	Not applicable	Y
ii	The number of outcome events in each analysis is specified (e.g. in the analysis of each model if more than one model is developed) <i>If outcomes are continuous, score Not applicable.</i>	Y	Not applicable	Y
14b	If done, report the unadjusted association between each candidate predictor and outcome.			0
i	The unadjusted associations between each predictor and outcome are reported <i>If any univariable analysis is mentioned in the methods but not in the results, score No.</i> <i>If nothing on univariable analysis (in methods or results) is reported, score this item as Not applicable.</i>	N	Not applicable	N
15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).			1
i	The regression coefficient (or a derivative such as hazard ratio, odds ratio, risk ratio) for each predictor in the model is reported	Y	Not applicable	Y
ii	The intercept or the cumulative baseline hazard (or baseline survival) for at least one time point is reported	Y	Not applicable	Y
15b	Explain how to use the prediction model.			0
i	An explanation (e.g. a simplified scoring rule, chart, nomogram of the model, reference to online calculator, or worked example) is provided to explain how to use the model for individualised predictions.	N	Not applicable	N
16	Report performance measures (with confidence intervals) for the prediction model. <i>These should be described in results section of the paper (item 10 addresses the reporting of the methods for model performance).</i>			1
i	A discrimination measure is presented <i>E.g. C-index / area under the ROC curve.</i>	Y	Y	Y

Table 5. Continued

ii	The confidence interval (or standard error) of the discrimination measure is presented	Y	Y	Y
iii	Measures for model calibration are described <i>E.g. calibration plot, calibration slope or intercept, calibration table, Hosmer Lemeshow test, O/E ratio.</i>	Y	Y	Y
iv	Other model performance measures are presented <i>E.g. R2, Brier score, predictive values, sensitivity, specificity, AUC difference, decision curve analysis, net reclassification improvement, integrated discrimination improvement, AIC.</i>	Y	Y	Y
17	If done, report the results from any model updating (i.e., model specification, model performance, recalibration). <i>If updating was not done, score this TRIPOD item as 'Not applicable.'</i>			Not applicable
0	Model updating was done <i>If "No", then answer 17i-17v with "Not applicable"</i>	Not applicable	N	N
i	The updated regression coefficients for each predictor in the model are reported <i>If model updating was described as 'not needed', score Yes.</i>	Not applicable	NA	NA
ii	The updated intercept or cumulative baseline hazard or baseline survival (for at least one time point) is reported <i>If model updating was described as 'not needed', score Yes.</i>	Not applicable	NA	NA
iii	The discrimination of the updated model is reported	Not applicable	NA	NA
iv	The confidence interval (or standard error) of the discrimination measure of the updated model is reported	Not applicable	NA	NA
v	The calibration of the updated model is reported	Not applicable	NA	NA
18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).			1
i	Limitations of the study are discussed <i>Stating any limitation is sufficient.</i>	Y	Y	Y
19a	For validation, discuss the results with reference to performance in the development data, and any other validation data.			1
i	Comparison of results to reported performance in development studies and/or other validation studies is given	Not applicable	Y	Y
19b	Give an overall interpretation of the results considering objectives, limitations, results from similar studies and other relevant evidence.			1
i	An overall interpretation of the results is given	Y	Y	Y
20	Discuss the potential clinical use of the model and implications for future research.			1

i	The potential clinical use is discussed <i>E.g. an explicit description of the context in which the prediction model is to be used (e.g. to identify high risk groups to help direct treatment, or to triage patients for referral to subsequent care).</i>	Y	Y	Y
ii	Implications for future research are discussed <i>E.g. a description of what the next stage of investigation of the prediction model should be, such as "We suggest further external validation".</i>	Y	Y	Y
21	Provide information about the availability of supplementary resources, such as study protocol, web calculator, and data sets.			
i	Information about supplementary resources is provided	Y	Y	Y
22	Give the source of funding and the role of the funders for the present study.			1
i	The source of funding is reported or there is explicit mention that there was no external funding involved	Y	Y	Y
ii	The role of funders is reported or there is explicit mention that there was no external funding	Y	Y	Y
Number of applicable TRIPOD items				34
Number of TRIPOD items adhered				26
OVERALL adherence to TRIPOD				76%

References

1. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*. 1973;**SMC-3**(6):610-21.
2. Galloway MM. Texture analysis using grey level run lengths. 1974 July 01, 1974.
3. Thibault G, Fertil B, Navarro C, Pereira S, Lévy N, Sequeira J, et al. Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification 2009.
4. Thibault G, Angulo J, Meyer F, editors. Advanced statistical matrices for texture characterization: Application to DNA chromatin and microtubule network classification. 2011 18th IEEE International Conference on Image Processing; 2011 11-14 Sept. 2011.
5. Sun C, Wee WG. Neighboring gray level dependence matrix for texture classification. *Computer Vision, Graphics, and Image Processing*. 1983;**23**(3):341-52.
6. Amadasun M, King R. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics*. 1989;**19**(5):1264-74.
7. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med Phys*. 2017;**44**(3):1050-62.
8. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med*. 2015;**162**(10):735-6.

CHAPTER 4



Computed tomography-derived radiomic signature of head and neck squamous cell carcinoma (peri) tumoral tissue for the prediction of locoregional recurrence and distant metastasis after concurrent chemo-radiotherapy

Simon Keek^{1*}, Sebastian Sanduleanu^{1*}, Frederik Wesseling², Reinout de Roest³, Michiel van den Brekel^{4,5}, Martijn van der Heijden^{6,7}, Conchita Vens^{8,9}, Calareso Giuseppina¹⁰, Lisa Licitra^{11,12}, Kathrin Scheckenbach¹³, Marije Vergeer¹⁴, C. René Leemans³, Ruud H Brakenhoff³, Irene Nauta³, Stefano Cavalieri⁷, Henry C. Woodruff^{1,15}, Tito Poli¹⁶, Ralph Leijenaar², Frank Hoebbers^{2*}, Philippe Lambin^{1,15*}

*Contributed equally to the article

1 The D-lab, Department of Precision Medicine, GROW – School for Oncology and Developmental Biology, Maastricht University, Maastricht, The Netherlands

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

3 Amsterdam UMC, Vrije Universiteit Amsterdam, Otolaryngology / Head and Neck Surgery, Cancer Center Amsterdam, The Netherlands

4 Department of Head and Neck Oncology and Surgery, The Netherlands Cancer Institute, Amsterdam, The Netherlands

5 Department of Oral and Maxillofacial Surgery, Academic Medical Center, Amsterdam, The Netherlands

6 Department of Head and Neck Oncology and Surgery, The Netherlands Cancer Institute, Amsterdam, The Netherlands

7 Division of Cell Biology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

8 Division of Cell Biology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

9 Department of radiation oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

10 Istituto nazionale dei tumori, Department of Radiology, Milan, Italy

11 Fondazione IRCCS Istituto Nazionale dei Tumori, Head and Neck Medical Oncology Department, Milan, Italy

12 University of Milan, Department of Oncology and Hematology-Oncology, Milan, Italy

13 Dept. of Otorhinolaryngology, Head and Neck Surgery, Heinrich-Heine-University, Düsseldorf, Germany

14 Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Radiation Oncology Amsterdam, The Netherlands

15 Department of Radiology and Nuclear Medicine, GROW – School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands

16 University of Parma, Department of Surgical Sciences, Parma, Italy

4.1 Abstract

4.1.1 Introduction

In this study, we investigate the role of radiomics for prediction of overall survival (OS), locoregional recurrence (LRR) and distant metastases (DM) in stage III and IV HNSCC patients treated by chemoradiotherapy. We hypothesize that radiomic analysis of (peri-) tumoral tissue may detect invasion of surrounding tissues indicating a higher chance of locoregional recurrence and distant metastasis.

4.1.2 Methods

Two comprehensive data sources were used: the Dutch Cancer Society Database (Alp 7072, DESIGN) and “Big Data To Decide” (BD2Decide). The gross tumor volumes (GTV) were delineated on contrast-enhanced CT. Radiomic features were extracted using the RadiomiX Discovery Toolbox (OncoRadiomics, Liege, Belgium). Clinical patient features such as age, gender, performance status etc. were collected. Two machine learning methods were chosen for their ability to handle censored data: Cox proportional hazards regression and random survival forest (RSF). Multivariable clinical and radiomic Cox/RSF models were generated based on significance in univariable cox regression/RSF analyses on the held out data in the training dataset. Features were selected according to a decreasing hazard ratio for Cox and relative importance for RSF.

4.1.3 Results

A total of 444 patients with radiotherapy planning CT-scans were included in this study: 301 head and neck squamous cell carcinoma (HNSCC) patients in the training cohort (DESIGN) and 143 patients in the validation cohort (BD2DECIDE). We found that the highest performing model was a clinical model that was able to predict distant metastasis in oropharyngeal cancer cases with an external validation C-index of 0.74 and 0.65 with the RSF and Cox models respectively. Peritumoral radiomics based prediction models performed poorly in the external validation, with C-index values ranging from 0.32 to 0.61 utilizing both feature selection and model generation methods.

4.1.4 Conclusion

Our results suggest that radiomic features from the peritumoral regions are not useful for the prediction of time to OS, LR and DM.

4.2 Introduction

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common malignant disease worldwide [1]. In the Netherlands, approximately 39,000 men and women were diagnosed with HNSCC between 2000 and 2015 [2]. Roughly two thirds of patients have advanced stage of disease at diagnosis with debilitating symptoms.

Major progress has been made in the treatment of advanced HNSCC throughout the last decade [6]. The “traditional” treatment of these advanced tumors consists of surgical excision followed by complementary (adjuvant) radiotherapy or chemoradiotherapy (CRT). CRT either applied upfront or postoperatively significantly improves survival in HNSCC patients with overall 5-year survival rates up to 61%, 41%, and 69% for oral, pharyngeal and laryngeal cancers, respectively [3–6]. The introduction of organ-preserving therapies (induction chemotherapy, upfront concomitant CRT, or molecular targeted drugs such as cetuximab) has notably changed treatment protocols of advanced stage HNSCC patients, especially in patients where surgical resection is considered too invasive and where severe problems with speech and swallowing are expected after surgery. Concomitant CRT consists of systemic administration of cisplatin in combination with locoregional radiotherapy and is the mainstay of organ-preserving treatment for advanced HNSCC.

It has been shown that 40% of patients treated upfront with CRT develop a locoregional recurrence or distant metastasis within 2 years after treatment and consequently have an unfavourable prognosis [7]. Several studies have found that advanced and human papillomavirus (HPV)-16-negative tumors respond poorly to CRT in contrast to HPV positive tumors, in particular in oropharyngeal HNSCC [4, 8]. TNM classifications are expected to support patient prognosis by clinicians but unfortunately, they are not helpful to accurately predict which HNSCC patients treated with CRT will develop locoregional recurrences and hence might have benefited from alternative treatment options. Several other potentially prognostic factors have been proposed, such as chemotherapy dose, radiotherapy dose, co-morbidity, World Health Organization (WHO) Performance Status (PS), and HPV-status. Through the use of machine learning algorithms, complex survival models can be created that take these clinical factors into account, while accounting for e.g. interaction between the predictors and right censored data [9].

Currently used biomarkers comprise tumor size, local tumor extent and a few molecular markers (e.g. p16 staining or HPV-PCR). Radiologic imaging, which is routinely performed prior to initiation of CRT, provides an additional source of information that can be exploited through the use of advanced image analysis methods such as radiomics. Radiomics turns radiographic images into a high-throughput data-mining format. The format of the extracted data is a set of features, including first-order intensity histogram statistics,

shape- and size statistics, and (filtered) texture features. Complex models that combine radiomics with clinical parameters may be better in detecting HNSCC patients that have a higher likelihood to relapse early after CRT [10].

A growing body of research shows that the tumor microenvironment is a key player in head and neck cancer development and progression [11,12] and hence the immediate surroundings of the tumor may be a source for the extraction of imaging biomarkers. One of the hypotheses is that information about underlying malignancy-associated changes (MAC's) in the tumor microenvironment can be detected by these imaging biomarkers. These MAC changes are subtle changes in the nuclear morphology and chromatin structure of seemingly normal cells located within the stroma distally to neoplastic lesions that have been shown to dictate its ability to grow and spread, evade the body's immune defenses, and resist therapeutic intervention [13].

In this study, we aim to investigate the role of radiomics for prediction of overall survival (OS), locoregional recurrence (LRR) and distant metastasis (DM) in stage III and IV HNSCC patients, both in a HPV-negative oropharyngeal cohort (high risk) as well as in the general HNSCC population. We hypothesize that radiomic analysis of peritumoral tissue detects changes associated with malignancy and therefore the likelihood of locoregional recurrence and distant metastasis following CRT.

4.3 Methods

4.3.1 Patient characteristics

Two sources of clinical and imaging data were available to us for this study: the Dutch Cancer Society Database (Alp 7072, acronym DESIGN) and "Big Data To Decide" (BD2Decide, NCT02832102). DESIGN is a Dutch multicenter clinical study to create predictive models for stage III and IV HPV-negative HNSCC patients treated by CRT. BD2Decide is a European multicenter clinical study to improve clinical decision making in stage III and IV HNSCC patients irrespective of treatment. In the present study, we included patients from both consortiums with pathologically-confirmed HNSCC, who received contrast-enhanced pre-treatment CT and have been treated upfront with CRT.

The DESIGN data consists of contrast enhanced CT images (and associated clinical data) acquired from 4 different centers: Amsterdam UMC location VUmc, Netherlands Cancer Institute (NKI), Maastricht Radiation Oncology Clinic (MAASTRO), and the University Medical Center Utrecht (UMCU). The BD2Decide data consists of contrast-enhanced CT images retrospectively acquired from 4 different centers: Fondazione IRCCS Istituto dei Tumori Milano (INT), Maastricht Radiation Oncology Clinic (MAASTRO), Amsterdam UMC,

location VUMc (VUMC), and the Heinrich-Heine-university in Dusseldorf. There were no over-lapping patients between DESIGN and BD2DECIDE.

Both DESIGN and BD2Decide data included clinical, pathological, radiologic imaging, and molecular markers for each case. After comparing datasets, a selection was made to include patients based on the overlap of available clinical data between the two cohorts. These consist of age, sex, performance status, ACE-27 baseline comorbidity, number of pack years, alcohol consumption, hemoglobin at baseline, chemotherapy regimen, HPV status (defined as p16-status) for oropharyngeal cancer, induction chemotherapy (yes/no), chemotherapy completion (yes/no), and RT dose to the high-risk clinical target volume (HR-CTV).

4.3.2 CT acquisition parameters and segmentation

Patients were selected according to the following inclusion criteria: (i) concomitant CRT of unresected HNSCC, (ii) hypopharyngeal, laryngeal or (HPV-negative on p16 staining) oropharyngeal, (iii) no prior treatment with chemotherapy or with radiotherapy in the head and neck area, (iv) availability of contrast-enhanced baseline planning CT imaging with a slice thickness ≤ 5 mm and artifacts in less than 50% of the GTV slices, and (v) availability of patient outcome data for OS, LRR, and DM. A large selection of different scanners were used to acquire the images (S1 Appendix).

GTVs were delineated in each center by an assigned radiation oncologist or radiologist. All contours were revised by a radiation oncologist with over 18 years experience, using MIM soft-ware version 6.9.0 (MIM, Cleveland, United States).

Tumor border regions of interest (ROI) extending 3mm and 5mm from the 3D GTV border were generated in MIM (outer ring expansion, see Fig 1). Afterwards, air and bone were filtered from the delineation by setting minimum and maximum thresholds, and manually adjusting the final ROI's border (peritumoral) regions.

4.3.3 Ethical approval

This study was performed following the guidelines of the Code of Conduct for Human Tissue and Medical Research (<https://www.federa.org/codes-conduct>) and the EU General Data Protection Regulation. Medical Ethics Committee approval was provided by the individual centers (full list provided in S2 Appendix). Written informed consent was given and was placed under the responsibility of the Principal Investigators of the relevant Clinical Participating Centers mentioned above and remain under the custodianship of the specific Participating Centers. For reproducibility purposes, our code can be found on: <https://github.com/PeritumoralRadiomics/Peritumoral-radiomics-HN.git>

4.3.4 Clinical outcome

The clinical endpoints evaluated in this study were overall survival (OS), locoregional recurrence (LRR) and distant metastasis (DM). The missForest (non-parametric missing value imputation using Random Forest) function within the R environment (<https://www.R-project.org/>) was used to impute missing data. Time to OS was defined as the time between CRT start date and date of death, or censored at the last follow-up date

Time to LRR was defined as the time between CRT start date and the first scan date of radiologically evident local or regional recurrence (event), or censored at the last follow-up date or date of death.

Time to DM was defined as the time between CRT start date and the first scan date of radiologically evident distant metastasis, or censored at the last follow-up date or date of death.

4.3.5 Image pre-processing, radiomic feature extraction and feature harmonization

International Biomarker Standardization Initiative (IBSI)-compliant radiomic features as well as other non-IBSI covered features were extracted with our in-house RadiomiX research soft-ware (supported by Oncoradiomics, Liège, Belgium) implemented in Matlab 2017a (Math-works, Natick, Mass). Hounsfield Unit (HU) intensities beyond -1024 and +3071 HU were clipped (assigned the value -1024 and +3071 respectively). An image intensity discretization applying a fixed bin width of 25HU was used for feature extraction in CT. Voxel size resampling was performed before feature extraction using cubic interpolation. Images were resampled to isotropic voxels of size 3 x 3 x 3 mm³ using cubic interpolation (upsampling to highest slice thickness).

Radiomic features were extracted consisting of five main groups: 1) fractal features 2) first order statistics, 3) shape and size, 4) texture descriptors including gray level co-occurrence (GLCM), gray level run-length (GLRLM) and gray level size-zone texture matrices (GLSZM), 5) features from groups 1, 3 and 4 after wavelet decomposition of the original image. There were no missing feature values. Definitions and detailed feature descriptions are described elsewhere [14].

Radiomic feature values are potentially sensitive to inter-scanner model, acquisition protocol and reconstruction settings variations. The ComBat statistical feature harmonization technique was employed in our analysis. This technique was initially developed by Johnson et al. [15] for gene expression microarray data (even for small sample sizes) and was recently applied in multicenter PET, MRI and CT radiomic studies [16,17]. Feature values were adjusted for the batch effect according to treatment center, without adjustment for

other covariates. Finally, features were normalized in the training dataset by the mean and standard deviation, which were subsequently used to normalize the validation dataset.

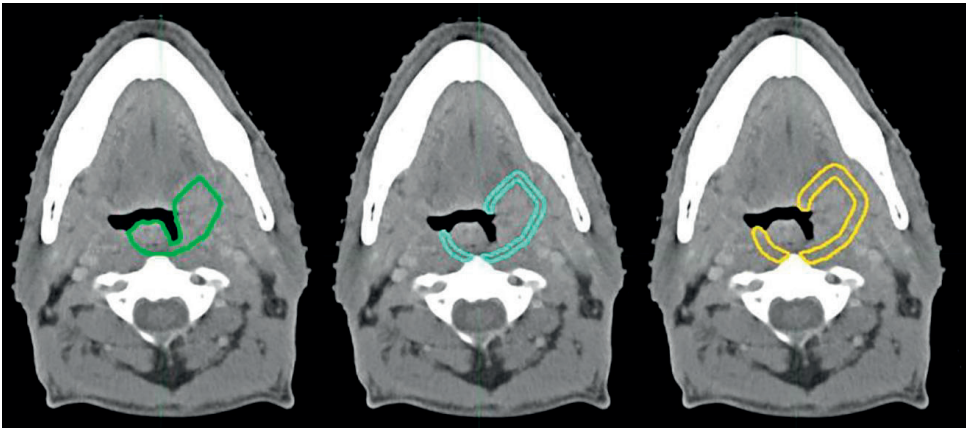


Figure 1. Contrast-enhanced CT image from an oropharyngeal cancer patient. Primary gross tumor volume (GTV1) border in green, blue: 3mm peritumoral border, yellow: 5mm peritumoral border.

4.3.6 Univariable analysis and generation of multivariable models

The prognostic value of the individual radiomic and clinical features was evaluated using concordance index (CI) with the survival package (Therneau T (2015). A Package for Survival Analysis in R. version 2.38, URL: <https://CRAN.R-project.org/package=survival>) and random-ForestSRC package (Ishwaran H (2017) Fast Unified Random Forests for Survival, Regression and Classification (RF-SRC) version 2.9.1, URL: <https://cran.r-project.org/web/packages/randomForestSR/>).

Noether's method was applied to assess the statistical significance of the computed CI from random chance ($CI = 0.5$) with the survcomp package (Benjamin Haibe-Kains (2017). Performance Assessment and Comparison for Survival Analysis in R. version 1.36.0, URL: <https://www.pmggenomics.ca/bhklab/>). To account for multiple testing, a false-discovery-rate (FDR) procedure by Benjamin and Hochberg was applied to adjust the p-values in univariate Cox-regression. Two machine learning methods were employed that are able to use censored survival data as inputs: Cox proportional hazards based and random survival forest (RSF). Multivariable radiomic Cox models were generated using the significant features selected through univariate cox modelling on the training dataset. In a 100-repeat 2-fold cross-validation on the training data, significant features were selected based on univariate significance ($p < 0.05$) adjusted for multiple testing. "These features were then ranked according to adjusted hazard ratios, where hazard ratios lower than 1 were inverted, and were gradually added to a multivariate cox model until the first peak in the cross-validation testing C-index or after the first peak until the C-index drops by

more than 0.02, depending if there is an oscillation or noise pattern leading to multiple peaks. The number of occurrences of each feature in all repetitions was determined, and a selection rate > 50% was used as threshold for the final set of features, ensuring that the selected features were chosen in the majority of the models." Multivariable clinical models included features selected through Cox-regression based on univariate significance ($p < 0.05$) adjusted for multiple testing. The selected clinical features were then used to train multivariable Cox or RSF models. Multivariable clinical RSF models were generated based on selecting all features with a relative feature importance > 0 in the Random Survival Forest. RSF strictly adheres to the prescription laid out by Breiman (2003) and requires taking into account the outcome (splitting criterion used in growing a tree must explicitly involve survival time and censoring information) in growing a random forest model. Further, the predicted value for a terminal node in a tree, the resulting ensemble predicted value from the forest, and the measure of prediction accuracy must all properly incorporate survival information. Multivariable radiomic RSF models were generated based on the optimal number of features corresponding to the first peak in C-index value in the out-of-bag cases OR after the first peak until the C-index drops by more than 0.02, depending if there is an oscillation or noise pattern leading to multiple peaks. Hereby features with decreasing relative importance in the Random Survival Forest were consecutively added.

4.4 Results

4.4.1 Clinical characteristics

Contrast enhanced CT images from a total of 444 patients were included in this study: The training cohort (DESIGN) consisted of 301 head and neck squamous cell carcinoma (HNSCC) patients and the validation cohort (BD2DECIDE) of 143 patients. At time of diagnosis, the median age in the training cohort (DESIGN) was 61 years (range: 36 to 80 years), while the median age in the external validation cohort (BD2DECIDE) was 60.5 years (range: 41 to 78 years).

In the training dataset the median OS time was 1118 days, the median time to LRR or last follow-up was 1042 days and the median time to DM or last follow-up was 1060 days. In the external validation dataset the median time to death or last follow-up was 1268 days, the median time to LRR or last follow-up was 1217 days and the median time to DM or last follow-up was 1189 days.

The full list of patient characteristics and time to progression is presented in Table 1.

Table 1. DESIGN/ BD2DECIDE patient characteristics.

	DESIGN training cohort (n = 301)	BD2DECIDE validation cohort (n = 143)	P-value
	Median (range)	Median (range)	
GTV_{prim} Volume (cm³)	21.28 (0.65-176.10)	19.82 (0.54-157.28)	0.82
Age (years)	61 (36-80)	60 (41-78)	0.52
	Number of pts (%)	Number of pts (%)	
WHO PS			<0.001
0	0 (0)	120 (83.9)	
1	79 (26.2)	20 (14.0)	
2	139 (46.2)	3 (2.1)	
3	10 (3.3)	0 (0)	
Missing	73 (24.3)	0 (0)	
Clinical TNM (T), 7th Edition			0.08
cTX	0 (0)	0 (0)	
cT1	14 (4.7)	3 (2.1)	
cT2	63 (20.9)	25 (17.5)	
cT3	106 (35.2)	68 (47.6)	
cT4	118 (39.2)	47 (32.9)	
Clinical Nodal stage (N), 7th Edition			0.01
cNX	1 (0.3)	0 (0)	
cN0	41 (13.6)	37 (25.9)	
cN1	41 (13.6)	19 (13.3)	
cN2 a-b-c	209 (69.5)	79 (55.2)	
cN3	9 (3.0)	8 (5.6)	
HPV status (P16 stain)			<0.001
Negative	207 (68.8)	64 (44.8)	
Positive/ Unknown	94 (31.2)	79 (55.2)	
Treatment			
Chemotherapy regimen			<0.001
Platin	292 (97.0)	81 (56.6)	
Platin + others	9 (3.0)	23 (16.1)	
Cetuximab	0 (0)	39 (27.3)	
Cumulative radiotherapy dose high-risk CTV	70 (60-84) Gy	70 (20-76) Gy	
Tumor site			
Oropharynx	145 (48.2)	49 (34.3)	0.02
Larynx	57 (18.9)	39 (27.3)	
Hypopharynx	99 (32.9)	55 (38.5)	

Clinical models (Tables 2 and 3) to predict OS, LR and DM ranged from a C-index of 0.61–0.85 in training with both methods and a C-index of 0.49–0.75 in external validation.

Table 2. Multivariable Cox Regression method, C-index and number of radiomic and (non)-treatment related prognostic clinical factors in validation dataset (BD2DECIDE).

	C-index Prognostic (No. feat)		C-index GTV _{prim} (No. feat)		C-index TB 3mm (No.feats)		C-index TB 5mm (No. feat)		C-index GTV _{prim} + TB 3mm + TB 5mm (No. feat)	
	Train	Val	Train	Val	Train	Val	Train	Val	Train	Val
Oropharynx										
Clinical-OS	0.61 (1)	0.49 (1)								
Clinical-LR	0.61 (1)	0.55 (1)								
Clinical-DM	0,67 (1)	0.65 (1)								
Radiomics-OS			0.65 (3)	0.57 (3)	0.69 (3)	0.52 (3)	0.79 (1)	0.60 (1)	0.70 (2)	0.56 (2)
Radiomics-LR			0.57 (1)	0.52 (1)	0.70 (2)	0.56 (2)	0.76 (6)	0.51(6)	0.72 (4)	0.48 (4)
Radiomics-DM			-	-	0.69 (2)	0.61 (2)	0.73 (3)	0.44 (3)	0.72 (2)	0.60 (2)
All subsites										
Clinical-OS	0.64 (4)	0.56 (4)								
Clinical-LR	-	-								
Clinical-DM	0.67 (1)	0.49 (1)								
Radiomics-OS			0.61 (1)	0.60 (1)	0.63 (4)	0.61 (4)	0.61 (2)	0.62 (2)	0.61 (3)	0.59 (3)
Radiomics-LR			0.66 (3)	0.51 (3)	0.67 (3)	0.51 (3)	0.58 (1)	0.47 (1)	0.61 (1)	0.47 (1)
Radiomics-DM			0.63 (2)	0.54 (2)	0.54 (2)	0,47 (4)	0.61 (2)	0.56 (2)	0.64 (3)	0.55(2)

Table 3. Random survival forest method, C-index and number of radiomic or (non)-treatment related prognostic clinical factors. Abbreviations GTV_{prim} – Primary Gross Tumor Volume, OS – Overall Survival, LR – Local recurrence, DM – Distant Metastasis

	C-index Prognostic (No. feat)		C-index Treatment (No. feat)		C-index GTV _{prim} (No. feat)		C-index TB 3mm (No. feat)		C-index TB 5mm (No. feat)		C-index GTV _{prim} + TB 5mm + TB 3mm (No. feat)	
	Train	Val	Train	Val	Train	Val	Train	Val	Train	Val	Train	Val
Oropharynx												
Clinical-OS	0.74 (5)	0.74 (5)	0.52 (1)	0.53 (1)								
Clinical-LR	0.81 (5)	0.81 (5)	0.51 (1)	0.51 (1)								
Clinical-DM	0.85 (4)	0.85 (4)	0.51 (1)	0.52 (1)								
Radiomics-OS					0.73 (3)	0.58 (3)	0.77 (6)	0.49 (6)	0.79 (5)	0.60 (5)	0.78 (6)	0.61 (6)
Radiomics-LR					0.77 (2)	0.49 (2)	0.83 (3)	0.43 (3)	0.83 (2)	0.59 (7)	0.71 (3)	0.57 (3)
Radiomics-DM					0.82 (2)	0.49 (2)	0.91 (8)	0.55 (8)	0.81 (3)	0.50 (3)	0.86 (4)	0.32 (4)
All subsites												
Clinical-OS	0.77 (7)	0.77 (7)	0.56 (1)	0.51 (1)								
Clinical-LR	0.79 (3)	0.79 (3)	0.56 (2)	0.49 (2)								
Clinical-DM	0.84 (4)	0.84 (4)	-	-								
Radiomics-OS					0.79 (7)	0.58 (4)	0.89 (4)	0.58 (4)	0.77 (5)	0.60 (5)	0.78 (6)	0.59 (6)
Radiomics-LR					0.81 (3)	0.52 (2)	0.80 (2)	0.52 (2)	0.86 (7)	0.59 (7)	0.83 (2)	0.53 (2)
Radiomics-DM					0.86 (3)	0.49 (3)	0.86 (4)	0.49 (4)	0.96 (3)	0.50 (3)	0.86 (3)	0.43 (3)
Ab												

Details on the clinical variable selected in the final Cox/ RSF models are presented in Table 4.

Table 4. Multivariable clinical Cox/ RSF models.

Outcome	Clinical Cox, all subsites Prognostic	Clinical Cox, Oropharynx Prognostic	Clinical RSF, all subsites Prognostic	Clinical RSF, Oropharynx Prognostic	Clinical RSF, all subsites Treatment	Clinical RSF, Oropharynx Treatment
OS	N-stage	N-stage	N-stage	N-stage	Chemotherapy regimen	Chemotherapy regimen
	Tumor site		Tumor site	Age	Chemotherapy completion	
	Gender		Hb baseline	Pack Years		
	Alcohol consumption		Age	Alcohol consumption		
			Pack-years	Gender		
LR	N-stage	Gender	Hb baseline	Gender	Chemotherapy regimen	Chemotherapy regimen
			Tumor site	Alcohol consumption	Chemotherapy completion	
			Gender	Age		
				Pack years		
				N-stage		
DM	N-stage	N-stage	N-stage	N-stage		Chemotherapy regimen
			T-stage	T-stage		
			Hb baseline	Age		
			Pack-years	Pack years		

The highest performing model in external validation was a clinical model (Oropharynx-DM). With this clinical model a significant survival split was found both in training (Fig 2a) but not in validation (Fig 2b) based on the median prediction probabilities in training according to the Cox model.

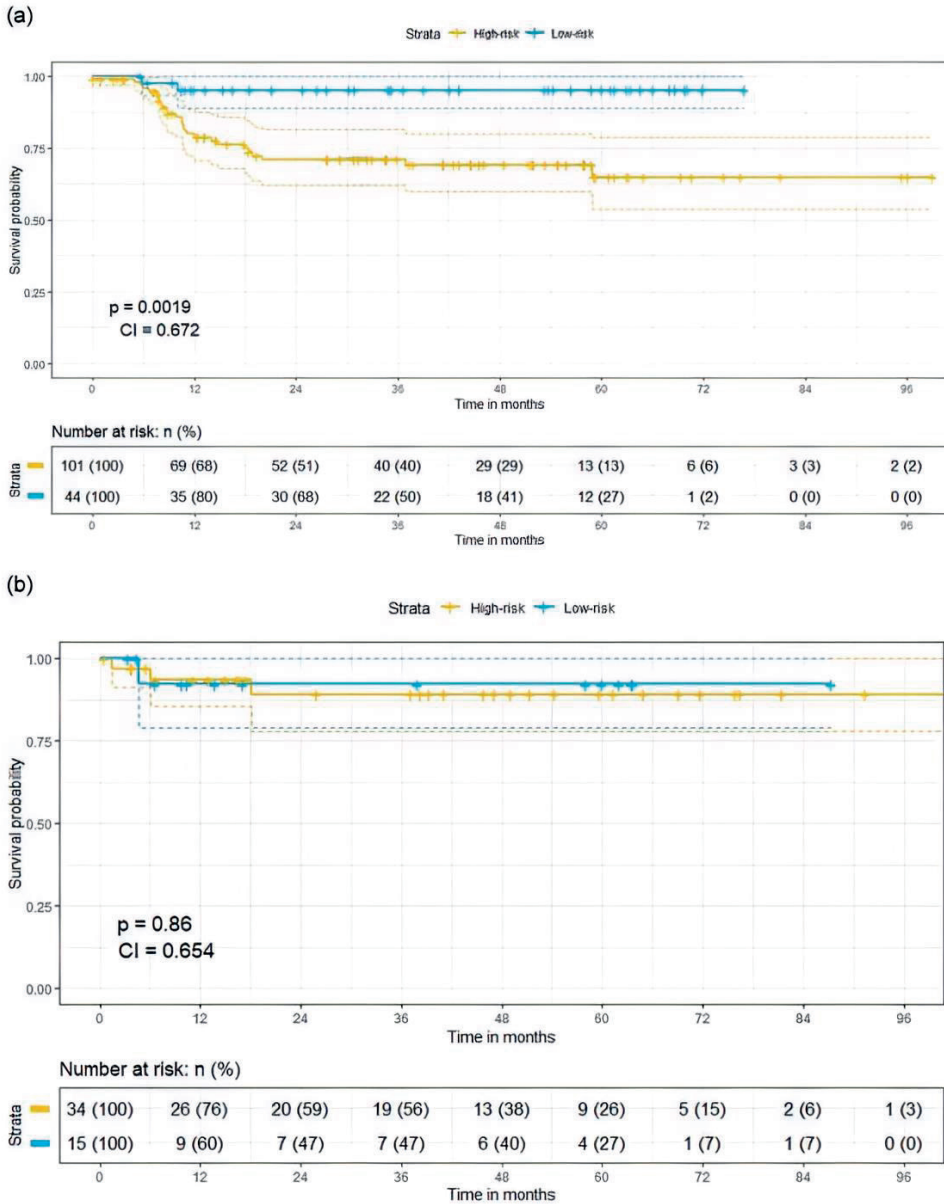


Figure 2. a. Training Kaplan-Meier (distant metastasis free) survival split for oropharyngeal patients (best performing clinical model in validation with Cox regression, oropharynx-DM) based on above (blue line) and below (yellow) median prediction probabilities. b. Validation Kaplan-Meier (distant metastasis free) survival split for oropharyngeal patients (best performing clinical model in validation with Cox regression, oropharynx-DM) based on above (blue line) and below (yellow) median prediction probabilities. Non-significant split in survival according to median in training, though in all of the above median cases the time to event is not observed (censoring).

4.4.2 Radiomics characteristics

A total of 1298 radiomic features were extracted from all contrast-enhanced CT-images. Results of training (DESIGN) and validation (BD2DECIDE) c-index metrics are provided in Tables 2 and 3. Both in oropharyngeal cases alone as well as in all tumor subsites combined peritumoral radiomics performed poorly in external validation, with C-index ranging from 0.32 to 0.61 with both feature selection and model generation methods. (Figs 3 and 4).

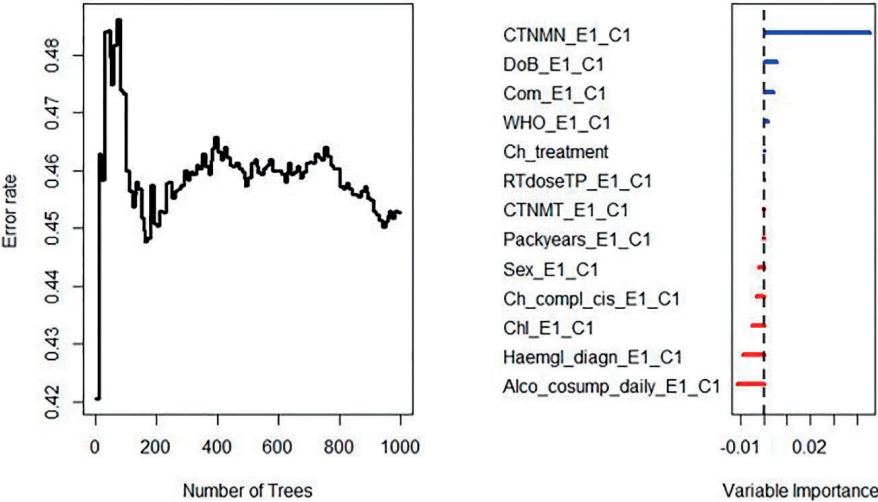


Figure 3. Error rate stabilizes with increasing number of trees. Features with an importance > 0 on an RFSRC model trained with all clinical variables in were eventually combined in the multivariable clinical (prognostic/ treatment-related) RFSRC model and externally validated on the BD2DECIDE dataset.

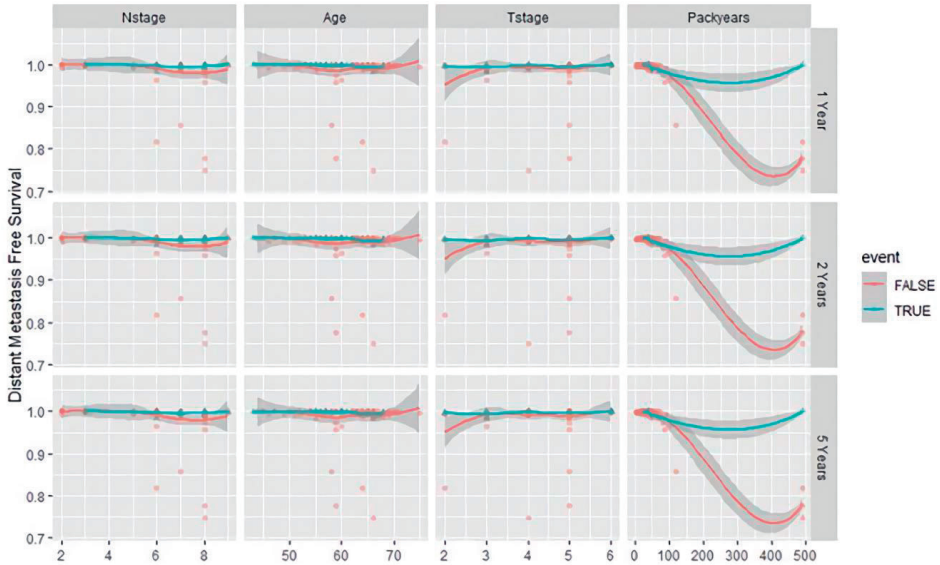


Figure 4. Variable dependence of predicted distant metastasis at 1, 2 and 5 years on the 4 clinical variables of interest (highest performing clinical model in validation, oropharyngeal-DM) according to Random Survival Forest. Individual cases are marked with blue triangles for censored cases and red circles for distant metastasis events. Loess smooth curve indicates the distant metastasis trend with increasing values of the individual clinical feature.

Volumetric information was calculated for GTV_{prim} and Spearman correlation coefficients between individual selected features and volume were calculated. With the Cox method these C-indexes were all <0.60 (all $P>0.05$ correlation with model features). With the RSF method these varied between 0.28–0.45 (all $P>0.05$ correlation with model features).

4.4.3 Radiomics quality assurance and TRIPOD statement

For quality assurance a radiomics quality score (RQS) was calculated [14] for this study. The RQS score for this specific study was 44% (most points allocated for external validation and use of feature reduction analysis). Scores were likewise calculated for the 22-item adherence data extraction checklist of the TRIPOD (Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis), which was in the range of 0.75–0.86 (See S3 Appendix).

4.5 Discussion

In this first peritumoral H&N radiomics study we found that the highest performing model in external validation was a clinical model which was able to predict distant metastasis in oropharyngeal cancer cases with an external validation c-index of 0.65 and 0.75 with the RSF and Cox models respectively. Both in oropharyngeal cases alone as well as in all tumor subsites peritumoral radiomics performed poorly in external validation, with C-index ranging from 0.32 to 0.61 with both feature selection and model generation methods.

The reasoning for choosing a 5mm tumor border is based on radiotherapy margins which are defined outside the visible/palpable or imaging-detectable (macroscopic) tumor GTV, the clinical target volume (CTV), whereby potential microscopic tumor spread is taken into account. Based on experience from pathological examination of surgical resections, the Danish Head and Neck Cancer (DAHANCA) group concluded that for primary tumors (GTV-T), the risk of subclinical microscopic spread was around 50% of which more than 99% was within 5 mm and 95% within 4 mm of the rim of GTV-T [18].

Previous studies on peritumoral radiomics in other tumor models have not been able to produce promising results in internal cross-validation either. We have not yet seen a peritumoral H&N radiomics study with an external validation dataset. Dou et al. [19] for instance found a testing C-index of 0.55 with a lung radiomic tumor border model in the prediction of distant metastasis, while Shan et al. [20] found that in predicting early recurrence in hepatocellular carcinoma (HCC), by comparing AUC values between training and validation cohorts, the prediction accuracy in the validation cohort was good for the peritumoral radiomics model (0.80 vs. 0.79, $P = 0.47$) but poor for the tumoral radiomics model (0.82 vs. 0.62, $P < 0.01$).

Despite the poor performance in external validation with both GTVprim, 3mm, and 5 mm tumor border radiomics, we have found a clinical model for the prediction of distant metastasis in oropharyngeal cancer patients performed the best in external validation.

We find an overlapping clinical parameter, namely node-stage, between these two clinical models. Indeed high node stage is hypothesized to be one of the major risk factors for the development of distant metastasis [21,22]. We also see some discrepancies between the two clinical models. For instance, T-stage, age, and packyears (the number of packs of cigarettes per day multiplied by the years spent smoking) are also selected as one of the predictors of distant metastasis in the RSF model.

Strengths of the current study include the use of an external validation dataset, the extensive clinical data and the rigorous feature selection methods that take into account time-to-event outcomes.

One of the limitations is the retrospective nature of the study, leading to several clinical variables (e.g. weight loss) to not be comparable between training and validation. Another limitation is the heterogeneity between the training and validation dataset, both in terms of WHO PS, N-stage, chemotherapy regimen (mostly platin alone regimens in DESIGN versus platin + other regimens in BD2DECIDE) as well as tumor site (DESIGN more oropharynx, less laryngeal cases compared to BD2DECIDE). We hypothesize that this has negatively impacted the model performance.

Another limitation is the omission of valuable semantic imaging features, qualitative imaging features that are defined by experienced radiologists (e.g. extracapsular growth, necrosis) as well as the omission of radiomics description of the GTV2 (positive lymph nodes).

Most radiomic features are designed to be extracted from a fully enclosed 3D volume, as is often the case with the primary tumor. In contrast, the peritumoral regions are rings with limited volume, especially the 3mm regions. Therefore, features such as those extracted from filtered images require a certain volume of the region of interest and therefore have limited application in small volumes or disjointed regions. These technical issues may have contributed to the relatively poor performance of peritumoral radiomics.

We believe that in the future, to improve clinical use of this kind of signatures, larger and more homogenous and prospectively collected data should be sought, taking into account imaging features derived from GTV2/ lymph node regions and gene expression profiles in order to construct more reliable prognostic biomarkers. An intrinsic problem might be that recurrences cannot be predicted well with bulk tumor characteristics. In a recent genetics study [23] it was shown that half of the local relapses of CRT treated HNSCCs, did not share genetic changes with the index tumors, suggesting that minor treatment resistant subclones determine outcome in many cases. Taking this into regard we believe that future radiomics studies should derive information not only from the planning CT's, but also during the multi-ple follow-up moments after treatment.

4.6 Conclusion

In this study, we have investigated whether clinical data as well as computer-extracted radio-mic features from peritumoral as well as inter-tumoral derived imaging features on CT can predict OS, LRR and DM. Our results show that radiomic features from the primary peritumoral regions, as well as from the primary inter-tumoral regions, do not predict OS, LRR and DM. More homogenous cohorts, both in patient and imaging characteristics, and the combination of clinical, radiomics, and genomics models may increase the generalizability and predictive power of prognostic models.

4.7 Acknowledgments

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015, n° 694812—Hypoximmuno), ERC-2018-PoC (n° 81320 –CL-IO). This research is also supported by the Dutch technology Foundation STW (grant n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. Authors also acknowledge financial support from SME Phase 2 (RAIL—n° 673780), EUROSTARS (DART, DECIDE), the European Program H2020-2015-17 (BD2Decide—PHC30-689715, PREDICT—ITN—n° 766276), TRANSCAN Joint Transnational Call 2016 (JTC2016 “CLEARLY”- n° UM 2017–8295), Interreg V-A Euregio Meuse-Rhine (“Eura-diagnostics”) and the Dutch Cancer Society.

Authors further acknowledge financial support by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018-2 and A6C 7072/2014-2.

4.8 References

- [1] Clayburgh Daniel R, Grandis Jennifer R et al. "Molecular Biology" Oral, Head and Neck Oncology and Reconstructive Surgery 2018: 79–89
- [2] Dutch Cancer Registry (NKR) from Integral Cancer Center Netherlands (IKNL): <https://www.iknl.nl/cijfers/de-nederlandse-kankerregistratie>.
- [3] Ang K. K., Harris J., Wheeler R., et al. Human papillomavirus and survival of patients with oropharyngeal cancer. *New England Journal of Medicine*. 2010; 363(1):24–35. <https://doi.org/10.1056/NEJMoa0912217> PMID: 20530316
- [4] Fakhry C., Westra W. H., Li S., et al. Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial. *JNCI Journal of the National Cancer Institute*. 2008; 100(4):261–269. <https://doi.org/10.1093/jnci/djn011> PMID: 18270337
- [5] Nahavandipour Arvin, Jakobsen Kathrine Kronberg, Gronhoj Christian et al. "Incidence and survival of laryngeal cancer in Denmark: a nation-wide study from 1980 to 2014." *Acta Oncologica* 2019; 58 (7)
- [6] Gregoire V., Lefebvre J.-L., Licitra L. et al. "Squamous cell carcinoma of the head and neck: EHS-ESMO-ESTRO Clinical Practice Guidelines for diagnosis, treatment and follow-up." *Ann Oncol*. 2010; 21 Suppl 5: v184–186.
- [7] Beaumont J., Acosta O., Devillers A. et al. "Voxel-based identification of local recurrence sub-regions from pre-treatment PET/CT for locally advanced head and neck cancers." *EJNMMI Res* 2019; 9: 90. <https://doi.org/10.1186/s13550-019-0556-z> PMID: 31535233
- [8] Trosman S. J., Koyfman S.A. Ward M.C. et al. "Effect of human papillomavirus on patterns of distant metastatic failure in oropharyngeal squamous cell carcinoma treated with chemoradiotherapy." *JAMA Otolaryngol Head Neck Surg*. 2015; 141(5): 457–462. <https://doi.org/10.1001/jamaoto.2015.136> PMID: 25742025
- [9] Leger S., Zwanenburg A., Pilz K. et al. "A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling." *Sci Rep*. 2017; 7(1): 13206. <https://doi.org/10.1038/s41598-017-13448-3> PMID: 29038455
- [10] Giraud P., Giraud P., Gasnier A. et al. "Radiomics and Machine Learning for Radiotherapy in Head and Neck Cancers." *Front Oncol* 2019; 9: 174. <https://doi.org/10.3389/fonc.2019.00174> PMID: 30972291
- [11] Alsahafi E., Begg K., Amelio I. et al. "Clinical update on head and neck cancer: molecular biology and ongoing challenges." *Cell Death Dis* 2019; 10(8): 540. <https://doi.org/10.1038/s41419-019-1769-9> PMID: 31308358
- [12] Peltanova B., Raudenska M., Masarik M. et al. "Effect of tumor microenvironment on pathogenesis of the head and neck squamous cell carcinoma: a systematic review." *Mol Cancer* 2019; 18(1): 63. <https://doi.org/10.1186/s12943-019-0983-5> PMID: 30927923
- [13] Gallagher M., Hogan J., Maire F. et al. "Intelligent Data Engineering and Automated Learning" *LNCS* 2005: 382–383

- [14] Lambin P, Leijenaar R.T.H., Deist T.M. et al. "Radiomics: the bridge between medical imaging and personalized medicine." *Nat Rev Clin Oncol* 2017; 14(12): 749–762. <https://doi.org/10.1038/nrclinonc.2017.141> PMID: 28975929
- [15] Johnson W. E., Li C., Rabinovic A. et al. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007; 8(1): 118–127. <https://doi.org/10.1093/biostatistics/kxj037> PMID: 16632515
- [16] Orlhac F, Frouin F, Nioche C. et al. Validation of a Method to Compensate Multicenter Effects Affecting CT Radiomics. *Radiology*. 2019; 291(1): 52–58.
- [17] Lucia F., Visvikis D., Vallieres M. et al. "External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy." *Eur J Nucl Med Mol Imaging*. 2019; 46(4): 864–877. <https://doi.org/10.1007/s00259-018-4231-9> PMID: 30535746
- [18] Campbell S., Poon I., Markel D. et al. "Evaluation of microscopic disease in oral tongue cancer using whole-mount histopathologic techniques: implications for the management of head-and-neck cancers." *Int J Radiat Oncol Biol Phys* 2012; 82(2): 574–581. <https://doi.org/10.1016/j.ijrobp.2010.09.038> PMID: 21300463
- [19] Dou T. H., Coroller T.P., van Griethuysen J.J.M. et al. "Peritumoral radiomics features predict distant metastasis in locally advanced NSCLC." *PLoS ONE* 2018; 13(11): e0206108. <https://doi.org/10.1371/journal.pone.0206108> PMID: 30388114
- [20] Shan Q. Y., Hu H., Feng S. et al. "CT-based peritumoral radiomics signatures to predict early recurrence in hepatocellular carcinoma after curative tumor resection or ablation." *Cancer Imaging* 2019; 19 (1): 11. <https://doi.org/10.1186/s40644-019-0197-5> PMID: 30813956
- [21] Garavello W., Ciardo A., Spreafico R. et al. "Risk factors for distant metastases in head and neck squamous cell carcinoma." *Arch Otolaryngol Head Neck Surg* 2006; 132(7): 762–766. <https://doi.org/10.1001/archotol.132.7.762> PMID: 16847186
- [22] Kim D. H., Kim W.T., Lee J.H. et al. "Analysis of the prognostic factors for distant metastasis after induction chemotherapy followed by concurrent chemoradiotherapy for head and neck cancer." *Cancer Res Treat* 2015; 47(1): 46–54. <https://doi.org/10.4143/crt.2013.212> PMID: 25327492
- [23] de Roest R. H., Mes S., Brink A. et al. "Molecular Characterization of Locally Relapsed Head and Neck Cancer after Concomitant Chemoradiotherapy." *Clin Cancer Res* 2019; OF1–OF9

4.9 Appendix A: Datasets, imaging parameters and missing data

Table 1. Datasets and imaging parameters in survival analysis

Dataset (center)	Nr. patients	Nr. With contrast enhanced CT (CECT)	Original slice thickness (mm, median)	Original pixel spacing (mm, median)
DESIGN (training)				
VUMC	88	88	2.5 mm	0.96x0.96
UMCU	81	81	2.0 mm	0.98x0.98
NKI	102	102	3.0 mm	0.98x0.98
MAASTRO	30	30	3.0 mm	0.98x0.98
BD2DECIDE (validation)				
VUMC	55	55	2.5 mm	0.96x0.96
UDUS	10	10	3.0 mm	4.60x4.60
INT	11	11	2.5 mm	4.60x4.60
MAASTRO	40	40	3.0 mm	0.98x0.98

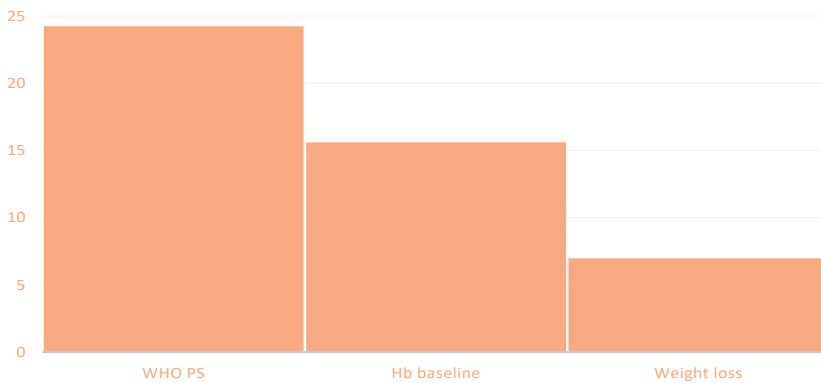


Table 2. Top 3 missing variables in training (DESIGN) cohort.



4.10 Appendix B: full names of ALL the ethics committees/institutional review boards that approved study

-Stichting VU-VUmc (VU/VUmc), NL6, 53815211 established in De Boelelaan 1105, 1081 HV, Amsterdam (The Netherlands), represented by Prof. Johannes Brug, Director and Dean

-Heinrich-Heine Universitaetet Duesseldorf (UDUS), CF10657730442, established in Universitaetsstrasse 1, 40225 Duesseldorf (Germany), DE811222416), represented by Dr. Martin Goch, Chancellor

-Fondazione IRCCS Istituto Nazionale dei tumori (INT), VIII/002398, established in Via Venezian 1, Milan 20133, Italy, IT0437635055, represented by mr. Enzo Lucchini, President

-Stichting Maastrto Radiation Oncology MAASTRO Clinic (MAASTRO) NL6, 41070330, established in Dr. Tanslaan 12, Maastricht 6229 ET, The Netherlands, represented by Mrs. Maria Jacobs, Administration Chief

-Stichting het Nederlands Kanker Instituut-Antoni van Leeuwenhoek ziekenhuis (Netherlands Cancer Institute/ Antoni van Leeuwenhoek Hospital), established at Plesmanlaan 121 1066 CX Amsterdam, in this matter duly represented by Prof. R Medema, P.h.d, in his capacity of Scientific Director and Chairman of the Board

-Universitair Medisch Centrum Utrecht, established at Heidelberglaan 100, 3584 CX Utrecht, in this matter duly represented by Prof. dr. ir. M.A. Viergever, Manager Research, and Mr. drs. H.K. Bouwer, Financila Manager

-Maastricht University, more specific its Faculty of Health, Medicine and Life Sciences, School of Oncology and Developmental Biology (GROW), having its principle office at Minderbroedersberg 4-6, 6211 LK Maastricht, The Netherlands, on behalf of the Executive Board represented by Prof Dr. Frans Ramaekers, Scientific Director GROW (third party, analysis of data).

4.11 Appendix C: TRIPOD adherence data extraction checklist



Prediction Model Development and Validation

B. TRIPOD ITEMS						
			[D] Develop ment	[V] External validatio n	[IV] Increme ntal value	[D+V] Developm ent and external validation (of same model)
Title and abstract						
<i>It is suggested to score items 1 and 2 (Title and Abstract) after scoring items 3 to 22, as only after reading the whole publication it can be judged whether the reporting in the title and abstract is complete.</i>						
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	Score 1 if all extraction items are scored as "Y"	Score 1 if all extraction items are scored as "Y"	Score 1 if all extraction items are scored as "Y"	Score 1 if all extraction items are scored as "Y"
	i	The words developing/development, validation/validating, incremental/added value (or synonyms) are reported in the title	0	0	0	0
	ii	The words prediction, risk prediction, prediction model, risk models, prognostic models, prognostic indices, risk scores (or synonyms) are reported in the title	1	1	0	1
	iii	The target population is reported in the title	1	1	0	0
	iv	The outcome to be predicted is reported in the title	1	1	0	0
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	Score 1 if all extraction items are scored as "Y" or "NA"	Score 1 if all extraction items are scored as "Y" or "NA"	Score 1 if all extraction items are scored as "Y" or "NA"	Score 1 if all extraction items are scored as "Y" or "NA"
	i	The objectives are reported in the abstract	1	1	0	0
	ii	Sources of data are reported in the abstract <i>E.g. Prospective cohort, registry data, RCT data.</i>	1	1	0	0
	iii	The setting is reported in the abstract <i>E.g. Primary care, secondary care, general population, adult care, or paediatric care. The setting should be reported for both the development and validation datasets, if applicable.</i>	0	0	0	0
	iv	A general definition of the study participants is reported in the abstract <i>E.g. patients with suspicion of certain disease, patients with a specific disease, or general eligibility criteria.</i>	1	1	0	1



	v	The overall sample size is reported in the abstract	1	1	0	1
	vi	The number of events (or % outcome together with overall sample size) is reported in the abstract <i>If a continuous outcome was studied, score Not applicable.</i>	0	0	0	0
	vii	Predictors included in the final model are reported in the abstract. For validation studies of well-known models, at least the name/acronym of the validated model is reported <i>Broad descriptions are sufficient, e.g. 'all information from patient history and physical examination'. Check in the main text whether all predictors of the final model are indeed reported in the abstract.</i>	1	1	1	1
	viii	The outcome is reported in the abstract	1	1	1	1
	ix	Statistical methods are described in the abstract <i>For model development, at least the type of statistical model should be reported. For validation studies a quote like "model's discrimination and calibration was assessed" is considered adequate. If done, methods of updating should be reported.</i>	1	1	1	1
	x	Results for model discrimination are reported in the abstract <i>This should be reported separately for development and validation if a study includes both development and validation.</i>	1	1	1	1
	xi	Results for model calibration are reported in the abstract <i>This should be reported separately for development and validation if a study includes both development and validation.</i>	0	0	0	0
	xii	Conclusions are reported in the abstract <i>In publications addressing both model development and validation, there is no need for separate conclusions for both; one conclusion is sufficient.</i>	1	1	1	1
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	if both extraction items are scored as "Y"	if both extraction items are scored as "Y"	if both extraction items are scored as "Y"	Score 1 if both extraction items are scored as "Y"
	i	The background and rationale are presented	1	1	1	1
	ii	Reference to existing models is included (or stated that there are no existing models)	1	1	1	1
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"
	i	It is stated whether the study describes development and/or validation and/or incremental (added) value	1	1	1	1
Methods						

		4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"
Source of data	i		The study design/source of data is described <i>E.g. Prospectively designed, existing cohort, existing RCT, registry/medical records, case control, case series.</i> <i>This needs to be explicitly reported; reference to this information in another article alone is insufficient.</i>	1	1	1	1
	4b		Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	Score 1 if all extraction items are scored as "Y", "NA", or "R"	Score 1 if all extraction items are scored as "Y", "NA", or "R"	Score 1 if all extraction items are scored as "Y", "NA", or "R"	Score 1 if all extraction items are scored as "Y", "NA", or "R"
	i		The starting date of accrual is reported	1	1	NA	1
	ii		The end date of accrual is reported	1	1	NA	1
	iii		The length of follow-up and prediction horizon/time frame are reported, if applicable <i>E.g. "Patients were followed from baseline for 10 years" and "10-year prediction of..."; notably for prognostic studies with long term follow-up. If this is not applicable for an article (i.e. diagnostic study or no follow-up), then score Not applicable.</i>	NA	NA	NA	NA
Participants	5a		Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	Score 1 if all extraction items are scored as "Y" or "R"	Score 1 if all extraction items are scored as "Y" or "R"	Score 1 if all extraction items are scored as "Y" or "R"	Score 1 if all extraction items are scored as "Y" or "R"
	i		The study setting is reported (e.g. primary care, secondary care, general population) <i>E.g.: 'surgery for endometrial cancer patients' is considered to be enough information about the study setting.</i>	1	1	1	1
	ii		The number of centres involved is reported <i>If the number is not reported explicitly, but can be concluded from the name of the centre/centres, or if clearly a single centre study, score Yes.</i>	1	1	1	1
	iii		The geographical location (at least country) of centres involved is reported <i>If no geographical location is specified, but the location can be concluded from the name of the centre(s), score Yes.</i>	1	1	1	1
	5b		Describe eligibility criteria for participants.	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"
	i		In-/exclusion criteria are stated <i>These should explicitly be stated. Reasons for exclusion only described in a patient flow is not sufficient.</i>	1	1	0	1



	5c	Give details of treatments received, if relevant.	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"
	i	Details of any treatments received are described <i>This item is notably for prognostic modelling studies and is about treatment at baseline or during follow-up. The 'if relevant' judgment of treatment requires clinical knowledge and interpretation. If you are certain that treatment was not relevant, e.g. in some diagnostic model studies, score Not applicable.</i>	1	1	1	1
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	Score 1 if all extraction items are scored as "Y" or "R"	Score 1 if all extraction items are scored as "Y" or "R"	Score 1 if all extraction items are scored as "Y" or "R"	Score 1 if all extraction items are scored as "Y" or "R"
	i	The outcome definition is clearly presented <i>This should be reported separately for development and validation if a publication includes both.</i>	1	1	1	1
	ii	It is described how outcome was assessed (including all elements of any composite, for example CVD [e.g. MI, HF, stroke]).	1	1	1	1
	iii	It is described when the outcome was assessed (time point(s) since T0)	1	1	1	1
	6b	Report any actions to blind assessment of the outcome to be predicted.	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"
i	Actions to blind assessment of outcome to be predicted are reported <i>If it is clearly a non-issue (e.g. all-cause mortality or an outcome not requiring interpretation), score Yes. In all other instances, an explicit mention is expected.</i>	1	1	1	1	
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	Score 1 if all extraction items are scored as "Y" or "R"	Score 1 if all extraction items are scored as "Y" or "R"	Score 1 if all extraction items are scored as "Y" or "R"	Score 1 if all extraction items are scored as "Y" or "R"
	i	All predictors are reported <i>For development, "all predictors" refers to all predictors that potentially could have been included in the 'final' model (including those considered in any univariable analyses). For validation, "all predictors" means the predictors in the model being evaluated.</i>	1	1	1	1
	ii	Predictor definitions are clearly presented	1	1	1	1

	iii	It is clearly described how the predictors were measured	1	1	1	1
	iv	It is clearly described when the predictors were measured	1	1	1	1
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	Score 1 if both extraction items are scored as "Y"	Score 1 if both extraction items are scored as "Y"	Score 1 if both extraction items are scored as "Y"	Score 1 if both extraction items are scored as "Y"
	i	It is clearly described whether predictor assessments were blinded for outcome <i>For predictors for which it is clearly a non-issue (e.g. automatic blood pressure measurement, age, sex) and for instances where the predictors were clearly assessed before outcome assessment, score Yes. For all other predictors an explicit mention is expected.</i>	0	0	0	0
	ii	It is clearly described whether predictor assessments were blinded for the other predictors	0	0	0	0
Sample size	8	Explain how the study size was arrived at.	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"
	i	It is explained how the study size was arrived at <i>Is there any mention of sample size, e.g. whether this was done on statistical grounds or practical/logistical grounds (e.g. an existing study cohort or data set of a RCT was used)?</i>	1	1	1	1
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	Score 1 if all extraction items are scored as "Y" or "NA"	Score 1 if all extraction items are scored as "Y" or "NA"	Score 1 if all extraction items are scored as "Y" or "NA"	Score 1 if all extraction items are scored as "Y" or "NA"
	i	The method for handling missing data (predictors and outcome) is mentioned <i>E.g. Complete case (explicit mention that individuals with missing values have been excluded), single imputation, multiple imputation, mean/median imputation. If there is no missing data, there should be an explicit mention that there is no missing data for all predictors and outcome. If so, score Yes. If it is unclear whether there is missing data (from e.g. the reported methods or results), score No. If it is clear there is missing data, but the method for handling missing data is unclear, score No.</i>	1	1	1	1
	ii	If missing data were imputed, details of the software used are given <i>When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.</i>	1	1	1	1
	iii	If missing data were imputed, a description of which variables were included in the imputation	1	1	1	1

		procedure is given. <i>When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.</i>				
	iv	If multiple imputation was used, the number of imputations is reported <i>When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.</i>	NA	NA	NA	NA
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.	Score 1 if all extraction items are scored as "Y" or "NA"	Not applicable	Score 1 if all extraction items are scored as "Y" or "NA"	Score 1 if all extraction items are scored as "Y" or "NA"
	i	For continuous predictors it is described whether they were modelled as linear, nonlinear (type of transformation specified) or categorized <i>A general statement is sufficient, no need to describe this for each predictor separately. If no continuous predictors were reported, score Not applicable.</i>	1	Not applicable	1	1
	ii	For categorical or categorized predictors, the cut-points were reported <i>If no categorical or categorized predictors were reported, score Not applicable.</i>	NA	Not applicable	NA	NA
	iii	For categorized predictors the method to choose the cut-points was clearly described <i>If no categorized predictors, score Not applicable.</i>	NA	Not applicable	NA	NA
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	Score 1 if all extraction items are scored as "Y" or "NA"	Not applicable	Score 1 if all extraction items are scored as "Y" or "NA"	Score 1 if all extraction items are scored as "Y" or "NA"
	i	The type of statistical model is reported <i>E.g. Logistic, Cox, other regression model (e.g. Weibull, ordinal), other statistical modelling (e.g. neural network)</i>	1	Not applicable	1	1
	ii	The approach used for predictor selection <u>before</u> modelling is described <i>'Before modelling' means before any univariable or multivariable analysis of predictor-outcome associations. If no predictor selection before modelling is done, score Not applicable. If it is unclear whether predictor selection before modelling is done, score No. If it is clear there was predictor selection before modelling but the method was not described, score No.</i>	1	Not applicable	1	1

	<p>iii The approach used for predictor selection <u>during</u> modelling is described <i>E.g. Univariable analysis, stepwise selection, bootstrap, Lasso.</i> <i>'During modelling' includes both univariable or multivariable analysis of predictor-outcome associations.</i> <i>If no predictor selection during modelling is done (so-called full model approach), score Not applicable.</i> <i>If it is unclear whether predictor selection during modelling is done, score No.</i> <i>If it is clear there was predictor selection during modelling but the method was not described, score No.</i></p>	1	Not applicable	1	1
	<p>iv Testing of interaction terms is described <i>If it is explicitly mentioned that interaction terms were not addressed in the prediction model, score Yes.</i> <i>If interaction terms were included in the prediction model, but the testing is not described, score No.</i></p>	0	Not applicable	0	0
	<p>v Testing of the proportionality of hazards in survival models is described <i>If no proportional hazard model is used, score Not applicable.</i></p>	1	Not applicable	1	1
	<p>vi Internal validation is reported <i>E.g. Bootstrapping, cross validation, split sample.</i> <i>If the use of internal validation is clearly a non-issue (e.g. in case of very large data sets), score Yes. For all other situations an explicit mention is expected.</i></p>	1	Not applicable	1	1
10c	For validation, describe how the predictions were calculated.	Not applicable	if extraction item is scored as "Y"	if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"
i.	<p>It is described how predictions for individuals (in the validation set) were obtained from the model being validated <i>E.g. Using the original reported model coefficients with or without the intercept, and/or using updated or refitted model coefficients, or using a nomogram, spreadsheet or web calculator.</i></p>	Not applicable	1	1	1
10d	<p>Specify all measures used to assess model performance and, if relevant, to compare multiple models.¹ <i>These should be described in the methods section of the paper (item 16 addresses the reporting of the results for model performance).</i></p>	Score 1 if extraction items 10di and 10dii are scored as "Y"	Score 1 if extraction items 10di and 10dii are scored as "Y"	Score 1 if all extraction items are scored as "Y"	Score 1 if extraction items 10di and 10dii are scored as "Y"
i	<p>Measures for model discrimination are described <i>E.g. C-index / area under the ROC curve.</i></p>	1	1	1	1
ii	<p>Measures for model calibration are described <i>E.g. calibration plot, calibration slope or intercept, calibration table, Hosmer Lemeshow test, O/E ratio.</i></p>	0	0	0	0

	iii	Other performance measures are described <i>E.g. R², Brier score, predictive values, sensitivity, specificity, AUC difference, decision curve analysis, net reclassification improvement, integrated discrimination improvement, AIC.</i>	1	1	1	1
	10e	Describe any model updating (e.g., recalibration) arising from the validation, if done.	Not applicable	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"
	i	A description of model-updating is given <i>E.g. Intercept recalibration, regression coefficient recalibration, refitting the whole model, adding a new predictor</i> <i>If updating was done, it should be clear which updating method was applied to score Yes.</i> <i>If it is not explicitly mentioned that updating was applied in the study, score this item as 'Not applicable'.</i>	Not applicable	NA	NA	NA
Risk groups	11	Provide details on how risk groups were created, if done.	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"
	i	If risk groups were created, risk group boundaries (risk thresholds) are specified <i>Score this item separately for development and validation if a study includes both development and validation.</i> <i>If risk groups were not created, score this item as not applicable.</i>	NA	NA	NA	NA
Development vs. validation	12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome and predictors.	Not applicable	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y" or "NA"	Score 1 if extraction item is scored as "Y"
	i	Differences or similarities in <u>definitions</u> with the development study are described <i>Mentioning of any differences in all four (setting, eligibility criteria, predictors and outcome) is required to score Yes.</i> <i>If it is explicitly mentioned that there were no differences in setting, eligibility criteria, predictors and outcomes, score Yes.</i> <i>For incremental value reports, in case additional predictors are not added to a previously developed prediction model but rather added to</i>	Not applicable	0	0	0

		<i>conventional predictors in a newly fitted model, score Not applicable.</i>				
Results						
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	Score 1 if all extraction items are scored as "Y" or "NA"	Score 1 if the extraction items are scored as "Y" or "NA"	Score 1 if all extraction items are scored as "Y" or "NA"	Score 1 if all extraction items are scored as "Y" or "NA"
	i	The flow of participants is reported	1	1	1	1
	ii	The number of participants with and without the outcome are reported <i>If outcomes are continuous, score Not applicable.</i>	1	1	1	1
	iii	A summary of follow-up time is presented <i>This notably applies to prognosis studies and diagnostic studies with follow-up as diagnostic outcome. If this is not applicable for an article (i.e. diagnostic study or no follow-up), then score Not applicable.</i>	1	1	1	1
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	Score 1 if all extraction items are scored as "Y"	Score 1 if all extraction items are scored as "Y"	Score 1 if all extraction items are scored as "Y"	Score 1 if all extraction items are scored as "Y"
	i	Basic demographics are reported	0	0	0	0
	ii	Summary information is provided for all predictors included in the final developed/validated model	1	1	1	1
	iii	The number of participants with missing data for predictors is reported	1	1	1	1
	iv	The number of participants with missing data for the outcome is reported	1	1	1	1
	13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	Not applicable	Score 1 if all extraction items are scored as "Y"	Score 1 if all extraction items are scored as "Y" or "NA"	Score 1 if all extraction items are scored as "Y"
	i	Demographic characteristics (at least age and gender) of the validation study participants are reported along with those of the original development study <i>For incremental value reports, in case additional predictors are not added to a previously developed prediction model but rather added to conventional predictors in a newly fitted model, score Not applicable.</i>	Not applicable	1	1	1
	ii	Distributions of predictors in the model of the validation study participants are reported along with those of the original development study <i>For incremental value reports, in case additional</i>	Not applicable	0	0	0



		<i>predictors are not added to a previously developed prediction model but rather added to conventional predictors in a newly fitted model, score Not applicable.</i>				
	iii	Outcomes of the validation study participants are reported along with those of the original development study <i>For incremental value reports, in case additional predictors are not added to a previously developed prediction model but rather added to conventional predictors in a newly fitted model, score Not applicable.</i>	Not applicable	1	1	1
Model development	14a	Specify the number of participants and outcome events in each analysis.	if both extraction items are scored as "Y" or "NA"	Not applicable	if both extraction items are scored as "Y" or "NA"	if both extraction items are scored as "Y" or "NA"
	i	The number of participants in each analysis (e.g. in the analysis of each model if more than one model is developed) is specified	1	Not applicable	1	1
	ii	The number of outcome events in each analysis is specified (e.g. in the analysis of each model if more than one model is developed) <i>If outcomes are continuous, score Not applicable.</i>	1	Not applicable	1	1
	14b	If done, report the unadjusted association between each candidate predictor and outcome.	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"	Not applicable	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"	Score 1 if extraction item is scored as "Y"; score <i>Not applicable</i> if extraction item is scored as "NA"
	i	The unadjusted associations between each predictor and outcome are reported <i>If any univariable analysis is mentioned in the methods but not in the results, score No. If nothing on univariable analysis (in methods or results) is reported, score this item as Not applicable.</i>	NA	Not applicable	NA	NA
Model specification	15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	Score 1 if both extraction items are scored as "Y"	Not applicable	Score 1 if both extraction items are scored as "Y"	Score 1 if both extraction items are scored as "Y"
	i	The regression coefficient (or a derivative such as hazard ratio, odds ratio, risk ratio) for each predictor in the model is reported	0	Not applicable	0	0
	ii	The intercept or the cumulative baseline hazard (or baseline survival) for at least one time point is reported	1	Not applicable	1	1
	15b	Explain how to use the prediction model.	Score 1 if extraction item is scored as	Not applicable	Score 1 if extraction item is scored as	Score 1 if extraction item is scored as



			"Y"		"Y"	"Y"
	i	An explanation (e.g. a simplified scoring rule, chart, nomogram of the model, reference to online calculator, or worked example) is provided to explain how to use the model for individualised predictions.	1	Not applicable	1	1
Model performance	16	Report performance measures (with confidence intervals) for the prediction model.² <i>These should be described in results section of the paper (item 10 addresses the reporting of the methods for model performance).</i>	Score 1 if extraction items are scored as "Y"	Score 1 if extraction items are scored as "Y"	Score 1 if all extraction items are scored as "Y"	Score 1 if extraction items are scored as "Y"
	i	A discrimination measure is presented <i>E.g. C-index / area under the ROC curve.</i>	1	1	1	1
	ii	The confidence interval (or standard error) of the discrimination measure is presented	1	1	1	1
	iii	Measures for model calibration are described <i>E.g. calibration plot, calibration slope or intercept, calibration table, Hosmer Lemeshow test, O/E ratio.</i>	1	1	1	1
	iv	Other model performance measures are presented <i>E.g. R², Brier score, predictive values, sensitivity, specificity, AUC difference, decision curve analysis, net reclassification improvement, integrated discrimination improvement, AIC.</i>	1	1	1	1
Model updating	17	If done, report the results from any model updating (i.e., model specification, model performance, recalibration). <i>If updating was not done, score this TRIPOD item as 'Not applicable'.</i>	Not applicable	Score 1 if all extraction items are scored as "Y"	Not applicable	Score 1 if all extraction items are scored as "Y"
	i	The updated regression coefficients for each predictor in the model are reported <i>If model updating was described as 'not needed', score Yes.</i>	Not applicable	NA	Not applicable	NA
	ii	The updated intercept or cumulative baseline hazard or baseline survival (for at least one time point) is reported <i>If model updating was described as 'not needed', score Yes.</i>	Not applicable	NA	Not applicable	NA
	iii	The discrimination of the updated model is reported	Not applicable	NA	Not applicable	NA
	iv	The confidence interval (or standard error) of the discrimination measure of the updated model is reported	Not applicable	NA	Not applicable	NA
	v	The calibration of the updated model is reported	Not applicable	NA	Not applicable	NA
Discussion						
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"
	i	Limitations of the study are discussed <i>Stating any limitation is sufficient.</i>	1	1	1	1

Interpretation	19a	For validation, discuss the results with reference to performance in the development data, and any other validation data.	Not applicable	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"
	i	Comparison of results to reported performance in development studies and/or other validation studies is given	Not applicable	1	1	1
	19b	Give an overall interpretation of the results considering objectives, limitations, results from similar studies and other relevant evidence.	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"	Score 1 if extraction item is scored as "Y"
	i	An overall interpretation of the results is given	1	1	1	1
Implications	20	Discuss the potential clinical use of the model and implications for future research.	Score 1 if both extraction items are scored as "Y"	Score 1 if both extraction items are scored as "Y"	Score 1 if both extraction items are scored as "Y"	Score 1 if both extraction items are scored as "Y"
	i	The potential clinical use is discussed <i>E.g. an explicit description of the context in which the prediction model is to be used (e.g. to identify high risk groups to help direct treatment, or to triage patients for referral to subsequent care).</i>	1	1	1	1
	ii	Implications for future research are discussed <i>E.g. a description of what the next stage of investigation of the prediction model should be, such as "We suggest further external validation".</i>	1	1	1	1
Other information						
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, web calculator, and data sets.	Not included in overall scoring	Not included in overall scoring	Not included in overall scoring	Not included in overall scoring
	i	Information about supplementary resources is provided	1	1	1	1
Funding	22	Give the source of funding and the role of the funders for the present study.	Score 1 if both extraction items are scored as "Y"	Score 1 if both extraction items are scored as "Y"	Score 1 if both extraction items are scored as "Y"	Score 1 if both extraction items are scored as "Y"
	i	The source of funding is reported or there is explicit mention that there was no external funding involved	1	1	1	1
	ii	The role of funders is reported or there is explicit mention that there was no external funding	1	1	1	1
Total Adherence	23	Calculates the total Adherence to the TRIPOD statement	0.86	0.86	0.75	0.80

Source: <https://www.tripod-statement.org/>

CHAPTER 5

5

Investigation of the added value of CT-based radiomics in predicting the development of brain metastases in patients with radically treated stage III NSCLC

Simon A. Keek*¹, Esmā Kayan*¹, Avishek Chatterjee¹, José S.A. Belderbos², Gerben Bootsma³, Ben van den Borne⁴, Anne-Marie C. Dingemans⁵, Hester A. Gietema⁶, Harry J.M. Groen⁷, Judith Herder⁸, Cordula Pitz⁹, John Praag¹⁰, Dirk De Ruyscher¹¹, Janna Schoenmaekers¹², Hans J.M. Smit¹³, Jos Stigt¹⁴, Marcel Westenend¹⁵, Haiyan Zeng¹¹, Henry C. Woodruff^{1,6}, Philippe Lambin^{1,6}, Lizza Hendriks¹²

*Contributed equally to the article

1 The D-Lab, Department of Precision Medicine, GROW-School for Oncology and Reproduction, Maastricht University, Universiteitssingel 40, 6229 ER Maastricht, The Netherlands

2 Department of Radiation Oncology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

3 Department of Pulmonary Diseases, Zuyderland Hospital, Henri Dunantstraat 5, 6419 PC Heerlen, The Netherlands

4 Department of Pulmonary Diseases, Catharina Hospital, Michelangelolaan 2 5623 EJ Eindhoven, The Netherlands

5 Department of Pulmonary Diseases, Erasmus MC, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands

6 Department of Radiology and Nuclear Medicine, GROW-School for Oncology, Maastricht University Medical Centre+, P.O. Box 5800, 6202 AZ Maastricht, The Netherlands

7 Department of Pulmonary Diseases, University Medical Center Groningen, University of Groningen, Hanzeplein 1 9713 GZ Groningen, The Netherlands

8 Department of Pulmonary Diseases, Meander Medical Center, Maatweg 3, 3813 TZ Amersfoort, The Netherlands

9 Department of Pulmonary Diseases, Laurentius Hospital, Mgr. Driessenstraat 6, 6043 CV Roermond, The Netherlands

10 Department of Radiotherapy, Erasmus MC, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands

11 Department of Radiation Oncology (Maastr), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands

12 Department of Pulmonary Diseases, GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, P.O. Box 5800, 6202 AZ Maastricht, The Netherlands

13 Department of Pulmonary Diseases, Rijnstate, Wagnerlaan 55, 6815 AD Arnhem, The Netherlands

14 Department of Pulmonary Diseases, Isala Hospital, Dokter van Heesweg 2, 8025 AB Zwolle, The Netherlands

15 Department of Pulmonary Diseases, VieCuri, Tegelseweg 210, 5912 BL Venlo, The Netherlands

5.1 Abstract

5.1.1 Introduction

Despite radical intent therapy for patients with stage III non-small cell lung cancer (NSCLC), cumulative incidence of brain metastases (BM) reaches 30%. Current risk stratification methods fail to accurately identify these patients. As radiomics features have been shown to have predictive value, this study aims to develop a model combining clinical risk factors with radiomics features for BM development in patients with radically treated stage III NSCLC.

5.1.2 Methods

Retrospective analysis of two prospective multicenter studies. Inclusion criteria: adequately staged (18-FDG-PET-CT, contrast-enhanced chest CT, contrast-enhanced brain MRI/CT) and radically treated stage III NSCLC, exclusion criteria: second primary within 2 years of NSCLC diagnosis, prior prophylactic cranial irradiation. Primary endpoint was BM development any time during follow-up (FU). CT-based radiomics features (N=530) were extracted from the primary lung tumor on 18-FDG-PET-CT images, and a list of clinical features (N=8) was collected. Univariate feature selection based on the area under the curve (AUC) of the receiver operating characteristic was performed to identify relevant features. Generalized linear models were trained using the selected features, and multivariate predictive performance was assessed through the AUC.

5.1.3 Results

In total 219 patients were eligible for analysis. Median FU was 59.4 months for the training cohort and 67.3 months for the validation cohort; 21 (15%) and 17 (22%) patients developed BM in the training and validation cohort, respectively. Two relevant clinical features (age and adenocarcinoma histology), and four relevant radiomics features were identified as predictive. The clinical model yielded the highest AUC value of 0.71 (CI 95% 0.58-0.84), better than radiomics or a combination of clinical parameters and radiomics (both an AUC of 0.62, 95% CIs of 0.47-0.76 and 0.48-0.76, respectively).

5.1.4 Conclusion

CT-based radiomics features of primary NSCLC in the current setup could not improve on a model based on clinical predictors (age and adenocarcinoma histology) of BM development in radically treated stage III NSCLC patients.

5.2 Introduction

The brain is a frequent site of disease relapse in patients with non-small cell lung cancer (NSCLC). Risk factors for brain metastases (BM) are advanced stage, adenocarcinoma histology, and younger age¹⁻³. For radically treated patients, locally advanced (stage III) NSCLC has the highest risk for BM, with a cumulative incidence of BM of approximately 30%⁴. The majority of BM present within two years of diagnosis, despite brain imaging without BM during initial staging for NSCLC⁴. Brain magnetic resonance imaging (MRI) is recommended in clinical guidelines (and if not possible, contrast enhanced computed tomography (CECT))⁵⁻⁸. The type of chemotherapy administered during chemoradiation therapy does not influence the incidence of BM². Curative treatment of (symptomatic) BM is seldom possible and for the overwhelming majority of patients overall survival (OS) is limited⁹. Moreover, BM are associated with a devastating impact on Quality of Life (QoL)^{10, 11}. Therefore, strategies to prevent BM and to predict who is at risk for their development are necessary, especially taking into consideration that treatments that reduce the incidence of BM are possible.

Prophylactic cranial irradiation (PCI) has been shown to reduce the incidence of BM in patients with NSCLC with a relative risk of 0.33⁴. PCI prolongs progression free survival (PFS) in stage III NSCLC, but not OS⁴. Furthermore, PCI leads to neuro-cognitive impairment (mostly grade 1-2) in about 25-27% of patients^{12, 13}. Ideally, only those patients with an a priori high risk of BM should undergo PCI and those with a low risk could avoid the risk of neurocognitive decline. An alternative approach to preventive treatment would be to closely monitor patients at high risk for BM through MRI surveillance, although there is no evidence that this improves outcome¹⁴. Hence, identifying predictive biomarkers, and thereby stratifying patients at high vs low risk for BM development, is key to personalize follow-up and treatment.

Although clinical risk factors are identified as described above, it remains challenging to discriminate between patients at high and low risk of BM^{15, 16}. Won et al. (2015) developed a prediction model using clinical and pathological risk factors, such as histology, pathological T- and N-stages, and smoking status to predict the probability of BM development after curative surgery in a large group of patients with NSCLC¹⁷. This study used dedicated brain imaging (majority brain MRI, subset brain CECT) at baseline to verify that no BM were present. However, the model only had a moderate discriminative power in predicting BM development at 2 and 5 years (Harrell's C-index (CI) of 0.670 and 0.674, respectively), and was verified only through internal validation, showing a clear need for more studies investigating BM prediction models.

Metastases develop through a “wiring” of the primary tumour to spread to certain organs (“seed and soil” hypothesis) ¹⁸⁻²⁰. Therefore, analysis of the primary tumour could provide valuable feedback in identifying those patients at risk of developing BM. Indeed, molecular biomarkers, such as microRNAs (miRNA) expression patterns were previously associated with BM development in patients with NSCLC^{21, 22}. However, these markers were not investigated in a prospective predictive study. Furthermore, they require invasive biopsies, and small tumour biopsies disregard the heterogeneous nature of tumours²³. Therefore, an approach that takes the entirety of the tumour into account (i.e. the whole primary tumour and not only a small biopsy) is preferred.

Radiomics refers to the extraction of quantitative data from medical images using mathematical algorithms and finding correlations with biological or clinical outcomes via machine learning techniques²⁴⁻²⁶. When radiomics is applied to oncology, radiological images (e.g., CT, MRI, or positron emission tomography [PET]) performed during routine clinical workflow can be used to non-invasively extract imaging features describing the tumour and patient phenotypes²⁷. These features can have significant diagnostic, prognostic, and predictive value, and hold the potential to assist clinical decision-making²⁸.

Coroller et al. (2015) found that a model based on the primary tumour in locally advanced adenocarcinomas of the lung was predictive of distant metastases²⁹. However, this study tried to predict distant metastases in general, not BM specifically. Three other studies showed that CT -based radiomics models on primary lung tumours might have positive value to predict BM in patients with NSCLC³⁰⁻³². Models of clinical features and radiomics features were compared and combined, and in all three studies complementary value for the radiomics models were found. However, sample sizes were small (N = 85-124), no external validation was performed, not all patients were adequately staged according to guidelines⁵⁻⁸, and patient groups included were heterogeneous (e.g. different disease stages), which may affect the reliability of the created models.

Therefore, the aim of the current study is to develop a prediction model for BM development (low vs high risk) in patients with adequately staged, radically treated stage III NSCLC, based on clinical patient characteristics only, and combined with CT-based radiomics analysis of the primary lung tumour. We hypothesize that a model based on CT-radiomics and clinical variables can assist medical professionals in the decision-making process, and facilitate precision medicine for the treatment of NSCLC.

5.3 Materials and methods

5.3.1 Study population

This was a post hoc analysis of two prospective, multicentre studies (NVALT-11, NCT01282437) [inclusion 2009-2015] and NL3335 [inclusion 2012-2017]) enrolling patients with stage III NSCLC (IASLC 7th edition). NCT01282437 (N = 175) was a multicentre randomized phase III study evaluating PCI vs no PCI in patients with radically treated stage III NSCLC. Primary endpoint was the development of symptomatic BM 24 months after randomization. Approximately half of these patients had baseline brain CECT, the remaining brain MRI. Only patients without baseline BM were eligible³³. NL3335 was a prospective multicentre observational study, evaluating whether performing a brain MRI after a negative dedicated CECT had additive value in the diagnosis of asymptomatic BM³⁴. One of the secondary endpoints was development of BM after radical treatment for stage III NSCLC. For NL3335, patients with stage III NSCLC and an available 18F-FDG-PET-CT were screened, and only those with a dedicated brain CT (with contrast, arms at thorax level, correct field of view, and delayed imaging³⁵) performed before or together with the 18F-FDG-PET-CT available, and followed by a brain MRI, were deemed eligible. For the current study, all patients who were staged with 18F-FDG-PET-CT and dedicated brain imaging (MRI and/or CECT), and treated with radical intent therapy (i.e. sequential or concurrent chemoradiation with/without surgery, or radical radiotherapy), were eligible. The collection of the imaging data for the current study was approved by the Medical Ethics Review Committee of Maastricht UMC+ (2017-0317), and, if applicable, by institutional review boards of the other participating centres. The ethics committee approved the waiver of informed consent. For both studies, additional eligibility criteria consisted of availability of baseline chest CECT (i.e. at diagnosis of stage III NSCLC), and a distinct primary tumour (primary tumour not detectable [Tx] or primary tumour not definable due to surrounding atelectasis were excluded). Furthermore, all patients that received PCI or had a second primary within 2 years of NSCLC diagnosis were excluded.

The dataset was split into a training and a validation dataset. The patient data obtained from the NL3335 study from the hospitals in Heerlen (Zuyderland MC) and Maastricht (Maastricht UMC+) were assigned to the training dataset. This dataset was used to select relevant features and to train the model. To test the performance on data not yet seen by the model, a validation dataset was also defined comprising data from one of the centres participating in the NL3335 study (VieCuri Medisch Centrum) and from the NVALT-11 study.

5.3.2 Patient characteristics

Baseline characteristics recorded in the two prospective studies and extracted for this analysis included age, gender, World Health Organization Performance Status (WHO

PS), smoking status, pack years, TNM-stage (IASLC 7th edition, IIIA vs. IIIB), histology, and follow-up data regarding BM development. The primary endpoint of this study was the development of BM (binary: yes/no), which was defined as disease progression to the brain assessed by MRI or CECT anytime during follow-up.

Image acquisition

Pre-treatment diagnostic chest CT-images were acquired with a Philips Gemini TF64 (Philips Medical Systems, Best, Netherlands), Siemens Somatom Force scanner (Siemens Healthineers, Erlangen, Germany), GE Discovery STE (GE Medical systems, Chicago, United States), and Toshiba Aquilion (Toshiba, Tokyo, Japan). The scanning parameters were 80-140 kVp tube voltage, 37-462 mAs tube current, and 512×512 matrix. An overview of the imaging characteristics can be found in supplementary figure S1. CT-images were obtained through the picture archiving and communication system in the Digital Imaging and Communications in Medicine format. For each patient an ¹⁸F-FDG-PET-CT with a non-diagnostic low-dose CT for attenuation correction and diagnostic CECT were available. Generally, the injection of contrast induces noise in the images and hence in some radiomics features due to differences between patients in diffusion of the contrast agent. However, the CECT scan was finally chosen for the analysis, as several tumours were difficult to contour on the low-dose CT due to mediastinal invasion and undefined tumour borders. Furthermore, the lower spatial resolution of low-dose CT could lead to the loss of important radiomics information. The CECT scans were obtained with different imaging parameters (e.g. spatial resolution, slice thickness, reconstruction kernel) due to variation in acquisition protocols of hospitals and different scanners available. Therefore, imaging parameters that were the most common throughout all images were set as the standard imaging parameters, e.g. 3mm slice thickness, soft reconstruction kernel, which were used to select the appropriate CECT scan for each patient accordingly.

5.3.4 Tumour segmentation

The region of interest (ROI), i.e. the primary lung tumour, was manually delineated on the CT-images using MIM Software Inc. (Version 6.9.4, Cleveland, Ohio, USA). ¹⁸F-FDG-PET-CT imaging was used alongside the CT-image to locate the tumour, and to identify tumour borders adjacent to atelectasis or tumours invading extra-pulmonary structures. The lung window was used to identify tumour-lung borders, while tumour regions adjacent to extra-pulmonary tissues were contoured in the mediastinal window. In cases of tumours completely (or for a greater part) surrounded by atelectasis (i.e. reliable contouring not possible) the CT-scan was excluded from radiomics analysis. All tumour segmentations were performed and checked for accurate delineation by an experienced pulmonary oncologist or thoracic radiologist.

5.3.5 Pre-processing and feature extraction

In order to homogenize the datasets prior to feature extraction all images were resampled to the mode of the unprocessed scans (1x1 mm² pixel size and 3mm slice thickness). Furthermore, to reduce noise and computational burden, the intensity values inside the ROI were discretized with a fixed bin width of 25 Hounsfield units (HU) which has been reported to yield the most reproducible radiomics features for CT images³⁶.

Feature extraction for every 3-D ROI on each baseline CECT was performed using PyRadiomics version 2.2.0 on both the original images as well as filtered images. Laplacian of Gaussian (LoG) convolution filtering was applied to the original image in order to highlight regions of intensity change within an image. The LoG was applied with five different Gaussian standard deviation values ranging from 1mm to 5mm resulting in five different LoG images. The radiomics features extracted from the images can be divided into 3 main groups: first-order intensity and histogram statistics features, shape and size features, and texture features. First-order intensity and histogram statistics features describe the voxel intensity distribution within the ROI. Shape- and size features describe the spatial characteristics of the ROI itself, such as volume and sphericity, and are thus independent of the image contents. Texture features describe the spatial relationships of voxel intensities and are derived from six different matrices that are defined over the ROIs: grey-level co-occurrence (GLCM), grey-level run length (GLRLM), grey-level size-zone (GLSZM), grey-level distance-zone (GLDZM), neighbourhood grey-level dependence (NGLDM), and neighbourhood grey-tone difference matrix (NGTDM).

The total number of features that can be extracted with the PyRadiomics package, without using highly correlating/depreciated features and without any further manipulation of the image is 107. However, the application of image filters, either Wavelet-based or Log-based with different kernel sizes can multiply this number to thousands of features. The wavelet-based features were omitted from this analysis, as with a relatively low number of patients adding more features would increase the risk of overfitting and finding spurious correlations, and because wavelet-based features have shown to have low reproducibility compared to Log-filtered images³⁷.

5.3.6 Feature selection and predictive modeling

The radiomics features were first normalized on the training dataset through z-score normalization: the mean and standard deviation (SD) of each feature were determined over the entire training population and used to perform normalization on the training dataset, as well as on the validation dataset. For the clinical features, a list of known clinical predictors for BM defined by Won et al (2015) were used¹⁷. These included histology (adenocarcinoma vs. others), age, stage (IIIA vs. IIIB), WHO PS (0 vs. 1 or higher, 0-1 vs. 2 or higher, and 0-2 vs. 3), smoking status (ever vs. never, and current vs. former or ever), packyears, and treatment received (concurrent chemoradiation vs. other). As the

volume of the tumour is also a radiomics feature, it was not included as a clinical variable. Dimensionality reduction through feature selection was performed on both the radiomics and clinical variables.

Feature selection and modelling were performed using R software (Version 3.3.2, R Core Team, Vienna, Austria) on the training dataset³⁸. Supervised univariate feature selection was performed on all clinical and radiomics features, using the occurrence of BM as the binary outcome. For each feature, the area under the curve (AUC) of the receiving operating characteristic (ROC) was calculated. The ROC-curve shows the sensitivity and specificity of the model at different classification thresholds on the feature score. The AUC of this curve was a metric of the predictive performance of the feature, ranging from 0.5 to 1, where 1 indicates a perfect prediction and 0.5 a prediction equal to chance. As an AUC > 0.6 indicates a feature has some predictive power, this cut-off was chosen to select features. Features that are highly correlated (Spearman's correlation > 0.8) were determined, and the feature with the highest average correlation with all other features remaining in the set was excluded. To verify that radiomics features are not simply surrogates for tumour volume, the correlation with volume was also determined. Three separate models were created: using the selected radiomics features, using the selected clinical features, and using a combination of selected radiomics and clinical features.

Using the selected features, a generalized linear model (GLM) was trained on the training dataset using BM status as outcome calculated. Without changing its parameters, the model was then validated on the validation dataset, and the prediction score created as output. This prediction score is the probability a patient will develop a BM, and ranges from 0 to 1. By selecting a threshold on this prediction score, the binary classification of the validation patients was performed.

5.3.7 Patient inclusion

A total of 467 patients with stage III NSCLC were reviewed for selection, and 248 patients were excluded for several reasons: not fully staged (N = 15, no adequate brain imaging, i.e. no brain MRI or dedicated brain CT as defined in the methods section); no radical therapy performed (N = 69); history of previous cancer (N = 10); no CECT of the chest available (N = 90); atelectasis surrounding primary tumour (N = 17); no detectable primary tumour (N = 8). Lastly, from the NVALT-11 study, all patients with available imaging who underwent PCI were excluded (N = 39). As a result, 219 patients with stage III NSCLC with segmented CECT images were included for radiomics analysis. The CONSORT diagram depicting the selection process is depicted in Figure 1.

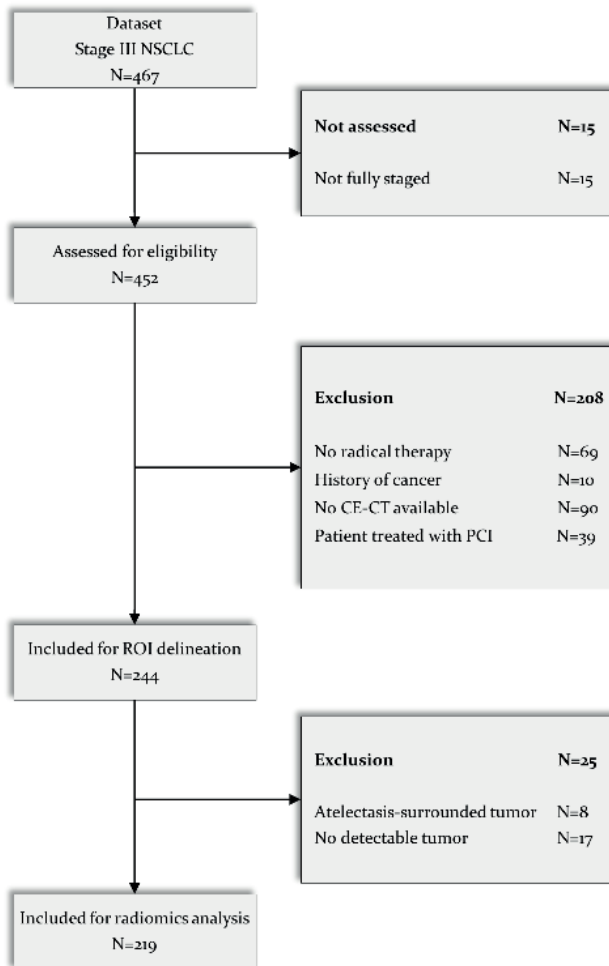


Figure 1. CONSORT diagram for patient selection. NSCLC, non-small cell lung cancer; CE, contrast-enhanced; CT, computed tomography; MRI, magnetic resonance imaging; ROI, region of interest; PCI, Prophylactic Cranial Irradiation.

5.3.8 Statistical analysis

Baseline patient characteristics were analysed using standard descriptive statistics. Statistical analysis of continuous variables was performed with the independent two-sample *t*-test, whereas differences in categorical variables were analysed using a χ^2 -test. The reported statistical significance levels were all two-sided set at a < 0.05 .

The predictive performance of the model was quantified through the AUC of the ROC. Calibration of the model on the external dataset was tested using the calibration curve, and a χ^2 -test to see if the slope and intercept are significantly different from 0 and 1, respectively. If this test is significant, it indicates the model does not fit on the external dataset. The ROC-curve

was plotted, and its confidence interval of 95% was calculated on 2000 stratified bootstrap replicates. Additionally, the binary classification was used to create a confusion matrix, which visualizes the performance of the model by comparing the predicted BM status to the true BM status. The binary classification was performed by determining an optimal threshold on the prediction score, calculated on 2000 stratified bootstrap replicates. The metric calculated to determine the optimal cut-off was the F1-score, which takes both precision and recall into account. From this binary prediction the sensitivity, specificity, precision, negative predictive value, accuracy, balanced accuracy, and F1-score were determined. Lastly, a two-proportion z-test was performed to determine if there was a significant difference between the true proportions of cases in the two predicted risk-groups.

The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines were adhered to³⁹. To test this adherence the adherence form was filled in, and the TRIPOD score is reported (supplementary materials table S1). This score is a grade from 0 to 100% that gives an indication of the compliance to the TRIPOD guidelines.

5.4 Results

5.3.1 Patient characteristics

Of the resulting 219 patients, 142 were assigned as the training dataset and 77 as the validation set. These datasets are completely independent. An overview of baseline patient characteristics is listed in Table 1. In the training set, 21 patients developed BM (incidence of 15%); in the validation dataset, 17 patients had BM development (22%). In the training dataset, 100% of the patients received a brain MRI at staging. For the validation dataset, 85.7% of the patients received an MRI, while the remaining 14.3% (11 patients) only received a CECT scan of the brain. Additionally, the median follow-up time in the training dataset was 59.4 months (interquartile range (IQR) 40.4-71.2), and in the validation dataset 67.3 months (IQR 42.0-83.3) ($p = 0.05$). In the entire population, patients were mostly male (61%) and mean age was 67 years at the time of NSCLC diagnosis, with 75% of patients > 60 years. The majority of patients (~88%) had a WHO performance score of 0 or 1. Most patients were either current (45%) or former smokers (50%), while 3% had never smoked (2% unknown smoking status). Patients were evenly distributed in the stages IIIA and IIIB (51% and 49%, respectively), and 38% had adenocarcinoma histology. No significant differences were found in patient characteristics between the training and validation sets, except for age, where the mean age was significantly higher ($p < 0.001$) and the proportion of patients over 60 years old was significantly larger (p of 0.005) in the training dataset. In addition, the validation dataset received a significantly lower proportion of brain MRI ($p < 0.001$).

Table 1. Baseline characteristics of patients assigned to training and validation sets.

Characteristic	Training set N = 142 (%)	Validation set N = 77 (%)	Total N = 219 (%)	p
Gender				0.939
Male	87 (61.3)	46 (59.7)	133 (60.7)	
Female	55 (38.7)	31 (40.3)	86 (39.3)	
Age (years)				
Mean ± SD	68.6 ± 8.3	63.6 ± 8.2	66.8 ± 8.6	< 0.001
Range	47.5-88.6	47.2-85.0	47.2-88.6	
< 60 years	26 (18.3)	28 (36.4)	54 (24.7)	0.005
> 60 years	116 (81.7)	49 (63.6)	165 (75.3)	
WHO PS				0.293
0	53 (37.3)	26 (33.8)	79 (36.1)	
1	68 (47.9)	45 (58.4)	113 (51.6)	
2	16 (11.3)	3 (3.9)	19 (8.7)	
3	2 (1.4)	2 (2.6)	4 (1.8)	
Unknown	3 (2.1)	1 (1.3)	4 (1.8)	
Smoking status				0.163
Never	5 (3.5)	2 (2.6)	7 (3.2)	
Former	64 (45.1)	45 (58.4)	109 (49.8)	
Current	69 (48.6)	30 (39.0)	99 (45.2)	
Unknown	4 (2.8)	0 (0)	4 (1.8)	
TNM-stage				0.415
IIIA	76 (53.5)	36 (46.8)	112 (51.1)	
IIIB	66 (46.5)	41 (53.2)	107 (48.9)	
Histology				0.382
Adenocarcinoma	55 (38.7)	28 (36.4)	83 (37.9)	
Squamous cell carcinoma	62 (43.7)	30 (39.0)	92 (42.0)	
Large cell carcinoma	5 (3.5)	7 (9.1)	12 (5.5)	
Sarcomatoid	1 (0.7)	0 (0)	1 (0.5)	
LCNEC	2 (1.4)	0 (0)	2 (0.9)	
NOS	17 (12.0)	12 (15.6)	29 (13.2)	
Brain metastasis diagnosed				0.241
Yes	21 (14.8)	17 (22.1)	38 (17.4)	
No	121 (85.2)	60 (77.9)	181 (82.6)	
Baseline brain MRI or brain CECT				< 0.001
MRI	142 (100)	66 (85.7)	208 (95)	
Only CECT	0 (0)	11 (14.3)	11 (5)	
Treatment received				0.233
CCRT +/- Surgery	100 (70.4)	61 (79.2)	161 (73.5)	
SCRT +/- Surgery	35 (24.6)	15 (19.5)	50 (22.8)	
Radical RT	7 (4.9)	1 (1.3)	8 (3.7)	

SD, standard deviation; TNM, tumour, node, metastasis; NOS, not otherwise specified; WHO PS, World Health Organization Performance Status: 0-1: Good, 2-3: Poor; LCNEC, large cell neuroendocrine carcinoma; NOS, not otherwise specified; CCRT, concurrent chemo radiotherapy; SCRT, sequential chemo radiotherapy; RT, radiotherapy.

5.3.2 Feature selection

In total, 530 radiomics features were extracted from each CT-image, and 8 clinical features were collected for each patient. After testing for univariate predictive performance and selecting features with AUC > 0.6, and excluding features with high correlation (Spearman correlation > 0.8), four relevant radiomics features (see supplementary materials section 1), and two relevant clinical features (adenocarcinoma vs other tumour types, and age as a continuous variable) were identified. None of the radiomics features showed high correlation (Spearman's correlation > 0.8) with tumour volume. Table 2 shows an overview of the selected features with their respective univariate AUC, and Spearman's correlation values with the volume.

Table 2. Selected clinical and radiomics features with corresponding univariate AUC, and Spearman's correlation with volume. LoG = Laplacian of Gaussian; GLSZM = grey-level size-zone matrix; GLCM = grey-level correlation matrix

Feature names		AUC	Correlation with volume
Clinical features	Adenocarcinoma vs. other tumour type	0.66	-
	Age (continuous)	0.73	-
Radiomics features	1mm LoG GLSZM normalized size-zone non-uniformity	0.60	-0.24
	2mm LoG GLCM correlation	0.62	0.52
	2mm LoG GLCM informational measure of correlation 1	0.61	-0.55
	2mm LoG GLCM informational measure of correlation 2	0.62	0.30

5.3.3 Clinical model

The performance of the predictive model built on the clinical features was evaluated in the validation set with an ROC curve, yielding an AUC of 0.71 (95% CI 0.58-0.84), as presented in Figure 2A. The calibration test yielded a p of 0.76, indicating the model fits on the external validation data. The calibration slope can be found in supplementary materials, figure S3. The binary prediction determined through bootstrapping gave a sensitivity and specificity of 0.82 and 0.57, respectively, which are shown in the figure represented by the dashed lines. The F1-score, the metric used to determine this cut-off, was 0.49.

The confusion matrix, shown in figure 2B, shows the number of correct and incorrect predictions. Of the control cases, 34 were predicted correctly; of the event cases, 14 were predicted correctly. The precision was 0.35, and the negative predictive value was 0.92. The accuracy and balanced accuracy were 0.62 and 0.70, respectively. Finally, the proportion of cases between predicted risk-groups were significantly different (p = 0.01).

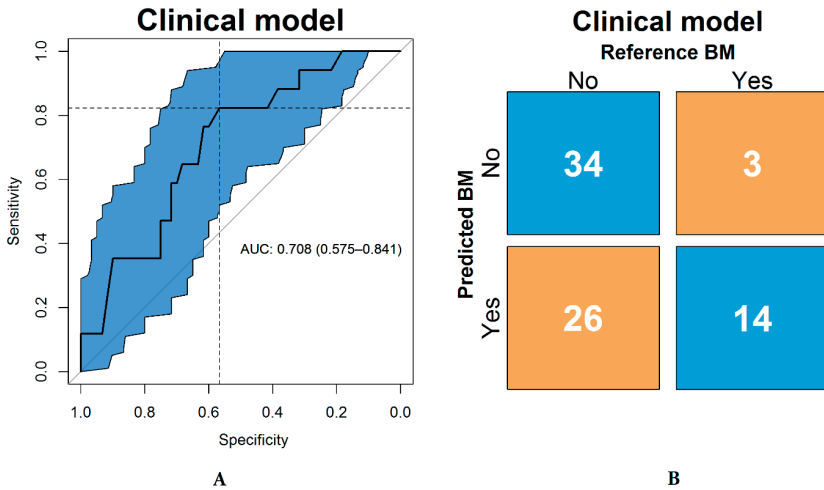


Figure 2. (A) Receiver operator characteristic curve and the corresponding confidence interval of 95% in blue of the clinical model, with area under the curve (AUC) and 95% confidence interval shown. On the y-axis is the sensitivity and on the x-axis the specificity of the model at different classification thresholds. The dashed lines show the sensitivity and specificity for the threshold that was used to make the binary prediction. **(B)** Confusion matrix with proportions of correct and wrong predictions made by the clinical model (y-axis) relative to the true labels (x-axis).

5.3.4 Radiomics model

The performance of the predictive model was evaluated in the validation set with an ROC curve, yielding an AUC of 0.62 (95% CI 0.47-0.76), as presented in Figure 3A. The calibration test yielded a $p < 0.001$, indicating the model does not fit on the external validation data. The calibration slope can be found in supplementary materials, figure S4. The binary prediction determined through bootstrapping gives a sensitivity and specificity of 0.65 and 0.6, respectively, which are shown in the figure represented by the dashed lines. The F1-score, the metric used to determine this cut-off, was 0.42.

The confusion matrix, shown in figure 3B, shows the number of correct and incorrect predictions. Of the control cases, 36 were predicted correctly; of the event cases, 11 were predicted correctly. The precision was 0.31, and the negative predictive value was 0.86. The accuracy and balanced accuracy were 0.61 and 0.62, respectively. Finally, the proportion of cases between predicted risk-groups were not significantly different ($p = 0.13$).

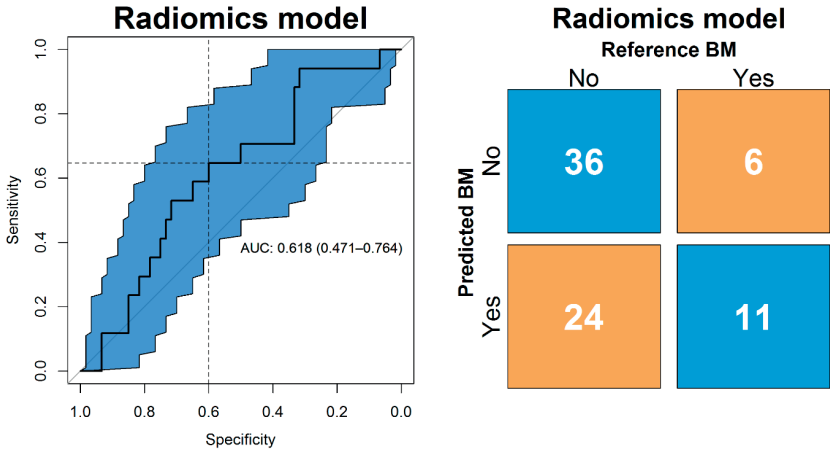


Figure 3. (A) Receiver operator characteristic curve and the corresponding confidence interval of 95% in blue of the radiomics model, with area under the curve (AUC) and 95% confidence interval shown. On the y-axis is the sensitivity and on the x-axis the specificity of the model at different classification thresholds. The dashed lines show the sensitivity and specificity for the threshold that was used to make the binary prediction. **(B)** Confusion matrix with proportions of correct and wrong predictions made by the radiomics model (y-axis) relative to the true labels (x-axis).

5.3.5 Radiomics & Clinical model

The performance of the predictive model was evaluated in the validation set with an ROC curve, yielding an AUC of 0.62 (95% CI 0.48-0.76), as presented in Figure 4A. The calibration test yielded a p of 0.03, indicating the model does not fit on the external validation data. The calibration slope can be found in supplementary materials, figure S5. The binary prediction determined through bootstrapping gives a sensitivity and specificity of 0.82 and 0.52, respectively, which are shown in the figure represented by the dashed lines. The F1-score, the metric used to determine this cut-off, was 0.47.

The confusion matrix, shown in figure 4B, shows the number of correct and incorrect predictions. Of the control cases, 31 were predicted correctly; of the event cases, 14 were predicted correctly. The precision was 0.33, and the negative predictive value was 0.91. The accuracy and balanced accuracy were 0.58 and 0.67, respectively. Finally, the proportion of cases between predicted risk-groups were significantly different (p = 0.03).

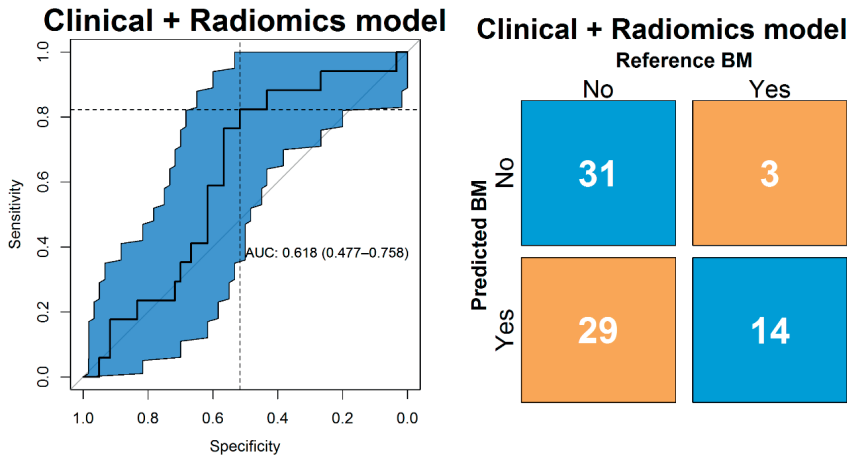


Figure 4. (A) Receiver operator characteristic curve and the corresponding confidence interval of 95% in blue of the clinical & radiomics model, with area under the curve (AUC) and 95% confidence interval shown. On the y-axis is the sensitivity and on the x-axis the specificity of the model at different classification thresholds. The dashed lines show the sensitivity and specificity for the threshold that was used to make the binary prediction. (B) Confusion matrix with proportions of correct and wrong predictions made by the clinical & radiomics model (y-axis) relative to the true labels (x-axis).

5.3.6 TRIPOD statement

The TRIPOD adherence for 22 guidelines was determined, and the adherence score was calculated to be 93%. The adherence form for this study can be found in supplementary materials table s1.

5.5 Discussion

The prediction and prevention of BM development in patients with radically treated stage III NSCLC is a major issue, as BM has a detrimental effect on survival and QoL^{10,11}. Preventive strategies such as PCI exist, but come at a cost of neurocognitive decline, and PCI has been shown to not be associated with an OS benefit in patients with stage III NSCLC not selected for BM risk⁴. Therefore, future studies evaluating new preventive treatments or the effects of regular screening should focus on those at high risk of BM. Patients with a low risk of BM could be spared PCI or intense imaging follow-up. This strategy requires a model that accurately separates high risk from low risk stage III NSCLC patients.

In this multicentre study, we developed a radiomics model based on four radiomics features extracted from the primary lung tumour on CECT-imaging and combined this with existing clinical predictors of BM. The first feature is based on a GLSZM matrix, which

quantifies the number and size of homogenous intensity patches found within the ROI. The normalized size-zone non-uniformity (NSZNU) feature based on this matrix measures variability of these size zones, with a higher score meaning less homogenous areas with the same intensity present in the ROI, i.e., more heterogeneity. The remaining three features are based on a GLCM matrix, which measures the frequency in which certain combinations of pixel intensity values are found. The features correlation, Informational Measure of Correlation 1 (IMC1), and Informational Measure of Correlation 2 (IMC2) based on this matrix all measure whether correlations between certain intensity values can be found within the ROI. A higher value would mean that more homogenous areas exist within the ROI, while a lower value means the intensity values are more randomly spread throughout the ROI, which is again a measure of heterogeneity.

We found that in a patient population of 219 (training N=142, validation N=77) the addition of radiomics was not able to improve the predictive performance of a model based solely on clinical factors. This result may indicate that, for the aforementioned population size, factors other than phenotypical characteristics of the tumour are more important in the incidence of BM, such as histology and age, as shown in the features selected for the clinical model.

To our knowledge, few studies have been undertaken on the topic of BM prediction using a combination of clinical and radiomics features. We found three radiomics studies with a comparable study design, shown contrasted to our study in Table 3²⁹⁻³¹. While one of the radiomics models has significantly higher performance (AUC of 0.85 vs. 0.62), these studies shared a low number of patients as well as BM events, a lack of external validation, and a lack of full staging compared to the current study, resulting in low reliability of the results.

Data quality should be a priority when selecting the study population⁴⁰. Especially, the large disease heterogeneity in stage III NSCLC emphasizes the importance of correct staging with the appropriate imaging modalities, as disease stage directly influences treatment options and prognosis⁵. For the previously reported studies either 18F-FDG-PET-CT or dedicated brain imaging (brain MRI or dedicated brain CT) were not mandatory, while in the present study only adequately staged patients were included for analysis. Therefore, in the previously reported studies, patients with occult BM could have been enrolled. For example, 15-21% of patients with stage III NSCLC have asymptomatic BM and without dedicated imaging, these will be missed^{41,42}. Asymptomatic BM are diagnosed on MRI in approximately 5% of patients that underwent a dedicated brain CT (with contrast and the correct field of view), and in 16% of patients that underwent an 18F-FDG-PET-CT with a low dose CT of the brain^{34,42}. All patients in our study received dedicated brain imaging, with 95% MRI and 5% CECT. Therefore, risk of bias due to undetected baseline BM is low in our study.

Table 3. Study parameters of radiomics studies on BM or DM prediction in NSCLC.

Study name	Coroller <i>et al.</i> (2015)	Chen <i>et al.</i> (2019)	Xu <i>et al.</i> (2019)	Present study (2021)
Study population	Stage II-III / adenocarcinoma	T1-stage / adenocarcinoma	Stage III-IV / ALK-positive	Stage IIIA/B
Sample size	N = 182	N = 89	N = 105	N = 219
Primary outcome	DM	BM	BM	BM
Number of events in FU	69 (37.9%)	35 (39.3%)	27 (25.7%)	38 (17.4%)
Staging	?	T1/N-stage based on non-CECT	'by medical images'	Full imaging
18F-FDG-PET-CT	-	-	?	+
Brain MRI / CECT (% MRI received)	- (N/A)	+ (not reported)	+ (not reported)	+ (95)
Chest CECT	-	-	+	+
Pathological analysis	Pathologically-confirmed lung adenocarcinoma	'Pathologically confirmed disease'	Pathologically confirmed ALK	-
Imaging modality	Planning CT + GTV (patients excluded if CTx/surgery was before RTx scheduled date)	Pretreatment non-CECT	Pretreatment CECT + RTstruct	Pretreatment CECT + RTstruct
Predictive performance (95% CI)	CI > 0.6 (-)	AUC 0.85 (0.767-0.933)	AUC 0.64 (0.501-0.783)	AUC 0.62 (0.47-0.76)
Strengths	(+) Pathologically confirmed (+) Pretreatment CT	(+) Pathologically confirmed (+) BM exclusion at baseline (+) Pretreatment CT	(+) Pathologically confirmed (+) BM exclusion at baseline (+) Diagnostic chest CECT / Pretreatment	(+) Pathologically confirmed (+) BM exclusion at baseline (+) Diagnostic chest CECT / Pretreatment (+) External validation
Limitations	(-) Unclear staging (-) Small sample size (-) GTV not specified (LN included?) (-) DM locations not specified (-) Planning CT	(-) Unclear staging; T1/N-stage determined with non-CECT (-) Small sample size	(-) Unclear staging; PET-CT not reported (-) Small sample size (-) GTV not specified (LN included?) (-) relatively low number of BM	(-) relatively low number of BM

NSCLC, non-small cell lung cancer, DM, distant metastasis; BM, brain metastasis; ¹⁸F-FDG-PET-CT, ¹⁸F-Fluorodeoxyglucose positron emission tomography-computed tomography; MRI, magnetic resonance imaging; (CE)-CT, contrast-enhanced computed tomography; CTx, chemotherapy; RTx, radiotherapy; T1, tumour stage 1; N, lymph node stage; LN, lymph node; ALK, anaplastic lymphoma kinase.

A further point of strength of this study is the use of ^{18}F -FDG-PET-CT alongside CECT images during contouring. In the field of radiation therapy, the differentiation of lung tumour from post obstructive atelectasis is a well-recognized problem, which even contrast enhancement cannot always resolve. As ^{18}F -FDG-PET-CT has proven utility during tumour delineation for radiation planning purposes, this may have significantly increased the delineation accuracy of the CECT-images in our study⁴³.

There may be a number of different reasons why the radiomics model failed to accurately predict patients at risk for BM. This study primarily focused on the selection of CECT images in consideration of delineation accuracy, as CECT is more specific in differentiating different tissue types, especially in case of mediastinal invasion, which often occurs in stage III NSCLC⁴⁴. However, this may have diminished the discriminatory performance of the model, since recent studies have found differences between CECT and non-CECT radiomics features^{45, 46}. In addition, CECT was associated with variability of radiomics features due to differences in contrast uptake; a concept which is strongly influenced by patient variables which impact contrast distribution, e.g. age and weight⁴⁷. Given that patient-related factors are a permanent source of variability (with any imaging modality), efforts should be directed at homogenizing datasets in terms of contrast-enhancement and investigating CECT robust features. Furthermore, despite the strict selection of CECT with the same reconstruction protocol and slice spacing, there were still differences in imaging parameters and the images were not fully standardized. The collected images were not standardized to one acquisition and reconstruction protocol before or during the studies. Furthermore, due to the retrospective nature of the study, we were not able to perform phantom scans on the different scanners. Performing phantom studies or applying a different harmonization method is likely needed to harmonize images and make reproducible models. This should be standard practice in a radiomics protocol⁴⁸⁻⁵⁰.

This study was performed on a homogenous patient group regarding stage, only including stage IIIA and IIIB tumours. However, stage III NSCLC is known for its heterogeneity regarding varying tumour sizes and the pattern of lymph node metastasis (e.g. a T1N3 vs a T4N0 tumour)⁵¹. This could further explain the inability of the model to predict BM, and while it was not in the scope of the current study due to a lack of data in the NCT01282437 study, investigating further clinical features that describe the risk of high T-status vs high N-status, or total tumour volume could be investigated, as Won et al (2015) have shown these features have predictive power¹⁷. The clinical features selected, age and histology, are not directly affected by this shortcoming. Although selection based on stage may increase homogeneity, it could also overlook the complexity of BM risk. For instance, primary tumour size alone is inadequate in predicting disseminating tumour behaviour, i.e. small tumours with extensive N-status have previously been described to metastasize early, whereas large tumours with limited N-stage may not at all⁵². Therefore, a critical

evaluation of the target population and the associated clinical implications is necessary in conducting relevant research.

Compared to previous studies that report a BM incidence of approximately 30%, the incidences of BM in the training and validation set were significantly lower at 15% and 22%, respectively⁴. Both NVALT11 and NL3335 had a median follow-up time largely exceeding two years, while most BM occur within two years of the initial staging of NSCLC³³. Therefore, inadequate follow-up time is not an explanation. For NVALT11 (control arm 28% BM in follow-up), not all scans could be retrieved, and indeed more scans were retrieved from patients without BM. In addition, almost all patients included had a baseline brain MRI and not only a CECT. It is known that MRI is slightly superior (in 5% of patients additional BM detected after negative CECT) in detecting asymptomatic BM in stage III NSCLC and this also could have resulted in a lower BM incidence in the follow-up³⁴.

The small sample size, even though larger datasets were used compared to previous studies, and different imaging parameters are both well-known sources of variability in radiomics that limit reproducibility⁵³. Furthermore, manual tumour delineations are prone to inter-observer variability, which affect the stability of radiomics features⁵⁴. Taken together, these aspects may explain the limited performance of the radiomics model and require further attention. Therefore, our future work will address these limitations by optimizing the radiomics model through expanding the sample size and reducing data heterogeneity, by using imaging phantoms and standardization methods in the radiomics pipeline, and through image and feature harmonization. While clinical factors seem to outperform radiomics features, with the current sample size the results are inconclusive with regard to the complementary predictive role of CT-based radiomics.

Future radiomics studies could also focus on utilizing the additional imaging performed during the standard diagnostic workup of patients with stage III NSCLC. These imaging modalities, e.g., dedicated brain MRI or CECT together with¹⁸F-FDG-PET-CT, may have additional value in BM prediction. For instance, brain MRI features might reveal micro metastases indiscernible to the human eye, and may aid in the early detection, whereas tumour heterogeneity captured by ¹⁸F-FDG-PET-CT uptake pattern may further characterize tumour aggressiveness⁵⁵. Accordingly, imaging modality-specific features could be integrated to form a robust radiomics signature.

-Finally, other AI approaches, such as deep learning models, have shown to be able to perform risk prediction on clinical images⁵⁶. While these methods usually require larger datasets to achieve significant results, they should be investigated in future studies for their complementary value in predicting the risk of BM. Other machine learning methods such as recursive feature elimination (RFE) or least absolute shrinkage and selection operator

(LASSO) to select features exist, which have shown to be able to improve performance of predictive models. However, with the current study setup and study population size, the feature selection through univariate predictive performance was found to achieve the highest performance.

5.6 Conclusion

A model based on known clinical predictors of BM development (age and tumour histology) is able to predict BM development in patients with radically treated stage III NSCLC with moderate precision, with an AUC of 0.71 (model available on www.ai4cancer.ai). This model did not improve with the addition of CT-based radiomics features. Future work will focus on optimizing the radiomics model by expanding the dataset, investigating more clinical features, other imaging modalities, data harmonization, and reducing data heterogeneity.

5.7 Funding

This study was funded by a Lung Foundation grant, n° 11.1.18.250. Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno), ERC-2020-PoC: 957565-AUTO.DISTINCT. Authors also acknowledge financial support from SME Phase 2 (RAIL n°673780), the European Union's Horizon 2020 research and innovation programme under grant agreement: ImmunoSABR n° 733008, MSCA-ITN-PREDICT n° 766276, CHAIMELEON n° 952172, EuCanImage n° 952103, Scholarship of China Scholarship Council (Grant No. : CSC 201909370087).

5.8 Conflicts of interest

P.L. reports, within and outside the submitted work, grants/sponsored research agreements from Radiomics SA, ptTheragnostic/DNAmito, Health Innovation Ventures. He received an advisor/presenter fee and/or reimbursement of travel costs/consultancy fee and/or in kind manpower contribution from Radiomics SA, BHV, Merck, Varian, Elekta, ptTheragnostic, BMS and Convert pharmaceuticals. Dr Lambin has minority shares in the company Radiomics SA, Convert pharmaceuticals, Comunicare Solutions and LivingMed Biotech, he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Radiomics SA and one issued patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, one non issued patent on LSRT (PCT/ P126537PC00) licensed to Varian Medical, three non-patented

invention (softwares) licensed to ptTheragnostic/DNAmito, Radiomics SA and Health Innovation Ventures and three non-issues, non licensed patents on Deep & handcrafted Radiomics (US P125078US00, PCT/NL/2020/050794, n° N2028271). He confirms that none of the above entities or funding was involved in the preparation of this paper.

LH: none related to current manuscript, outside of current manuscript: research funding Roche Genentech, Boehringer Ingelheim, AstraZeneca (all institution, furthermore Takeda and Beigene in negotiation [institution]); advisory board: BMS, Eli Lilly, Roche Genentech, Pfizer, Takeda, MSD, Boehringer Ingelheim, Amgen, Janssen (all institution, Roche one time self); speaker: MSD (institution); travel/conference reimbursement: Roche Genentech (self); mentorship program with key opinion leaders: funded by AstraZeneca; fees for educational webinars: Benecke, Medtalks, VJOnco (self), high5oncology (institution); interview sessions funded by Roche Genentech, Bayer (institution); local PI of clinical trials: AstraZeneca, Novartis, BMS, MSD /Merck, GSK, Takeda, Blueprint Medicines, Roche Genentech, Janssen Pharmaceuticals, Mirati.

DdR: none related to current manuscript, outside of current manuscript: grants from BMS, AstraZeneca, Seattle Genetics, Philips, Olink, BeiGene. Advisory board (no personal fees: AstraZeneca, Philips.

HW: has minority shares in the company Radiomics SA.

H.G: Advisory board (no personal fees): Roche, Astra-Zeneca, Boehringer-Ingelheim, Lilly, Novartis.

5.9 References

1. Siegel RL, Miller KD and Jemal A. Cancer Statistics, 2017. *CA Cancer J Clin* 2017; 67: 7-30. 2017/01/06. DOI: 10.3322/caac.21387.
2. Hendriks LE, Brouns AJ, Amini M, et al. Development of symptomatic brain metastases after chemoradiotherapy for stage III non-small cell lung cancer: Does the type of chemotherapy regimen matter? *Lung cancer* 2016; 101: 68-75. DOI: 10.1016/j.lungcan.2016.09.008.
3. Govindan R, Bogart J and Vokes EE. Locally advanced non-small cell lung cancer: the past, present, and future. *J Thorac Oncol* 2008; 3: 917-928. 2008/08/02. DOI: 10.1097/JTO.0b013e318180270b.
4. Witlox WJA, Ramaekers BLT, Zindler JD, et al. The Prevention of Brain Metastases in Non-Small Cell Lung Cancer by Prophylactic Cranial Irradiation. *Front Oncol* 2018; 8: 241. DOI: 10.3389/fonc.2018.00241.
5. Eberhardt WE, De Ruyscher D, Weder W, et al. 2nd ESMO Consensus Conference in Lung Cancer: locally advanced stage III non-small-cell lung cancer. *Ann Oncol* 2015; 26: 1573-1588. 2015/04/22. DOI: 10.1093/annonc/mdv187.
6. Remon J, Soria JC, Peters S, et al. Early and locally advanced non-small-cell lung cancer: an update of the ESMO Clinical Practice Guidelines focusing on diagnosis, staging, systemic and local therapy. *Ann Oncol* 2021; 32: 1637-1642. 2021/09/05. DOI: 10.1016/j.annonc.2021.08.1994.
7. Specialisten FM. Niet kleincellig longcarcinoom, https://richtlijndatabase.nl/richtlijn/niet_kleincellig_longcarcinoom/startpagina_-_niet-kleincellig_longcarcinoom.html (2020).
8. Network NCC. NCCN Guidelines - Non-Small Cell Lung Cancer, <https://www.nccn.org/guidelines/guidelines-detail?category=1&id=1450> (2022).
9. Sperduto PW, Yang TJ, Beal K, et al. Estimating Survival in Patients With Lung Cancer and Brain Metastases: An Update of the Graded Prognostic Assessment for Lung Cancer Using Molecular Markers (Lung-molGPA). *JAMA Oncol* 2017; 3: 827-831. DOI: 10.1001/jamaoncol.2016.3834.
10. Roughley A, Damonte E, Taylor-Stokes G, et al. Impact of Brain Metastases on Quality of Life and Estimated Life Expectancy in Patients with Advanced Non-Small Cell Lung Cancer. *Value Health J Int Soc Pharmacoeconomics Outcomes Res* 2014; 17: A650.
11. Peters S, Bexelius C, Munk V, et al. The impact of brain metastasis on quality of life, resource utilization and survival in patients with non-small-cell lung cancer. *Cancer Treat Rev* 2016; 45: 139-162. 2016/03/29. DOI: 10.1016/j.ctrv.2016.03.009.
12. Sun A, Hu C, Wong SJ, et al. Prophylactic Cranial Irradiation vs Observation in Patients With Locally Advanced Non-Small Cell Lung Cancer: A Long-term Update of the NRG Oncology/RTOG 0214 Phase 3 Randomized Clinical Trial. *JAMA Oncol* 2019; 5: 847-855. 2019/03/15. DOI: 10.1001/jamaoncol.2018.7220.
13. Belderbos JSA, De Ruyscher DKM, De Jaeger K, et al. Phase 3 Randomized Trial of Prophylactic Cranial Irradiation With or Without Hippocampus Avoidance in SCLC (NCT01780675). *J Thorac Oncol* 2021; 16: 840-849. 2021/02/06. DOI: 10.1016/j.jtho.2020.12.024.

14. Jena A, Taneja S, Talwar V, et al. Magnetic resonance (MR) patterns of brain metastasis in lung cancer patients: correlation of imaging findings with symptom. *J Thorac Oncol* 2008; 3: 140-144. 2008/02/28. DOI: 10.1097/JTO.0b013e318161d775.
15. Mujoondar A, Austin JH, Malhotra R, et al. Clinical predictors of metastatic disease to the brain from non-small cell lung carcinoma: primary tumor size, cell type, and lymph node metastases. *Radiology* 2007; 242: 882-888. 2007/01/19. DOI: 10.1148/radiol.2423051707.
16. Ji Z, Bi N, Wang J, et al. Risk factors for brain metastases in locally advanced non-small cell lung cancer with definitive chest radiation. *Int J Radiat Oncol Biol Phys* 2014; 89: 330-337. 2014/04/15. DOI: 10.1016/j.ijrobp.2014.02.025.
17. Won YW, Joo J, Yun T, et al. A nomogram to predict brain metastasis as the first relapse in curatively resected non-small cell lung cancer patients. *Lung cancer* 2015; 88: 201-207. 2015/03/03. DOI: 10.1016/j.lungcan.2015.02.006.
18. Paget S. The distribution of secondary growths in cancer of the breast. 1889. *Cancer Metastasis Rev* 1989; 8: 98-101. 1989/08/01.
19. Srinivasan ES, Tan AC, Anders CK, et al. Salting the Soil: Targeting the Microenvironment of Brain Metastases. *Mol Cancer Ther* 2021; 20: 455-466. 2021/01/07. DOI: 10.1158/1535-7163.MCT-20-0579.
20. Tang WF, Wu M, Bao H, et al. Timing and Origins of Local and Distant Metastases in Lung Cancer. *J Thorac Oncol* 2021; 16: 1136-1148. 2021/03/17. DOI: 10.1016/j.jtho.2021.02.023.
21. Weidle UH, Birzele F and Nopora A. MicroRNAs as Potential Targets for Therapeutic Intervention With Metastasis of Non-small Cell Lung Cancer. *Cancer Genomics Proteomics* 2019; 16: 99-119. 2019/03/10. DOI: 10.21873/cgp.20116.
22. Dong J, Zhang Z, Gu T, et al. The role of microRNA-21 in predicting brain metastases from non-small cell lung cancer. *Onco Targets Ther* 2017; 10: 185-194. 2017/01/18. DOI: 10.2147/OTT.S116619.
23. Ramon YCS, Sese M, Capdevila C, et al. Clinical implications of intratumor heterogeneity: challenges and opportunities. *J Mol Med (Berl)* 2020; 98: 161-177. 2020/01/24. DOI: 10.1007/s00109-020-01874-2.
24. Parmar C, Leijenaar RT, Grossmann P, et al. Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer. *Sci Rep* 2015; 5: 11044. 2015/08/08. DOI: 10.1038/srep11044.
25. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer* 2012; 48: 441-446. 2012/01/20. DOI: 10.1016/j.ejca.2011.11.036.
26. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* 2017; 14: 749-762. DOI: 10.1038/nrclinonc.2017.141.
27. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 2014; 5: 4006. 2014/06/04. DOI: 10.1038/ncomms5006.

28. Khorrani M, Khunger M, Zagouras A, et al. Combination of Peri- and Intratumoral Radiomic Features on Baseline CT Scans Predicts Response to Chemotherapy in Lung Adenocarcinoma. *Radiol Artif Intell* 2019; 1: e180012. 2020/02/23. DOI: 10.1148/ryai.2019180012.
29. Coroller TP, Grossmann P, Hou Y, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol* 2015; 114: 345-350. 2015/03/10. DOI: 10.1016/j.radonc.2015.02.015.
30. Chen A, Lu L, Pu X, et al. CT-Based Radiomics Model for Predicting Brain Metastasis in Category T1 Lung Adenocarcinoma. *AJR Am J Roentgenol* 2019; 213: 134-139. 2019/04/02. DOI: 10.2214/AJR.18.20591.
31. Xu X, Huang L, Chen J, et al. Application of radiomics signature captured from pretreatment thoracic CT to predict brain metastases in stage III/IV ALK-positive non-small cell lung cancer patients. *J Thorac Dis* 2019; 11: 4516-4528. 2020/01/07. DOI: 10.21037/jtd.2019.11.01.
32. Sun F, Chen Y, Chen X, et al. CT-based radiomics for predicting brain metastases as the first failure in patients with curatively resected locally advanced non-small cell lung cancer. *Eur J Radiol* 2021; 134: 109411. 2020/11/28. DOI: 10.1016/j.ejrad.2020.109411.
33. De Ruyscher D, Dingemans AC, Praag J, et al. Prophylactic Cranial Irradiation Versus Observation in Radically Treated Stage III Non-Small-Cell Lung Cancer: A Randomized Phase III NVALT-11/DLCRG-02 Study. *J Clin Oncol* 2018; 36: 2366-2377. 2018/05/23. DOI: 10.1200/JCO.2017.77.5817.
34. Schoenmaekers J, Hofman P, Bootsma G, et al. Screening for brain metastases in patients with stage III non-small-cell lung cancer, magnetic resonance imaging or computed tomography? A prospective study. *Eur J Cancer* 2019; 115: 88-96. 2019/05/28. DOI: 10.1016/j.ejca.2019.04.017.
35. de Jong EEC, Hendriks LEL, van Elmpt W, et al. What you see is (not) what you get: tools for a non-radiologist to evaluate image quality in lung cancer. *Lung cancer* 2018; 123: 112-115. 2018/08/10. DOI: 10.1016/j.lungcan.2018.07.014.
36. Zwanenburg A, Vallieres M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* 2020; 191145. 2020/03/11. DOI: 10.1148/radiol.2020191145.
37. Jha AK, Mithun S, Jaiswar V, et al. Repeatability and reproducibility study of radiomic features on a phantom and human cohort. *Sci Rep* 2021; 11: 2055. 2021/01/23. DOI: 10.1038/s41598-021-81526-8.
38. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 3.6.0 ed. Vienna, Austria 2019.
39. Heus P, Damen J, Pajouheshnia R, et al. Uniformity in measuring adherence to reporting guidelines: the example of TRIPOD for assessing completeness of reporting of prediction model studies. *BMJ Open* 2019; 9: e025611. 2019/04/27. DOI: 10.1136/bmjopen-2018-025611.
40. Papanikolaou N, Matos C and Koh DM. How to develop a meaningful radiomic signature for clinical use in oncologic patients. *Cancer Imaging* 2020; 20: 33. 2020/05/03. DOI: 10.1186/s40644-020-00311-4.
41. Hochstenbag MMH, Twijnstra A, Hofman P, et al. MR-imaging of the brain of neurologic asymptomatic patients with large cell or adenocarcinoma of the lung. Does it influence

- prognosis and treatment? *Lung cancer* 2003; 42: 189-193. DOI: [https://doi.org/10.1016/S0169-5002\(03\)00291-5](https://doi.org/10.1016/S0169-5002(03)00291-5).
42. Hendriks LE, Bootsma GP, de Ruyscher DK, et al. Screening for brain metastases in patients with stage III non-small cell lung cancer: Is there additive value of magnetic resonance imaging above a contrast-enhanced computed tomography of the brain? *Lung cancer* 2013; 80: 293-297. 2013/03/23. DOI: 10.1016/j.lungcan.2013.02.006.
 43. Ganem J, Thureau S, Gardin I, et al. Delineation of lung cancer with FDG PET/CT during radiation therapy. *Radiat Oncol* 2018; 13: 219. 2018/11/14. DOI: 10.1186/s13014-018-1163-2.
 44. Bhalla AS, Das A, Naranje P, et al. Imaging protocols for CT chest: A recommendation. *Indian J Radiol Imaging* 2019; 29: 236-246. 2019/11/20. DOI: 10.4103/ijri.IJRI_34_19.
 45. Kakino R, Nakamura M, Mitsuyoshi T, et al. Comparison of radiomic features in diagnostic CT images with and without contrast enhancement in the delayed phase for NSCLC patients. *Phys Med* 2020; 69: 176-182. 2020/01/10. DOI: 10.1016/j.ejmp.2019.12.019.
 46. He L, Huang Y, Ma Z, et al. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. *Sci Rep* 2016; 6: 34921. 2016/10/11. DOI: 10.1038/srep34921.
 47. Bae KT. Intravenous contrast medium administration and scan timing at CT: considerations and approaches. *Radiology* 2010; 256: 32-61. 2010/06/25. DOI: 10.1148/radiol.10090908.
 48. Mali SA, Ibrahim A, Woodruff HC, et al. Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods. *J Pers Med* 2021; 11 2021/09/29. DOI: 10.3390/jpm11090842.
 49. Ibrahim A, Refaee T, Leijenaar RTH, et al. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *Plos One* 2021; 16: e0251147. 2021/05/08. DOI: 10.1371/journal.pone.0251147.
 50. Ibrahim A, Refaee T, Primakov S, et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers (Basel)* 2021; 13 2021/05/01. DOI: 10.3390/cancers13081848.
 51. Huber RM, De Ruyscher D, Hoffmann H, et al. Interdisciplinary multimodality management of stage III nonsmall cell lung cancer. *Eur Respir Rev* 2019; 28 2019/07/10. DOI: 10.1183/16000617.0024-2019.
 52. De Leyn P, Vansteenkiste J, Lievens Y, et al. Survival after trimodality treatment for superior sulcus and central T4 non-small cell lung cancer. *J Thorac Oncol* 2009; 4: 62-68. 2008/12/20. DOI: 10.1097/JTO.0b013e3181914d52.
 53. Park JE, Park SY, Kim HJ, et al. Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives. *Korean J Radiol* 2019; 20: 1124-1137. 2019/07/05. DOI: 10.3348/kjr.2018.0070.
 54. Pavic M, Bogowicz M, Wurms X, et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol* 2018; 57: 1070-1074. 2018/03/08. DOI: 10.1080/0284186X.2018.1445283.

55. Cherezov D, Goldgof D, Hall L, et al. Revealing Tumor Habitats from Texture Heterogeneity Analysis for Classification of Lung Cancer Malignancy and Aggressiveness. *Sci Rep* 2019; 9: 4500. 2019/03/16. DOI: 10.1038/s41598-019-38831-0.
56. Xu Y, Hosny A, Zeleznik R, et al. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clin Cancer Res* 2019; 25: 3266-3275. 2019/04/24. DOI: 10.1158/1078-0432.CCR-18-2495.

5.10 Supplementary materials

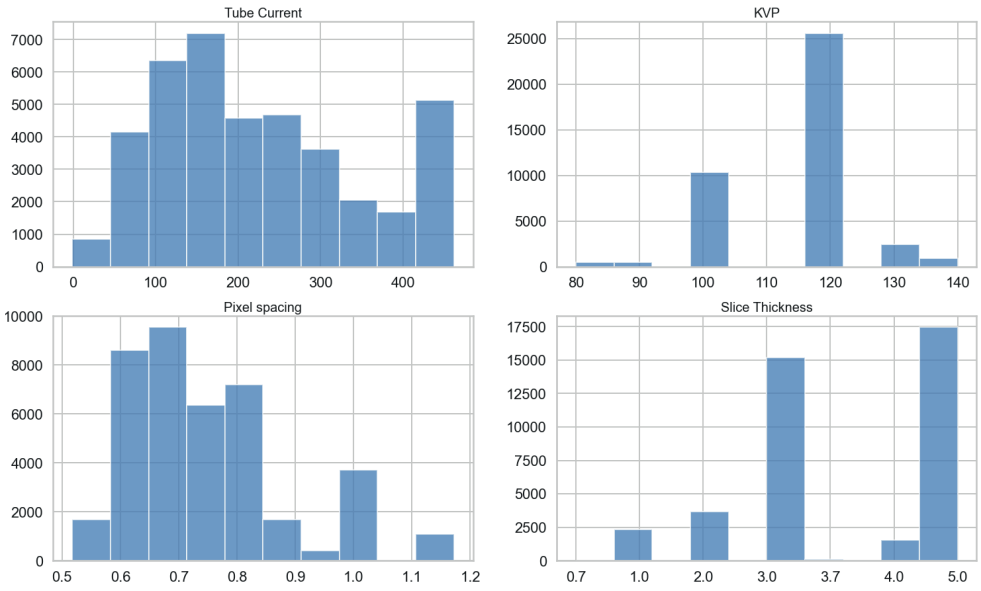


Figure S1. Tube current, peak kilo voltage peak (KVP) pixel spacing, and slice thickness histograms for the entire patient cohort.

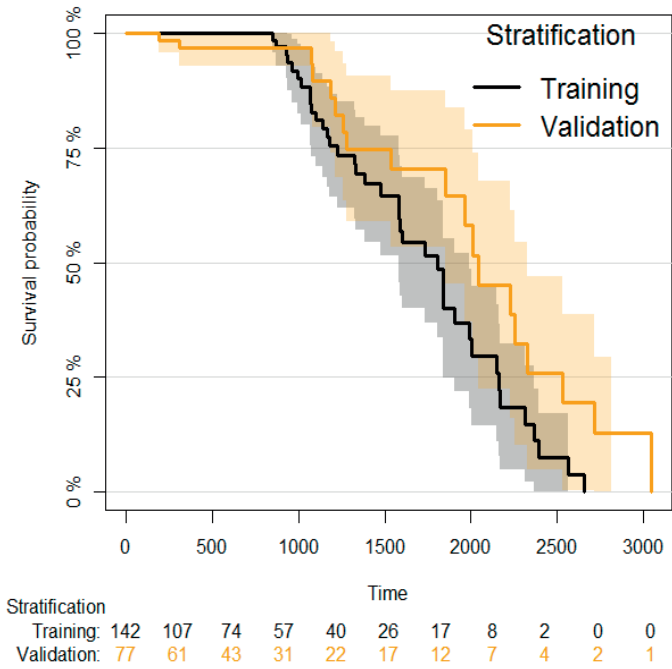


Figure S2. Reverse Kaplan-Meier plot of the training and validation dataset.

Section 1

The four selected radiomics features were: 1mm sigma LoG-filtered three dimensional GLSZM normalized size-zone non-uniformity (NSZNU), 2mm sigma LoG-filtered three-dimensional GLCM correlation, 2mm sigma LoG-filtered three-dimensional GLCM informational measure of correlation (IMC)1, and 2mm sigma LoG-filtered three-dimensional GLCM IMC2,

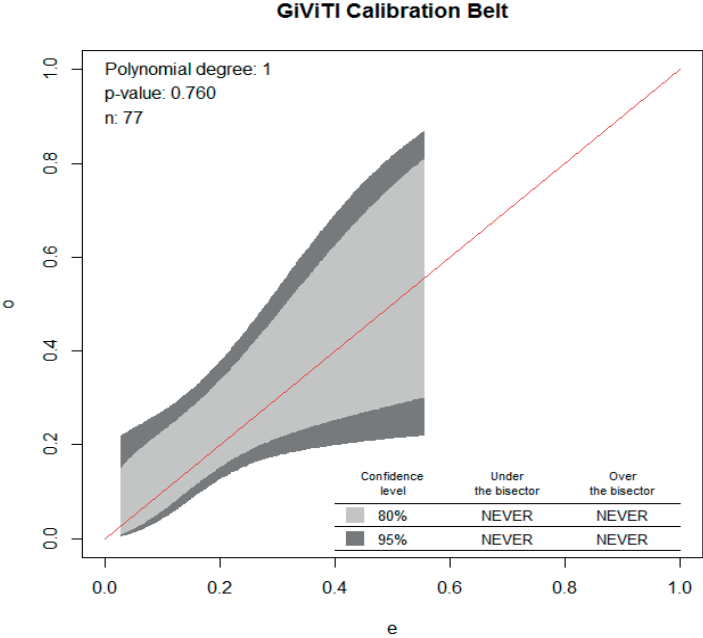


Figure S3. Calibration belt plot of the clinical model on the external validation dataset, with p-value of the calibration slope test.

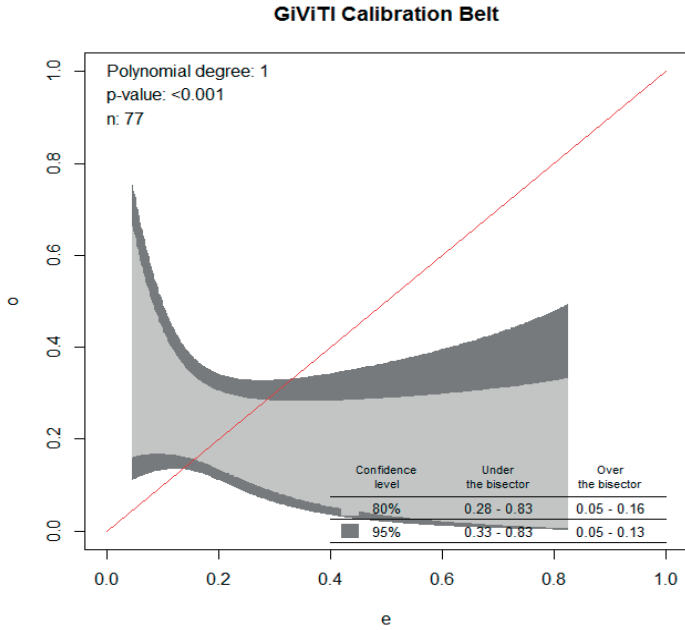


Figure S4. Calibration belt plot of the radiomics model on the external validation dataset, with p-value of the calibration slope test.

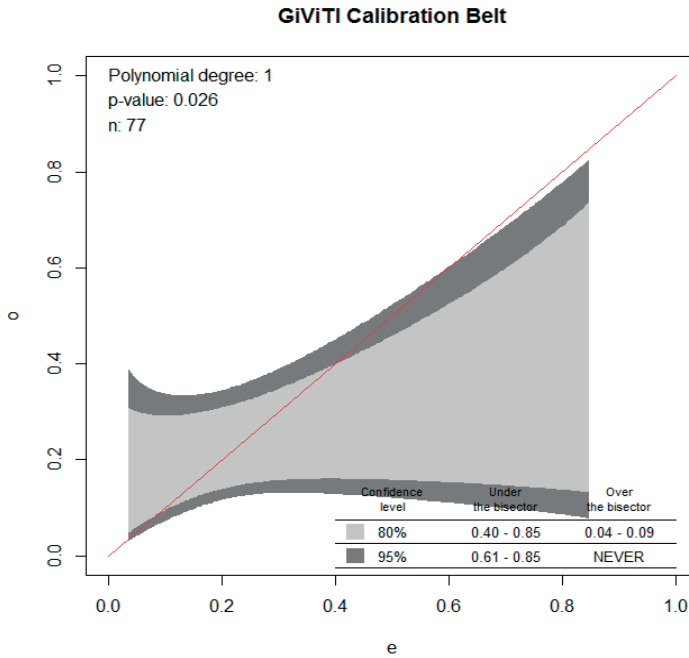


Figure S5. Calibration belt plot of the combined model on the external validation dataset, with p-value of the calibration slope test.

Table S1. TRIPOD adherence form.

Y=yes; N=no; R=referenced; NA=not applicable		[Study ID]
Title and abstract		
1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
i	The words developing/development, validation/validating, incremental/added value (or synonyms) are reported in the title	Y
ii	The words prediction, risk prediction, prediction model, risk models, prognostic models, prognostic indices, risk scores (or synonyms) are reported in the title	Y
iii	The target population is reported in the title	Y
iv	The outcome to be predicted is reported in the title	Y
2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	1
i	The objectives are reported in the abstract	Y
ii	Sources of data are reported in the abstract <i>E.g. Prospective cohort, registry data, RCT data.</i>	Y
iii	The setting is reported in the abstract <i>E.g. Primary care, secondary care, general population, adult care, or paediatric care. The setting should be reported for both the development and validation datasets, if applicable.</i>	Y
iv	A general definition of the study participants is reported in the abstract <i>E.g. patients with suspicion of certain disease, patients with a specific disease, or general eligibility criteria.</i>	Y
v	The overall sample size is reported in the abstract	Y
vi	The number of events (or % outcome together with overall sample size) is reported in the abstract <i>If a continuous outcome was studied, score Not applicable (NA).</i>	Y
vii	Predictors included in the final model are reported in the abstract. For validation studies of well-known models, at least the name/acronym of the validated model is reported <i>Broad descriptions are sufficient, e.g. 'all information from patient history and physical examination'. Check in the main text whether all predictors of the final model are indeed reported in the abstract.</i>	Y
viii	The outcome is reported in the abstract	Y
ix	Statistical methods are described in the abstract <i>For model development, at least the type of statistical model should be reported. For validation studies a quote like "model's discrimination and calibration was assessed" is considered adequate. If done, methods of updating should be reported.</i>	Y
x	Results for model discrimination are reported in the abstract <i>This should be reported separately for development and validation if a study includes both development and validation.</i>	Y
xi	Results for model calibration are reported in the abstract <i>This should be reported separately for development and validation if a study includes both development and validation.</i>	Y
xii	Conclusions are reported in the abstract <i>In publications addressing both model development and validation, there is no need for separate conclusions for both; one conclusion is sufficient.</i>	Y
3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	1
i	The background and rationale are presented	Y
ii	Reference to existing models is included (or stated that there are no existing models)	Y

3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	1
i	It is stated whether the study describes development and/or validation and/or incremental (added) value	Y
Methods		
4a	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	1
i	The study design/source of data is described <i>E.g. Prospectively designed, existing cohort, existing RCT, registry/medical records, case control, case series.</i> <i>This needs to be explicitly reported; reference to this information in another article alone is insufficient.</i>	Y
4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	1
i	The starting date of accrual is reported	Y
ii	The end date of accrual is reported	Y
iii	The length of follow-up and prediction horizon/time frame are reported, if applicable <i>E.g. "Patients were followed from baseline for 10 years" and "10-year prediction of..."; notably for prognostic studies with long term follow-up.</i> <i>If this is not applicable for an article (i.e. diagnostic study or no follow-up), then score Not applicable (NA).</i>	Y
5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	1
i	The study setting is reported (e.g. primary care, secondary care, general population) <i>E.g.: 'surgery for endometrial cancer patients' is considered to be enough information about the study setting.</i>	Y
ii	The number of centres involved is reported <i>If the number is not reported explicitly, but can be concluded from the name of the centre/centres, or if clearly a single centre study, score Yes.</i>	Y
iii	The geographical location (at least country) of centres involved is reported <i>If no geographical location is specified, but the location can be concluded from the name of the centre(s), score Yes.</i>	Y
5b	Describe eligibility criteria for participants.	1
i	In-/exclusion criteria are stated <i>These should explicitly be stated. Reasons for exclusion only described in a patient flow is not sufficient.</i>	Y
5c	Give details of treatments received, if relevant. <i>(i.e. notably for prognostic studies with long term follow-up)</i>	1
i	Details of any treatments received are described <i>This item is notably for prognostic modelling studies and is about treatment at baseline or during follow-up. The 'if relevant' judgment of treatment requires clinical knowledge and interpretation.</i> <i>If you are certain that treatment was not relevant, e.g. in some diagnostic model studies, score Not applicable.</i>	Y
6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	1
i	The outcome definition is clearly presented <i>This should be reported separately for development and validation if a publication includes both.</i>	Y
ii	It is described how outcome was assessed (including all elements of any composite, for example CVD [e.g. MI, HF, stroke]).	Y

Table S1. Continued

iii	It is described when the outcome was assessed (time point(s) since T0)	Y
6b	Report any actions to blind assessment of the outcome to be predicted.	1
i	Actions to blind assessment of outcome to be predicted are reported <i>If it is clearly a non-issue (e.g. all-cause mortality or an outcome not requiring interpretation), score Yes. In all other instances, an explicit mention is expected.</i>	Y
7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	1
i	All predictors are reported <i>For development, "all predictors" refers to all predictors that potentially could have been included in the "final" model (including those considered in any univariable analyses). For validation, "all predictors" means the predictors in the model being evaluated.</i>	Y
ii	Predictor definitions are clearly presented	Y
iii	It is clearly described how the predictors were measured	Y
iv	It is clearly described when the predictors were measured	Y
7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	1
i	It is clearly described whether predictor assessments were blinded for outcome <i>For predictors for which it is clearly a non-issue (e.g. automatic blood pressure measurement, age, sex) and for instances where the predictors were clearly assessed before outcome assessment, score Yes. For all other predictors an explicit mention is expected.</i>	Y
ii	It is clearly described whether predictor assessments were blinded for the other predictors	Y
8	Explain how the study size was arrived at.	1
i	It is explained how the study size was arrived at <i>Is there any mention of sample size, e.g. whether this was done on statistical grounds or practical/logistical grounds (e.g. an existing study cohort or data set of a RCT was used)?</i>	Y
9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	1
i	The method for handling missing data (predictors and outcome) is mentioned <i>E.g. Complete case (explicit mention that individuals with missing values have been excluded), single imputation, multiple imputation, mean/median imputation. If there is no missing data, there should be an explicit mention that there is no missing data for all predictors and outcome. If so, score Yes. If it is unclear whether there is missing data (from e.g. the reported methods or results), score No. If it is clear there is missing data, but the method for handling missing data is unclear, score No.</i>	Y
ii	If missing data were imputed, details of the software used are given <i>When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.</i>	Y
iii	If missing data were imputed, a description of which variables were included in the imputation procedure is given <i>When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.</i>	Y
iv	If multiple imputation was used, the number of imputations is reported <i>When under 9i explicit mentioning of no missing data, complete case analysis or no imputation applied, score Not applicable.</i>	Y

10a	Describe how predictors were handled in the analyses.	1
i	For continuous predictors it is described whether they were modelled as linear, nonlinear (type of transformation specified) or categorized <i>A general statement is sufficient, no need to describe this for each predictor separately. If no continuous predictors were reported, score Not applicable.</i>	Y
ii	For categorical or categorized predictors, the cut-points were reported <i>If no categorical or categorized predictors were reported, score Not applicable.</i>	Y
iii	For categorized predictors the method to choose the cut-points was clearly described <i>If no categorized predictors, score Not applicable.</i>	Y
10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	0
i	The type of statistical model is reported <i>E.g. Logistic, Cox, other regression model (e.g. Weibull, ordinal), other statistical modelling (e.g. neural network)</i>	Y
ii	The approach used for predictor selection <u>before</u> modelling is described <i>'Before modelling' means before any univariable or multivariable analysis of predictor-outcome associations. If no predictor selection before modelling is done, score Not applicable. If it is unclear whether predictor selection before modelling is done, score No. If it is clear there was predictor selection before modelling but the method was not described, score No.</i>	Y
iii	The approach used for predictor selection <u>during</u> modelling is described <i>E.g. Univariable analysis, stepwise selection, bootstrap, Lasso. 'During modelling' includes both univariable or multivariable analysis of predictor-outcome associations. If no predictor selection during modelling is done (so-called full model approach), score Not applicable. If it is unclear whether predictor selection during modelling is done, score No. If it is clear there was predictor selection during modelling but the method was not described, score No.</i>	NA
iv	Testing of interaction terms is described <i>If it is explicitly mentioned that interaction terms were not addressed in the prediction model, score Yes. If interaction terms were included in the prediction model, but the testing is not described, score No.</i>	N
v	Testing of the proportionality of hazards in survival models is described <i>If no proportional hazard model is used, score Not applicable.</i>	NA
vi	Internal validation is reported <i>E.g. Bootstrapping, cross validation, split sample. If the use of internal validation is clearly a non-issue (e.g. in case of very large data sets), score Yes. For all other situations an explicit mention is expected.</i>	Y
10c	For validation, describe how the predictions were calculated.	Not applicable
10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	1
	<i>These should be described in methods section of the paper (item 16 addresses the reporting of the results for model performance).</i>	
i	Measures for model discrimination are described <i>E.g. C-index / area under the ROC curve.</i>	Y
ii	Measures for model calibration are described <i>E.g. calibration plot, calibration slope or intercept, calibration table, Hosmer Lemeshow test, O/E ratio.</i>	Y

Table S1. Continued

iii	Other performance measures are described <i>E.g. R2, Brier score, predictive values, sensitivity, specificity, AUC difference, decision curve analysis, net reclassification improvement, integrated discrimination improvement, AIC.</i>	Y
10e	Describe any model updating (e.g., recalibration) arising from the validation, if done.	Not applicable
11	Provide details on how risk groups were created, if done. <i>If risk groups were not created, score this item as Yes.</i>	1
i	If risk groups were created, risk group boundaries (risk thresholds) are specified <i>Score this item separately for development and validation if a study includes both development and validation.</i> <i>If risk groups were not created, score this item as not applicable.</i>	Y
12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome and predictors.	Not applicable
Results		
13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	1
i	The flow of participants is reported	Y
ii	The number of participants with and without the outcome are reported <i>If outcomes are continuous, score Not applicable.</i>	NA
iii	A summary of follow-up time is presented <i>This notably applies to prognosis studies and diagnostic studies with follow-up as diagnostic outcome.</i> <i>If this is not applicable for an article (i.e. diagnostic study or no follow-up), then score Not applicable.</i>	Y
13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	1
i	Basic demographics are reported	Y
ii	Summary information is provided for all predictors included in the final developed/ validated model	Y
iii	The number of participants with missing data for predictors is reported	Y
iv	The number of participants with missing data for the outcome is reported	Y
13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	Not applicable
14a	Specify the number of participants and outcome events in each analysis.	1
i	The number of participants in each analysis (e.g. in the analysis of each model if more than one model is developed) is specified	Y
ii	The number of outcome events in each analysis is specified (e.g. in the analysis of each model if more than one model is developed) <i>If outcomes are continuous, score Not applicable.</i>	Y
14b	If done, report the unadjusted association between each candidate predictor and outcome.	1
i	The unadjusted associations between each predictor and outcome are reported <i>If any univariable analysis is mentioned in the methods but not in the results, score No.</i> <i>If nothing on univariable analysis (in methods or results) is reported, score this item as Not applicable.</i>	Y

15a	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	1
i	The regression coefficient (or a derivative such as hazard ratio, odds ratio, risk ratio) for each predictor in the model is reported	Y
ii	The intercept or the cumulative baseline hazard (or baseline survival) for at least one time point is reported	Y
15b	Explain how to use the prediction model.	0
i	An explanation (e.g. a simplified scoring rule, chart, nomogram of the model, reference to online calculator, or worked example) is provided to explain how to use the model for individualised predictions.	N
16	Report performance measures (with confidence intervals) for the prediction model. <i>These should be described in results section of the paper (item 10 addresses the reporting of the methods for model performance).</i>	1
i	A discrimination measure is presented <i>E.g. C-index / area under the ROC curve.</i>	Y
ii	The confidence interval (or standard error) of the discrimination measure is presented	Y
iii	Measures for model calibration are described <i>E.g. calibration plot, calibration slope or intercept, calibration table, Hosmer Lemeshow test, O/E ratio.</i>	Y
iv	Other model performance measures are presented <i>E.g. R2, Brier score, predictive values, sensitivity, specificity, AUC difference, decision curve analysis, net reclassification improvement, integrated discrimination improvement, AIC.</i>	Y
17	If done, report the results from any model updating (i.e., model specification, model performance, recalibration). <i>If updating was not done, score this TRIPOD item as 'Not applicable'.</i>	Not applicable
Discussion		
18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	1
i	Limitations of the study are discussed <i>Stating any limitation is sufficient.</i>	Y
19a	For validation, discuss the results with reference to performance in the development data, and any other validation data.	Not applicable
19b	Give an overall interpretation of the results considering objectives, limitations, results from similar studies and other relevant evidence.	1
i	An overall interpretation of the results is given	Y
20	Discuss the potential clinical use of the model and implications for future research.	1
i	The potential clinical use is discussed <i>E.g. an explicit description of the context in which the prediction model is to be used (e.g. to identify high risk groups to help direct treatment, or to triage patients for referral to subsequent care).</i>	Y
ii	Implications for future research are discussed <i>E.g. a description of what the next stage of investigation of the prediction model should be, such as "We suggest further external validation".</i>	Y
Other information		
21	Provide information about the availability of supplementary resources, such as study protocol, web calculator, and data sets.	
i	Information about supplementary resources is provided	Y
22	Give the source of funding and the role of the funders for the present study.	1

Table S1. Continued

i	The source of funding is reported or there is explicit mention that there was no external funding involved	Y
ii	The role of funders is reported or there is explicit mention that there was no external funding	Y
	Number of applicable TRIPOD items	30
	Number of TRIPOD items adhered	28
	OVERALL adherence to TRIPOD	93%

CHAPTER 6

6

Keywords

brain metastases, radiation necrosis, deep learning, radiomics, MRI, adverse

Predicting adverse radiation effects in brain tumors after stereotactic radiotherapy with deep learning and radiomics

Simon A. Keek^{1**}, Manon Beauque^{1**}, Sergey Primakov¹,
Henry C. Woodruff^{1,2}, Avishek Chatterjee¹, Janita E. van Timmeren³,
Martin Vallières^{4,5}, Lizza E.L. Hendriks⁶, Johannes Kraft^{3,7},
Nicolaus Andratschke³, Steve E. Braunstein⁸, Olivier Morin^{8**},
Philippe Lambin^{1,2* , **}

1 The D-Lab, Department of Precision Medicine, GROW- School for Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands

2 Department of Radiology and Nuclear Medicine, GROW – School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands

3 Department of Radiation Oncology, University Hospital of Zurich, University of Zurich, Zurich, Switzerland.

4 Medical Physics Unit, McGill University, Montréal, Canada

5 Department of Computer Science, Université de Sherbrooke, Sherbrooke, Canada

6 Department of Pulmonary Diseases, GROW – School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, The Netherlands

7 Department of Radiation Oncology, University Hospital Würzburg, Würzburg, Germany

8 Department of Radiation Oncology, University of California San Francisco, San Francisco, USA

**Equal contribution

6.1 Abstract

6.1.1 Introduction

There is a cumulative risk of 20-40% of developing brain metastases (BM) in solid cancers. Stereotactic radiotherapy (SRT) enables application of high focal doses of radiation to a volume, and is often used for BM treatment. However, SRT can cause adverse radiation effects (ARE) such as radiation necrosis, sometimes causing irreversible damage to the brain. It is therefore of clinical interest to identify patients at high-risk of developing ARE. We hypothesized that models trained with radiomics features, deep learning (DL) features, patient characteristics, or a combination, can predict ARE risk in patients with BM before SRT.

6.1.2 Methods

Gadolinium-enhanced T1-weighted MRIs and characteristics from patients treated with SRT for BM were collected for a training and testing cohort (N=1404), and a validation cohort (N=237) from a separate institute. From each lesion in the training set, radiomics features were extracted and used to train an extreme gradient boosting (XGBoost) model. A DL model was trained on the same cohort to make a separate prediction, and to extract the last layer of features. Different models using XGBoost were built using only radiomics features, DL features and patient characteristics, or a combination of them. Evaluation was performed using the area under the curve (AUC) of the receiver operating characteristic curve on the external dataset. Predictions for individual lesions and per-patient developing ARE were investigated.

6.1.3 Results

The best performing XGBoost model on a lesion-level was trained on a combination of radiomics features and DL features (AUC of 0.71, recall of 0.80). On a patient-level, a combination of radiomics features, DL features, and patient characteristics obtained the best performance (AUC of 0.72, recall of 0.84). The DL model achieved an AUC of 0.64 and recall of 0.85 per-lesion, and an AUC of 0.70 and recall of 0.60 per-patient.

6.1.4 Conclusion

Machine learning models built on radiomics features and DL features extracted from BM combined with patient characteristics show potential to predict ARE at the patient and lesion level. These models could be used in clinical decision making, informing patients on their risk of ARE, and allowing physicians to opt for different therapies.
radiation effects

6.2 Introduction

Brain metastases (BM) are the most common intracranial malignancies, accounting for more than 50% of all brain tumors, and occurring in 10 to over 40% of patients with solid malignancies (Walker, Robins and Weinfeld, 1985; Johnson and Young, 1996; Wen and Loeffler, 1999). BM occur most often in patients with lung cancer, breast cancer, and melanoma, which have a cumulative risk ranging from 20% to 40% of developing BM (Schouten *et al.*, 2002; Barnholtz-Sloan *et al.*, 2004; Rangachari *et al.*, 2015; Huber *et al.*, 2020). BM can be treated locally by surgery or radiotherapy, or with systemic anticancer therapy. Treatment depends on several factors, such as patient performance status, number and volume of metastases, presence of extracranial metastases, symptoms, and presumed efficacy of available systemic therapy ('Systemic therapy for brain metastases', 2018; Vogelbaum *et al.*, 2022). Radiotherapy of BM can be either stereotactic radiotherapy (SRT) or whole brain radiotherapy (WBRT), with SRT being guideline recommended treatment for a limited number of BM. As WBRT is associated with neurocognitive deterioration, SRT is increasingly used in multiple BM as well (McTyre, Scott and Chinnaiyan, 2013; Kraft *et al.*, 2019, 2021). SRT is either delivered in a single fraction, with stereotactic radiosurgery (SRS), or as fractionated stereotactic radiotherapy (FSRT), and results in a high dose within the target volume with a steep dose gradient to the surrounding healthy tissue (Badiyan, Regine and Mehta, 2016).

Even though most of the healthy brain is spared from high doses of radiation, a major shortcoming of SRT is a chance of high toxicity in the immediate surrounding tissues, which may lead to adverse radiation effects (ARE) such as radiation necrosis (RN), subacute edema, structural changes in the white matter, and vascular lesions (Walker *et al.*, 2014). ARE is a relatively late reaction to irradiation of healthy tissues where either reversible or irreversible injury has occurred (Sneed *et al.*, 2015). Risk of ARE after SRT and SRS is found to be similar, and ranges from 5-10% at patient level (Gerosa *et al.*, 2002; Lawrence *et al.*, 2010; Minniti *et al.*, 2014; Vellayappan *et al.*, 2018), or approximately 3% at lesion level (Sneed *et al.*, 2015). Known predictors of ARE are tumor volume, isodose volume, and previous SRT to the same lesion (Sneed *et al.*, 2015). ARE of the tumor area and tumor progression (TP) as two different post-therapeutic events require different treatment strategies: while steroids are often indicated for the initial treatment of ARE, true progression or relapse requires repeated radiotherapy, surgery, or effective intracranial systemic therapy for tumor control. Being able to differentiate between ARE and TP is therefore of utmost clinical interest.

Unfortunately, the (neurological) symptoms of ARE and TP are usually indistinguishable. Furthermore, the appearances of ARE and TP are very difficult to discern through qualitative radiological imaging, requiring multiple successive magnetic resonance images (MRI),

specialized MRI sequences such as perfusion-weighted or MR spectroscopy, and trained experts to evaluate the findings (Petrovich *et al.*, 2002; Vellayappan *et al.*, 2018). The clinical workflow is time- and labor-intensive, and while it is unfeasible to perform for every lesion, a definitive confirmation of the presence of ARE requires tissue acquisition (Vellayappan *et al.*, 2018).

SRT requires routine pretreatment MRI for accurate target volume delineation. This imaging provides a source of non-invasively acquired information about BM and brain phenotypes that could be investigated for their potential to determine before treatment which patient has a high risk of developing ARE. The early identification of these patients is an unmet clinical need which may help in clinical decision making by informing the patients of the risk of ARE, early risk stratification of patients that may develop ARE, and consideration of ARE-risk mitigating strategies such as deferring radiotherapy for central nervous system-penetrant systemic therapy.

Advanced quantitative medical image analysis methods such as radiomics and deep learning (DL) extract large amounts of imaging features and associate these with biological and/or clinical outcomes using machine learning (ML) techniques (Lambin *et al.*, 2012; Aerts *et al.*, 2014; Zhou *et al.*, 2018; Morin *et al.*, 2019; Avanzo *et al.*, 2020; Rogers *et al.*, 2020). Thus, radiological images from routine imaging procedures could potentially be used to non-invasively quantify the lesion phenotype providing clinically necessary information for patient management decisions. Several studies have indicated that MRI radiomics analysis is able to differentiate BM from glioblastoma (Abidin *et al.*, 2019; Dong *et al.*, 2020), to predict local recurrence (Huang *et al.*, 2020; Mouraviev *et al.*, 2020), to predict the origin of metastases (Ortiz-Ramón *et al.*, 2018; Kniep *et al.*, 2019), and overall survival (Bhatia *et al.*, 2019; Della Seta *et al.*, 2019). DL has also shown potential in predicting treatment response on brain MRI (Cho *et al.*, 2021). Moreover, DL and radiomics can have a complementary value, potentially establishing a more robust classifier (Parekh and Jacobs, 2019).

We hypothesize that models trained with radiomics features, DL features, and patient characteristics, or a combination thereof, can predict occurrence of ARE in patients with BM, both lesion specific and patient specific.

6.3 Materials and methods

6.3.1 Patient characteristics

All data from patients with BM treated with SRT between 1997 and 2017 for which imaging, outcome data, and patient data were available were collected retrospectively from the University of California-San Francisco (UCSF) medical center's picture archiving and communication system. Available imaging data, outcome data, and patient data of all patients with BM treated with SRS/SRT between 2014 and 2019 at the University Hospital Zürich (USZ) were collected retrospectively. The data included clinical and biological information for both the patient and the lesion. The eligibility criteria included radical treatment for metastatic brain cancer using Gamma Knife SRS for the UCSF patients and SRS/FSRT for the USZ patients. The inclusion of patients was regardless of the number of BM, but pathohistological or imaging-based confirmation of ARE during the follow-up was required in addition to pathohistological confirmation of the primary tumor. For the USZ cohort, in case of imaging-based suspicion of RN, positron emission tomography imaging was additionally used to exclude TP. The effort obtained ethical approval for observational research using anonymized linked care data for supporting medical purposes that are in the interests of individuals and the wider public. UCSF Institutional Review Board (<https://irb.ucsf.edu>) and Cantonal Ethics Committee Zurich approval with waiver of informed consent was obtained.

The UCSF dataset was divided randomly into sub-cohorts for training (70%) and testing (30%) while maintaining the ratios of events to non-events equal in both groups. The USZ dataset was used as an independent external validation dataset, i.e., it was entirely unseen by the models during the training and testing phases. The binary outcome used in training and validation was ARE per lesion, defined as either pathologically or imaging-based confirmation of RN occurring at any time after treatment. For both the UCSF and USZ patients, ARE was confirmed by histopathology when treated with open surgery. In all other cases, ARE was confirmed either at routine re-staging 3 months after radiotherapy for asymptomatic patients or at the onset of new symptoms. When patients presented new symptoms, imaging was performed usually after awaiting the effects of cortisone administration. As the time of BM formation is unknown, the outcome was not defined as right-censored. As every lesion is able to independently develop ARE after treatment, every lesion was considered to be an independent sample. The probability of ARE occurring for any lesion within a patient as an outcome was also investigated, whereby each patient was treated as an independent sample instead.

6.3.2 MR acquisition parameters and lesion segmentation

All images were axial gadolinium-enhanced T1-weighted MRI acquired prior to the treatment of BM. All included lesions were three-dimensionally delineated for curative

Gamma Knife SRS treatment purposes for the UCSF cohort and for curative SRS/ FSRT purposes for the USZ cohort according to local protocols by an experienced radiation oncologist. Figure 1 shows two T1-weighted gadolinium-enhanced MRI with lesions delineated for SRT purposes.

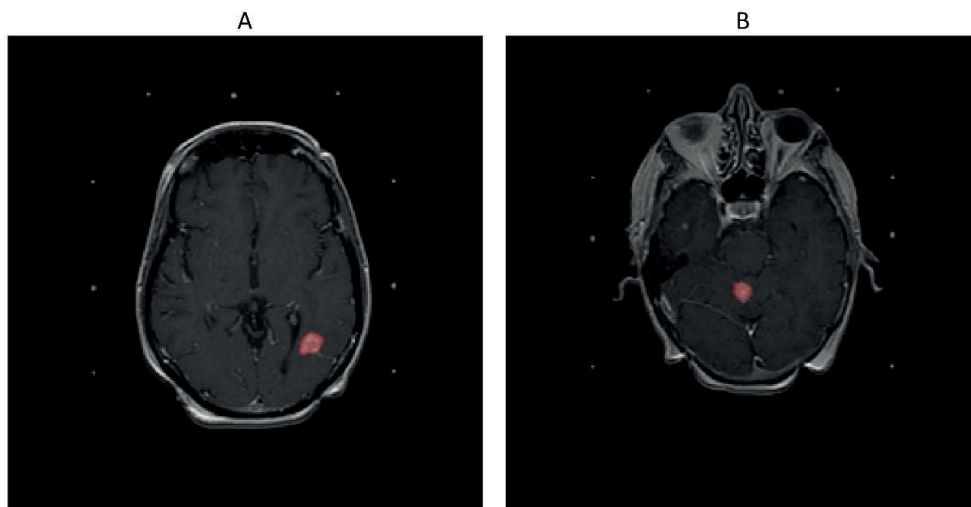


Figure 1. T1-weighted gadolinium enhanced MRIs of the brain with delineated in red (A) a lesion that developed adverse radiation effects after stereotactic radiotherapy and (B) a lesion that did not develop adverse radiation effects after stereotactic radiotherapy.

To perform segmentations of the brain and the ventricles on the entire dataset, an atlas-based segmentation strategy was chosen. To create the atlas in the MIM software package (MIM v. 6.9.4, MIM Software Inc., Cleveland, OH, USA), 50 randomly chosen MRI were manually segmented by an expert radiologist.

6.3.3 Pre-processing of brain MRI data

Bias-field correction was performed in the MIM software package using the N4 algorithm, which required brain segmentations (37). A bias field is a low-frequency signal distributed over an MR image, which is caused by inhomogeneities in the magnetic field of the MRI scanner. This causes shifts of intensity value ranges across the image (38). The ventricle mask was subtracted from the brain mask to obtain a white- and gray-matter segmentation. This segmentation was used to determine and correct the bias field present in the image using the N4 algorithm (37) using the MIM software package.

Following the bias correction, all remaining pre-processing, feature extraction, model training, and evaluation were performed in Python (version 3.7). The different Python packages used during this study can be found in Supplementary Table S1. Pre-processing

of MRI is essential for ML purposes, for reducing scanner dependence, and for ensuring reproducibility(39–41) As there is, to date, no consensus regarding the best way to pre-process MRI for our purposes, three different pre-processing workflows were applied and compared: “minimalist”, standardization, and “harmonization”. The descriptions of these pre-processing workflows can be found in the supplementary materials (Section 1)

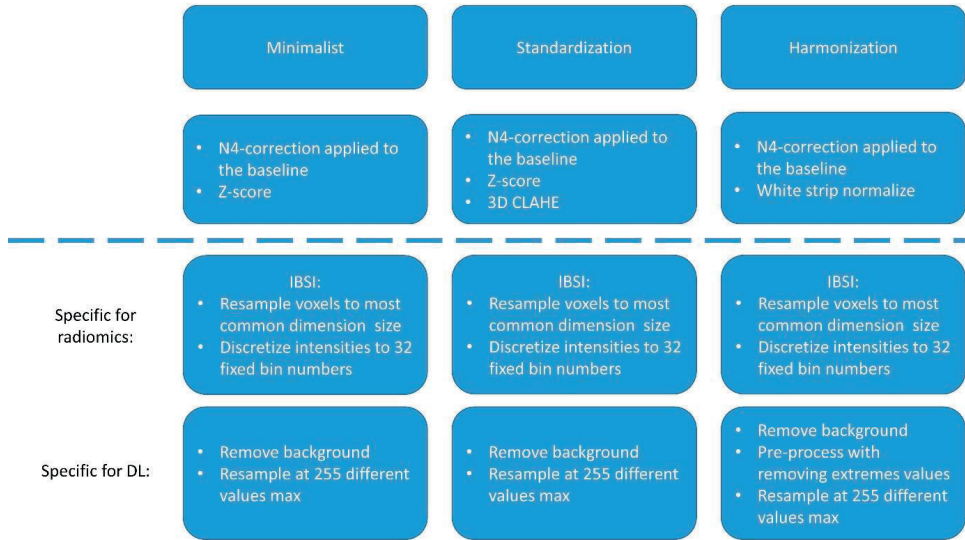


Figure 2. Pre-processing strategies for the “minimalist”, “standardization” and “harmonization” approaches.

6.3.4 Pre-processing for radiomics and feature

Feature extraction was performed according to the image biomarker standardization initiative (IBSI) guidelines (Duron et al., 2019; Carré et al., 2020; Zwanenburg et al., 2020) on the three different sets of processed MRI scans using the BM segmentations. All images were resampled to uniform 1x1x1 mm³ voxels using the ‘sitkBSpline’ interpolator to correct for differences in pixel size and slice spacing. The choice for voxel dimensions was made based on majority ruling, as it was found that most patients had a pixel spacing of ~1 mm. To achieve isotropic voxels, the choice for resampling in the z-direction was also chosen as 1mm. Pixel intensity values were resampled to a fixed number of 64 bins, as the number of gray levels was found to affect interchangeability of MRI radiomics features, and a fixed bin number of 64 has been found recommended in previous studies (Duron et al., 2019; Carré et al., 2020; Zwanenburg et al., 2020).

A total of 106 IBSI features were extracted from each segmentation. Features were extracted from the BM segmentations of the pre-processed images and can be divided into first-order intensity, histogram statistics, shape, and texture features. A full list and

description of the features can be found in the PyRadiomics documentation (Radiomic Features — pyradiomics documentation, 2019) , and a description of the feature groups in supplementary materials section 2.Pre-processing for deep learning

6.3.5 Pre-processing for deep learning

To inform the DL model on the location and extension of the lesions, lesion masks were used to highlight the ROI. A Gaussian smoothing filter was applied to the image, gradually decreasing the intensity values around the lesion from a factor of 1.0 to 0.2 to still include information of the voxels immediately around the lesion masks.

Otsu thresholding was performed to create a mask containing the brain and skull. This mask was used to determine the largest three-dimensional bounding box containing the brain and skull to crop the images. Anything outside this mask was defined as the image background, for which all pixel values were set at 0. For the “minimalist” and the “standardization” datasets, the intensities were resampled in a range between 0 and 255. Finally the scans were rescaled at 256x256x64 with spline interpolation order 3. As an example, the steps of the pre-processing workflow for the “minimalist” normalization is illustrated in figure 3.

6.3.6 Machine learning models

The mean and SD of each feature over the entire training population were determined. These values were used to apply z-score normalization to the features of the training, testing, and external validation datasets (Chatterjee *et al.*, 2019). Next, features with low variance (<0.01) were determined and excluded from the dataset. Lastly, the correlation between features was determined using absolute pairwise Spearman rank correlation. As highly correlated features (>0.85) were assumed to contain overlapping information about the outcome, the feature with the highest mean absolute correlation with the rest of the features was excluded. Lastly, supervised feature selection was performed through recursive feature elimination (RFE). RFE uses a ML algorithm to build a multivariate model and determine predictive performance using the currently selected features. It recursively drops and adds features, determining the optimal number of features and the selection of most predictive features.

An extreme gradient boosting (XGBoost) model was used for RFE and ARE prediction. A description of the XGBoost architecture, and the methodology to determine the optimal hyperparameters for the trained models, can be found in supplementary materials section 3.

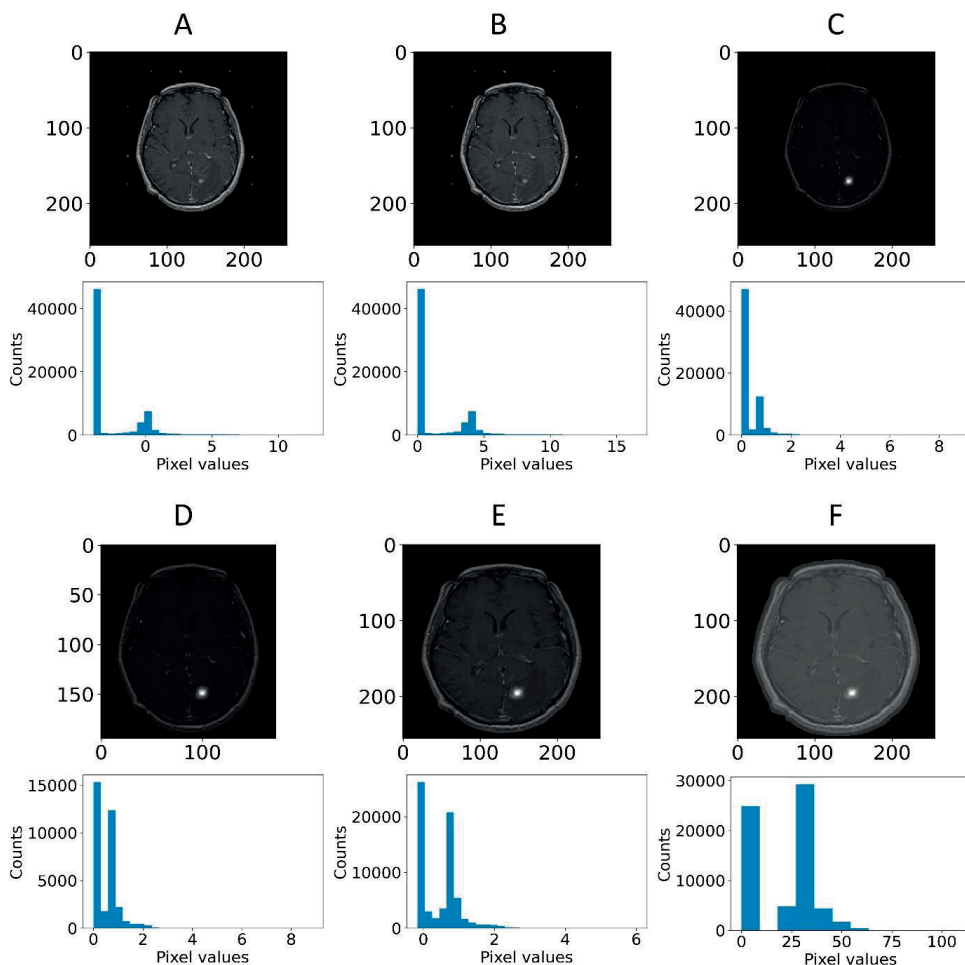


Figure 3. Example of pre-processing strategy: deep learning on the “minimalist” approach. The different steps of pre-processing were (A) z-score normalization, (B) shift to positive values only (C) pixel attenuations with Gaussian smoothing filtering (D) cropping around the largest bounding box and background set to 0 (E) resizing at 256x256 (F) rescaling the pixel value range to 0-255.

6.3.7 Deep learning model

An Xception three-dimensional model was trained and tested on the same datasets as the handcrafted radiomics-based model. Xception is the extreme version of an Inception model (Chollet 2017), which uses depthwise separable convolutions. The architecture can be found in supplementary figure 1. Adam optimization was used (Kingma and Ba, 2014) with an initial learning rate of 10^{-5} which updated the learning rate during training and used for loss function binary cross-entropy. This model produced a score ranging from 0 to 1, indicating the estimated probability that a lesion develops ARE. The area under the curve (AUC) of the receiver operating characteristic (ROC) was monitored on the test

dataset. The ROC displays the discriminative performance of a model expressed through the sensitivity and specificity as the threshold for binary classification is shifted. The AUC of the ROC is a metric from 0 to 1, where 1 means the model has perfect predictive performance and 0.5 is equivalent to guessing. To limit the imbalance of the outcomes to affect the model training, the model was only trained on lesions for those patients who had at least a single ARE, and tested on the scans of the patients who had ARE in the test dataset. To combine DL and radiomics, the last fully connected layer consisting of 256 features obtained after training the model was extracted. These features were then used to train a ML model similarly to using radiomics features, and used in models combining radiomics features and patient characteristics.

6.3.8 Clinical and treatment-related feature model

As the training and testing datasets contained patient characteristics not available in the external validation dataset, any feature not overlapping between these datasets was dropped. The list of remaining features was: primary tumor location, primary tumor histology, primary tumor controlled, extra-cranial metastases (ECM) present, patient age, patient sex, SRS to same location, prior external beam radiotherapy (EBRT), prior radiosurgery (RS), neurological symptoms, headaches, seizures, hypertension, diabetes, connective tissue disorder (CTD), Karnofsky performance status (KPS), prescription dose, and isodose-lines. For XGBoost to be able to handle categorical variables, one-hot encoding was performed on two categorical clinical features (primary tumor location and primary tumor histology).

Missing values were imputed using MissForest. MissForest is an imputation algorithm that uses RandomForest to train a model on the non-missing data for each feature with missing values, to predict the missing values. In the first iteration, all values are set to the mean value present for each variable (i.e., each column). Then, over multiple iterations, each data column with missing values will be predicted, using all the data except for the rows containing the missing values in question. This process is repeated over several iterations.

6.3.9 Metrics used for data analysis

Patient and tumor characteristics in the UCSF and USZ cohorts were assessed through a two-proportion z-test to test for significant differences in categorical variables between the cohorts, or the unpaired two-sample t-test to test for significant differences in numerical variables. For the latter, the assumptions of the data having a normal distribution and possessing the same variance in both cohorts were tested through Shapiro-Wilk's test and f-test, respectively. The significance level was set at 5%.

To determine which method ensured best performance for the radiomics-based and DL models, models were trained on the three different pre-processed datasets, and the best AUC of the ROC on the testing set was used to determine the best pre-processing methods for ML and DL separately. The 95% confidence intervals (CI) displayed on the ROC curves were obtained using bootstrapping (n=2000). For the radiomics-based model, the results were reported on the full train dataset and the entire test dataset. For the DL model, the results were reported on the balanced train dataset (which served to train the different DL models) and the full test dataset.

Once the best models were selected, the models were validated on the external dataset. Predictive performance of each model was expressed through the ROC curve, and its AUC, on the training, testing, and external data. By determining an optimal threshold value using the Youden’s J statistic (Youden, 1950) based on the training dataset, a binary classification was performed on the external dataset. From this binary classification, the balanced accuracy, precision, recall, and F1-score were determined. The confusion matrices were also derived from the binary classification. To determine model performance and to compare between models, the recall was investigated specifically, which is the proportion of true positives of the total number of true cases. As the number of events was relatively low and not missing any patients at risk of ARE is crucial, a high recall of the models was desirable. The CI obtained for all metrics were obtained using bootstrapping, resampling the results 2000 times. Moreover, analysis of the agreement prediction between the DL model and the radiomics-based model was performed. To give a prediction per patient, the maximum prediction of ARE among the different lesion predictions of the patient was selected. The ground truth to which the prediction was compared to was the ARE status of the patient, meaning the patient had at least one ARE lesion. An overview of the models tested can be found in figure 4.

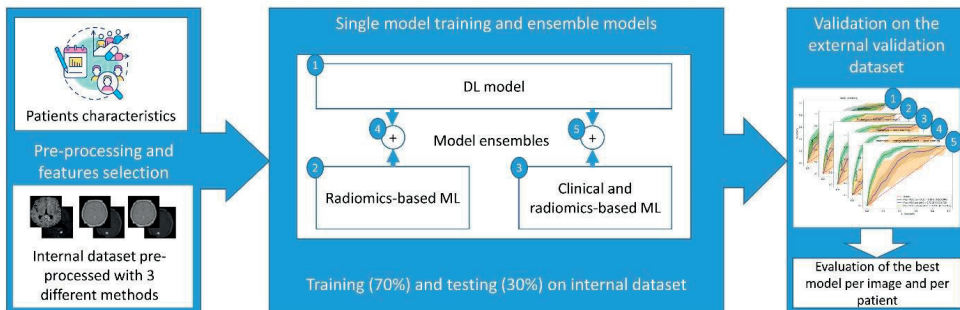


Figure 4. general workflow of the model training process: first the MRI data was pre-processed, using 3 pre-processing, selecting the most suitable pre-processed set of images according to the radiomics-based model or the DL model performance on the internal test dataset, then models are ensemble or trained separately, and finally the performance of each model is computed on the external dataset.

We evaluated on the external dataset for which cases the DL model and the best radiomics classifier obtained the same predictions and reported the amount of cases for which those models agreed on the label. The metrics based on the data for which the models agreed was also reported.

6.4 Results

6.4.1 Patient characteristics

A total of 1404 patients with 7974 lesions were included from UCSF, and 237 patients with 646 lesions from USZ. Table 1 shows an overview of the patient characteristics of the UCSF and USZ data. Significant differences between the proportion of males and females between the datasets ($P < 0.01$), median age ($P = 0.03$), KPS status ($P < 0.01$), and the number of lesions per patient at treatment ($P < 0.01$) were found. Furthermore, the proportions of primary tumor (lung, melanoma, and breast) were different between the datasets, and the data from USZ did not have kidney, GI, sarcoma, or other types of primary locations that were present in the UCSF dataset. For the histology of the primary tumor, only the melanoma histology subtype was found to be present in a significantly different proportion.

Table 1. Patient characteristics of University of California - San Francisco (UCSF) and University Hospital Zurich (USZ) datasets. P value of two-proportion z-test or unpaired two-sample t-test for significant differences between datasets was reported for each characteristic if applicable.

Patient/Tumor Characteristic		Total UCSF data	USZ data	P
		N = 1404	N = 237	
Sex (%)	Male	571 (41)	128 (54)	<0.01
	Female	833 (59)	109 (46)	
Median Age \pm SD		59 (13)	62 (12)	0.03
KPS (%)	80-100	1053 (75)	198 (83)	<0.01
	40-80	351 (25)	37 (16)	<0.01
	10-40	0 (0)	2 (1)	-
Primary tumor location (%)	Lung	530 (38)	136 (58)	<0.01
	Breast	357 (25)	27 (11)	<0.01
	Melanoma	272 (19)	74 (31)	<0.01
	Kidney	91 (7)	0 (0)	-
	Gastrointestinal	57 (4)	0 (0)	-
	Gynecologic	27 (2)	0 (0)	-
	Sarcoma	20 (1)	0 (0)	-
	Other	50 (4)	0 (0)	-

Histology primary tumor (%)	Adenocarcinoma	802 (57)	124 (52)	0.17
	Melanoma	272 (19)	74 (31)	<0.01
	Renal cell carcinoma	88 (6)	0 (0)	-
	Small cell carcinoma	44 (3)	0 (0)	-
	Squamous cell carcinoma	40 (3)	10 (4)	0.26
	Sarcoma	18 (1)	0 (0)	-
	Large cell carcinoma	9 (0.6)	2 (1)	0.72
	Bone carcinoma	8 (0.6)	0 (0)	-
	Adeno squamous carcinoma	6 (0.4)	0 (0)	-
	Broncho alveolar cell carcinoma	5 (0.4)	0 (0)	-
	Germ cell carcinoma	2 (0.1)	0 (0)	-
	Lymphoma	1 (0.1)	0 (0)	-
	Other/NOS	109 (8)	27 (11)	0.06
Primary controlled		974 (70)	149 (63)	0.05
ECM present		1097 (78)	190 (80)	0.48
#Lesions per patient at treatment	Median ± SD	3 (7)	2 (3)	<0.01
Symptoms	Headaches	437 (31)	31 (13)	<0.01
	Hypertension	407 (29)	0 (0)	<0.01
	Seizures	134 (10)	16 (7)	0.17
	Diabetes	98 (7)	13 (6)	0.4
	CTD	21 (2)	2 (1)	0.43
#Lesions in total		7974	646	-
#ARE cases (% of total lesions)		217 (2.7)	20 (3.1)	0.61
#Patients with ARE (% of total patients)		155 (11)	19 (8)	0.16
Prescription dose ± SD (Gy)		18.5 (1.5)	20 (5.0)	-

SD = standard deviation; KPS = Karnofsky performance score: 80-100 good performance, 50-70 medium performance, 10-40 bad performance; ECM = extracranial metastasis; BM = brain metastasis; CTD = connective tissue disorder; ARE = adverse radiation effect; Gy = gray.

6.4.2 Radiomics-based model and DL model results based on the three different preprocessing of the dataset

The best AUC on the test dataset for the radiomics-based models was found using the “harmonization” normalization, with an AUC of 0.76 (CI of 0.70-0.81), compared to 0.75 (CI of 0.70-0.80) and 0.73 (CI of 0.67-0.79) for “minimalist” and “standardization” methods, respectively.

The best AUC on the test dataset for the DL models was found using the “standardization” normalization, with an AUC of 0.72 (CI of 0.66-0.78), compared to 0.63 (CI of 0.57-0.70) and 0.65 (CI of 0.58-0.71) for “minimalist” and “harmonization” methods, respectively. Figure

5 shows the ROC-curves of the training and testing datasets for the three different pre-processing methods for radiomics based ML and for DL.

6.4.3 Results of the combined best performing models

We calculated the AUC and CI for each model combination on the external validation dataset. The DL model, built on images pre-processed with the “standardization” method, achieved an AUC of 0.64 (CI of 0.50-0.76). The model built on radiomics features, extracted from the images pre-processed with the “harmonization” method, achieved an AUC of 0.73 (CI of 0.63-0.83). The model was built on 20 features selected through RFE. Supplementary figure 2A provides an overview of the selected features and the corresponding importance in the XGBoost model. Supplementary table 2 provides an overview of the hyperparameters determined through grid search cross-validation. The model based on the combination of the DL features extracted from the last layer and radiomics features achieved an AUC of 0.71 (CI of 0.60-0.82). The model was built on 10 features selected through RFE. Supplementary figure 2B provides an overview of the selected features and the corresponding importance in the XGBoost model. The model built on radiomics features, extracted from images pre-processed with the “harmonization” method, combined with patient characteristic features achieved an AUC of 0.70 (CI of 0.57-0.80). The model was built on 19 features selected through RFE. Supplementary figure 2C provides an overview of the selected features and the corresponding importance in the XGBoost model. Finally, the model built on radiomics features, extracted from images pre-processed with the “harmonization” method, combined with DL features, extracted from images pre-processed with the “standardization” method, and patient characteristics achieved an AUC of 0.69 (CI of 0.56-0.81). The model was built on 20 features selected through RFE. Supplementary figure 2D provides an overview of the selected features and the corresponding importance in the XGBoost model. Figure 6 shows the ROC-curves with CI of the training, testing datasets, and validation datasets for these models.

The combination of radiomics and DL features achieved the highest combination of balanced accuracy and recall of 0.67 (CI of 0.56-0.76) and 0.80 (CI of 0.62-0.96), respectively, of the externally validated models. For a patient-level prediction, the DL model achieved an AUC of 0.70 (CI of 0.56-0.80), and the radiomics model an AUC of 0.72 (CI of 0.60-0.83). A combination of radiomics and DL achieved an AUC 0.71 (CI of 0.57-0.83), a combination of radiomics and patient characteristics an AUC of 0.71 (CI of 0.59-0.81), and a combination of radiomics features, DL features, and patient characteristics an AUC of 0.72 (CI of 0.58-0.84). The model combining radiomics features, DL features, and patient characteristics achieved the highest combination of balanced accuracy and recall of 0.65 (CI of 0.55-0.74) and 0.84 (CI of 0.65-1.00), respectively, of the externally validated models.

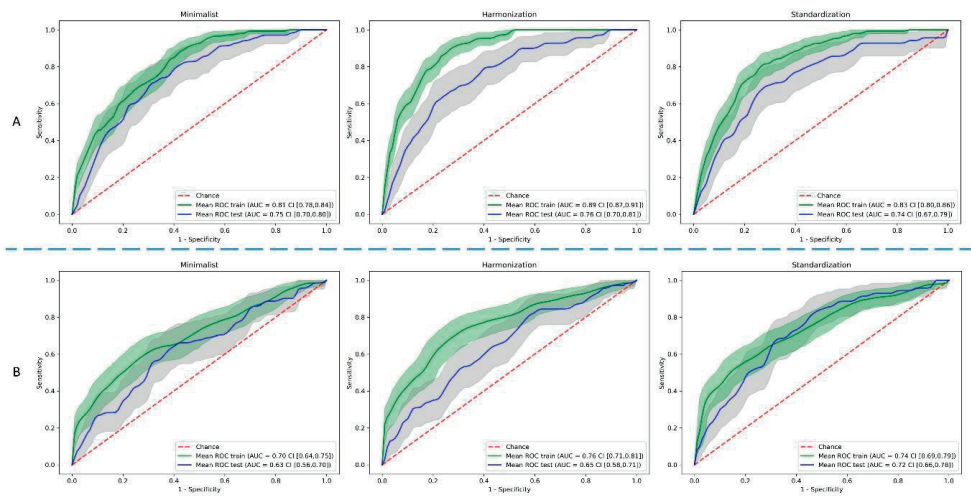


Figure 5. Comparison of predictive performance through receiver operating characteristic curves for (A) radiomics-based machine learning and (B) deep learning models using three different pre-processed image datasets. The shaded areas represent the 95% confidence intervals of the corresponding receiver operating characteristic curves.

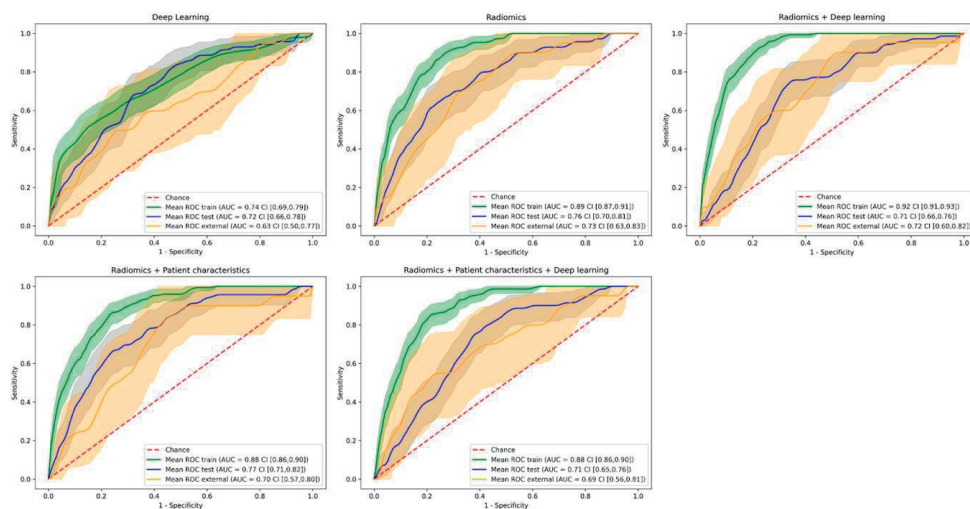


Figure 6. Receiver operating characteristic curves of the training, testing, and external validation datasets for the different model combinations. The shaded areas represent the 95% confidence intervals of the corresponding receiver operating characteristic curves.

The DL model predictions and the radiomics-based model predictions per lesion agreed for 32% of the external dataset. For the per-patient classification, the DL model predictions and the radiomics combined with clinical feature-based model predictions agreed for 19% of the external dataset. Because the amount of patients for which the models agreed was low (47 patients, 6 with ARE), no CI could be derived. Table 2 provides an overview of the AUC, balanced accuracy, precision, recall, and F1 score metrics for all DL and ML models, on a lesion and patient level, and for the agreed labels on the external validation. Supplementary tables 3 and 4 contain the same metrics for the training and testing datasets, respectively. The corresponding confusion matrices are in supplementary figure 3 and 4, respectively.

6.5 Discussion

Patients with BM treated with SRT are at risk of developing ARE, such as RN. Early identification of these patients can help in clinical decision making. The MRIs required for SRT planning provide an opportunity to identify these patients through quantitative imaging methods. In this large-scale study ML models that can successfully predict ARE were trained on T1-weighted MR imaging features from secondary brain tumors treated with SRT. As no consensus to harmonize MR images within and between centers exists, multiple methods were tested for the DL and ML pipeline, resulting in two optimal pre-processing methods (“harmonization” for the ML pipeline, “standardization” for the DL pipeline). A ML model trained with radiomics features combined with DL features yielded the highest predictive performance, with a combination of ROC AUC, balanced accuracy, and recall of 0.71, 0.67 and 0.80, respectively. At the patient level the best performing ML model was clearly a combination of radiomics, clinical (age at treatment, prior RS, and sex), and DL features achieving the highest predictive performance (AUC of 0.72), a balanced accuracy of 0.65, and recall of 0.84.

Performing an aggregate prediction (i.e., using only those predictions that agreed on the outcome) did not improve predictive performance for the lesion level prediction (AUC of 0.67), nor the binary prediction (balanced accuracy of 0.65). However, using this method the highest recall of 0.90 was achieved, making this method very robust in detecting true positives.

The models pave the way for clinical decision making of patients at risk of ARE before treatment. The information on the risk of an individual patient may be used by clinicians to inform patients of the risk of ARE when SRT is used as treatment. Furthermore, this information may be used to perform an early stratification of those patients at high risk, or may allow the patient and clinician to pursue alternative therapy, such as systemic therapy or alternate radiotherapy approaches (e.g., dose de-intensified SRT or WBRT) if the risk of ARE outweighs the possible benefits of SRT (Alvarez-Breckenridge *et al.*, 2022).

Table 2. AUC, balanced accuracy, precision, recall, and F1 metrics with CI on the external validation on patient and lesion levels.

Approaches	Per-lesion classification					Per-patient classification					
	AUC	Balanced accuracy	Precision	Recall	F1 score	Approaches	AUC	Balanced accuracy	Precision	Recall	F1 score
Best deep learning model	0.64 CI [0.50,0.76]	0.57 CI [0.48,0.64]	0.04 CI [0.02,0.05]	0.85 CI [0.67,1.00]	0.07 CI [0.04,0.10]	Best deep learning model	0.70 CI [0.56,0.83]	0.63 CI [0.52,0.73]	0.17 CI [0.09,0.25]	0.60 CI [0.39,0.78]	0.26 CI [0.16,0.37]
Best radiomics model	0.73 CI [0.63,0.83]	0.62 CI [0.51,0.74]	0.07 CI [0.03,0.11]	0.45 CI [0.23,0.67]	0.12 CI [0.05,0.19]	Best radiomics model	0.72 CI [0.60,0.83]	0.59 CI [0.51,0.69]	0.40 CI [0.09,0.75]	0.21 CI [0.05,0.43]	0.28 CI [0.07,0.48]
Radiomics and DL	0.71 CI [0.60,0.82]	0.67 CI [0.56,0.76]	0.05 CI [0.03,0.08]	0.80 CI [0.62,0.96]	0.10 CI [0.06,0.14]	Radiomics and DL	0.71 CI [0.57,0.83]	0.66 CI [0.54,0.77]	0.14 CI [0.07,0.22]	0.63 CI [0.40,0.84]	0.23 CI [0.13,0.34]
Radiomics and patient characteristics	0.70 CI [0.57,0.80]	0.62 CI [0.51,0.74]	0.06 CI [0.03,0.10]	0.50 CI [0.28,0.73]	0.11 CI [0.05,0.17]	Radiomics and patient characteristics	0.71 CI [0.59,0.81]	0.57 CI [0.48,0.68]	0.16 CI [0.04,0.30]	0.26 CI [0.08,0.47]	0.20 CI [0.05,0.35]
Radiomics, DL, and patient characteristics	0.69 CI [0.56,0.81]	0.64 CI [0.53,0.74]	0.05 CI [0.03,0.08]	0.70 CI [0.48,0.89]	0.09 CI [0.05,0.14]	Radiomics, DL, and patient characteristics	0.72 CI [0.58,0.84]	0.65 CI [0.55,0.74]	0.12 CI [0.07,0.17]	0.84 CI [0.65,1.00]	0.21 CI [0.13,0.29]
Agreed labels	0.67 CI [0.53,0.81]	0.65 CI [0.53,0.73]	0.07 CI [0.03,0.12]	0.90 CI [0.67,1.00]	0.13 CI [0.06,0.21]	Agreed labels	NA	NA	NA	NA	NA

To our knowledge, this is the first study that performs pre-treatment prediction of ARE using quantitative image analysis. Several studies have investigated the possibility of differentiating between tumor recurrence and RN after treatment, which is nominally similar in purpose to identify those patients who may have ARE. Zhang et al. (Zhang et al. 2018) used radiomics features extracted from four different MR sequences (T1, T1 post-contrast, T2, and fluid-attenuated inversion recovery (FLAIR)) at two different time-points during follow-up to differentiate RN from TP, confirmed pathologically. A model was built on a dataset of 87 patients with 97 lesions using 5 delta-radiomics features from T1 and T2 sequences. The AUC and binary prediction accuracy of the model were both 0.73. However, this result was obtained using leave-one-out cross validation, as no external validation was used. Similarly, Peng et al. created a model on radiomics features extracted from T1 and T2 FLAIR, on 66 patients with 77 lesions in total (Peng et al. 2018). The model was compared to a neuroradiologist's performance. No external validation was used, and instead a leave-one-out cross validation was performed, which gave an AUC of 0.81. The sensitivity and specificity of the neuroradiologist were 0.97 and 0.17, compared to 0.65 and 0.87 for the radiomics-based model. In (Park et al. 2021), the study compared the results obtained after training radiomics-based models using different MRI sequences (T1, T2 and apparent diffusion coefficient (ADC)). The models were trained using the data from 86 patients and tested on an external dataset of 41 patients. The best AUC was found on the ADC-based data with 0.80, when the other sequences had AUCs around 0.65. These results are similar or higher than the results obtained with our model, though within the range of the confidence intervals for the model based on radiomics and DL, and the lack of an external dataset on two of the studies makes the validity of these models difficult to determine (Peng *et al.*, 2018). Most other studies have a similar lack of external validation and total number of included patients, further making the results difficult to compare to the present study (Salvestrini *et al.*, 2022). These results show that the model presented in this study is able to perform similarly or even outperform models that perform classification (post treatment) instead of prediction (pre-treatment) of ARE.

One of the strengths of the present study is the large number of included patients and subsequent lesions, with 7974 lesions (2.7% ARE) of 1404 patients in training and testing, and 646 lesions (3.1% ARE) of 237 patients in the external validation. This provides a large volume of data for our models to train on, ensuring it covers the wide variability found between patients. In addition, the inclusion of an external validation is another strength, especially seeing the general lack of one in most other studies investigating ARE. This ensures that the reported result is not too optimistic, and shows that our model can be generalizable to populations from a different hospital in a different country, and even with different treatment from the training and testing set. While the difference in treatment between the training (exclusively SRS) and external validation (a mix of SRS and FSRT) may induce variability due to small differences in treatment planning for these methods,

literature has shown that these methods carry the same risk of ARE and were therefore considered interchangeable (Gerosa *et al.*, 2002; Lawrence *et al.*, 2010; Vellayappan *et al.*, 2018).

The large confidence interval on the external validation is partially due to the low number of positive findings in this dataset (n=20). This is because of the large imbalance in outcomes for both ARE and tumor failure. One of the major problems that may arise from this imbalance is a skewed view of predictive performance. However, this was addressed in the present study through multiple measures. The DL model was trained on a balanced subset of the data that only included patients that suffered at least 1 ARE. For ML, the XGBoost model was trained while scaling the weights of positive and negative classes and the respective proportion of the labels. Finally, through analysis of the confusion matrix, precision recall curves, and the recall metric, we ensured that the performance of the model was not entirely driven by labeling the data as the majority class.

While the models have been successfully validated on a dataset from an external center, further validation on multiple centers is required to ensure the models are generalizable. Future research could therefore focus on validating the present model on other datasets, potentially with recalibration of the model. At a later stage, a clinical trial to test the efficacy of the model is needed to be able to incorporate the model in a clinical setting.

In the present study, only one sequence of the MRI scan was used. Previous studies showed that a combination of radiomics computed on T1 and T2 sequences perform the best to differentiate ARE and TP (Peng *et al.*, 2018; Zhang *et al.*, 2018), and ADC sequence seems to also show a higher performance (Park *et al.*, 2021). Investigating more sequences in a future study may therefore improve performance of the imaging based models.

Lastly, for ARE (and to a lesser degree TP), treatment is one of the primary factors. In this study multiple dose treatment related variables have been included, such as prior treatments to the same patients, as well as dose variables and the volumes encompassing certain dose levels. However, a more thorough 'dosiomics' analysis would probably improve prediction of ARE. Liang *et al.* (2019) described a method to extract spatial and texture radiomics features from dose maps (Liang *et al.*, 2019). They found several radiomics features which have significant predictive value of radiation pneumonitis. Using a similar method for ARE in BM may result in improved prediction results. Our predictions could also be combined with models automatically classifying tumors and RN on brain MRI such as in (Zhang *et al.*, 2018), potentially strengthening the results of those studies.

6.6 Conclusion

Radiomics is able to predict lesions at high risk of ARE, especially when combined with DL features. When predicting ARE on a patient level, the highest performance was found using a combination of radiomics, DL, clinical, and treatment-related features. These models could potentially be used to aid clinical decision making for patients with BM treated with either Gamma Knife or EBRT.

6.7 Authors contribution

MB and SAK performed all the ML/DL analysis and wrote the manuscript. SAK, MV, SEB and OM collected and curated the imaging and patient data from UCSF. SP helped with the ML/DL analysis and study design. HCW supervised the progression of the project and the writing of this article and guaranteed the integrity of the analysis and results presented. AC and MV helped with the ML analysis. JEvT, JK, NA collected the imaging and patient data from USZ. LELH and SEB aided with the clinical aspects of the study. PL and OM devised the project's aim, and supervised the progression of the project. All authors have participated in writing the manuscript.

6.8 Abbreviations

ARE = Adverse radiation effects
 AUC = area under the curve
 AUCPR = Area under the precision recall-curve
 BM = brain metastasis
 CI = Confidence interval
 CLAHE = contrast limited adaptive histogram equalization
 CTD = Connective tissue disorder
 DICOM = Digital Imaging and Communications in Medicine
 DL = Deep learning
 EBRT = External beam radiotherapy
 ECM = Extra cranial metastases
 FSRT = Fractionated stereotactic radiotherapy
 GLCM = Gray-level co-occurrence matrix
 GLDZM = Gray-level distance-zone matrix
 GLRLM = Gray-level run length matrix
 GLSZM = Gray-level size-zone matrix
 HU = Hounsfield unit

IBSI = Image biomarker standardization initiative

KPS = Karnofsky performance score

ML = Machine learning

NGLDM = Neighborhood gray-level dependence matrix

NGTDM = Neighborhood gray-tone difference matrix

NRRD = Nearly raw raster data

RFE = Recursive feature elimination

RN = Radiation necrosis

ROC = receiver operating characteristic

ROI = Region of interest

RS = Radiosurgery

SD = standard deviation

SRS = Stereotactic radiosurgery

SRT = Stereotactic radiotherapy

TP = Tumor progression

UCSF = University of California – San Francisco

USZ = University Hospital Zürich

WBRT = Whole brain radiotherapy

XGBoost = extreme gradient boosting

6.9 Bibliography

- Abidin, A.Z. et al. (2019) 'Investigating a quantitative radiomics approach for brain tumor classification', in *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*. Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging, SPIE, pp. 36–45.
- Aerts, H.J.W.L. et al. (2014) 'Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach', *Nature communications*, 5, p. 4006.
- Alvarez-Breckenridge, C. et al. (2022) 'Emerging Systemic Treatment Perspectives on Brain Metastases: Moving Toward a Better Outlook for Patients', *American Society of Clinical Oncology educational book*. American Society of Clinical Oncology. Annual Meeting, 42, pp. 1–19.
- Amadasun, M. and King, R. (1989) 'Textural features corresponding to textural properties', *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 1264–1274. doi:10.1109/21.44046.
- Avanzo, M. et al. (2020) 'Machine and deep learning methods for radiomics', *Medical Physics*. doi:10.1002/mp.13678.
- Badiyan, S.N., Regine, W.F. and Mehta, M. (2016) 'Stereotactic Radiosurgery for Treatment of Brain Metastases', *Journal of oncology practice / American Society of Clinical Oncology*, 12(8), pp. 703–712.
- Barnholtz-Sloan, J.S. et al. (2004) 'Incidence proportions of brain metastases in patients diagnosed (1973 to 2001) in the Metropolitan Detroit Cancer Surveillance System', *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 22(14), pp. 2865–2872.
- Bhatia, A. et al. (2019) 'MRI radiomic features are associated with survival in melanoma brain metastases treated with immune checkpoint inhibitors', *Neuro-oncology*, 21(12), pp. 1578–1586.
- Carré, A. et al. (2020) 'Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics', *Scientific reports*, 10(1), p. 12340.
- Chatterjee, A. et al. (2019) 'Creating Robust Predictive Radiomic Models for Data From Independent Institutions Using Normalization', *IEEE Transactions on Radiation and Plasma Medical Sciences*, pp. 210–215. doi:10.1109/trpms.2019.2893860.
- Cho, J. et al. (2021) 'Deep Learning-Based Computer-Aided Detection System for Automated Treatment Response Assessment of Brain Metastases on 3D MRI', *Frontiers in oncology*, 11, p. 739639.
- Chollet, F. (2017) 'Xception: Deep learning with depthwise separable convolutions', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258.
- Della Seta, M. et al. (2019) 'A 3D quantitative imaging biomarker in pre-treatment MRI predicts overall survival after stereotactic radiation therapy of patients with a singular brain metastasis', *Acta radiologica*, 60(11), pp. 1496–1503.
- Dong, F. et al. (2020) 'Differentiation of supratentorial single brain metastasis and glioblastoma by using peri-enhancing oedema region-derived radiomic features and multiple classifiers', *European radiology*, 30(5), pp. 3015–3022.

- Duron, L. et al. (2019) 'Gray-level discretization impacts reproducible MRI radiomics texture features', *PloS one*, 14(3), p. e0213459.
- Galloway, M.M. (1975) 'Texture analysis using gray level run lengths', *Computer Graphics and Image Processing*, pp. 172–179. doi:10.1016/s0146-664x(75)80008-6.
- Gerosa, M. et al. (2002) 'Gamma knife radiosurgery for brain metastases: a primary therapeutic option', *Journal of Neurosurgery*, pp. 515–524. doi:10.3171/jns.2002.97.supplement_5.0515.
- Haralick, R.M., Shanmugam, K. and Dinstein, I. 'hak (1973) 'Textural Features for Image Classification', *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 610–621. doi:10.1109/tsmc.1973.4309314.
- Huang, C.-Y. et al. (2020) 'Radiomics as prognostic factor in brain metastases treated with Gamma Knife radiosurgery', *Journal of neuro-oncology*, 146(3), pp. 439–449.
- Huber, R.M. et al. (2020) 'Brigatinib in Crizotinib-Refractory ALK+ NSCLC: 2-Year Follow-up on Systemic and Intracranial Outcomes in the Phase 2 ALTA Trial', *Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer*, 15(3). doi:10.1016/j.jtho.2019.11.004.
- Johnson, J.D. and Young, B. (1996) 'Demographics of Brain Metastasis', *Neurosurgery Clinics of North America*, pp. 337–344. doi:10.1016/s1042-3680(18)30365-6.
- Juntu, J. et al. (2005) 'Bias Field Correction for MRI Images', *Advances in Soft Computing*, pp. 543–551. doi:10.1007/3-540-32390-2_64.
- Kingma, D.P. and Ba, J. (2014) 'Adam: A Method for Stochastic Optimization', *arXiv [cs.LG]*. Available at: <http://arxiv.org/abs/1412.6980>.
- Kniep, H.C. et al. (2019) 'Radiomics of Brain MRI: Utility in Prediction of Metastatic Tumor Type', *Radiology*, 290(2), pp. 479–487.
- Kraft, J. et al. (2019) 'Stereotactic Radiosurgery for Multiple Brain Metastases', *Current treatment options in neurology*, 21(2), p. 6.
- Kraft, J. et al. (2021) 'Management of multiple brain metastases: a patterns of care survey within the German Society for Radiation Oncology', *Journal of neuro-oncology*, 152(2), pp. 395–404.
- Lambin, P. et al. (2012) 'Radiomics: extracting more information from medical images using advanced feature analysis', *European journal of cancer*, 48(4), pp. 441–446.
- Lawrence, Y.R. et al. (2010) 'Radiation Dose–Volume Effects in the Brain', *International Journal of Radiation Oncology*Biophysics*, pp. S20–S27. doi:10.1016/j.ijrobp.2009.02.091.
- Liang, B. et al. (2019) 'Dosiomics: Extracting 3D Spatial Features From Dose Distribution to Predict Incidence of Radiation Pneumonitis', *Frontiers in oncology*, 9, p. 269.
- Masoudi, S. et al. (2021) 'Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research', *Journal of medical imaging (Bellingham, Wash.)*, 8(1), p. 010901.
- McTyre, E., Scott, J. and Chinnaiyan, P. (2013) 'Whole brain radiotherapy for brain metastasis', *Surgical neurology international*, 4(Suppl 4), pp. S236–44.
- Minniti, G. et al. (2014) 'Fractionated stereotactic radiosurgery for patients with brain metastases', *Journal of neuro-oncology*, 117(2), pp. 295–301.

- Moradmand, H., Aghamiri, S.M.R. and Ghaderi, R. (2020) 'Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma', *Journal of applied clinical medical physics / American College of Medical Physics*, 21(1), pp. 179–190.
- Morin, O. et al. (2019) 'Integrated models incorporating radiologic and radiomic features predict meningioma grade, local failure, and overall survival', *Neuro-oncology advances*, 1(1), p. vdz011.
- Mouraviev, A. et al. (2020) 'Use of radiomics for the prediction of local control of brain metastases after stereotactic radiosurgery', *Neuro-oncology*, 22(6), pp. 797–805.
- Nyúl, L.G., Udupa, J.K. and Zhang, X. (2000) 'New variants of a method of MRI scale standardization', *IEEE transactions on medical imaging*, 19(2), pp. 143–150.
- Ortiz-Ramón, R. et al. (2018) 'Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study', *European radiology*, 28(11), pp. 4514–4523.
- Parekh, V.S. and Jacobs, M.A. (2019) 'Deep learning and radiomics in precision medicine', *Expert review of precision medicine and drug development*, 4(2), pp. 59–72.
- Park, Y.W. et al. (2021) 'Differentiation of recurrent glioblastoma from radiation necrosis using diffusion radiomics with machine learning model development and external validation', *Scientific reports*, 11(1), p. 2913.
- Peng, L. et al. (2018) 'Distinguishing True Progression From Radionecrosis After Stereotactic Radiation Therapy for Brain Metastases With Machine Learning and Radiomics', *International journal of radiation oncology, biology, physics*, 102(4), pp. 1236–1243.
- Petrovich, Z. et al. (2002) 'Survival and pattern of failure in brain metastasis treated with stereotactic gamma knife radiosurgery', *Journal of neurosurgery*, 97(5 Suppl), pp. 499–506.
- Primakov, S. et al. (2022) 'Precision-medicine-toolbox: An open-source python package for facilitation of quantitative medical imaging and radiomics analysis', *arXiv [eess.IV]*. Available at: <http://arxiv.org/abs/2202.13965>.
- Radiomic Features — pyradiomics v3.0.1.post9+gdfe2c14 documentation (2019). Available at: <https://pyradiomics.readthedocs.io/en/latest/features.html> (Accessed: 21 October 2021).
- Rangachari, D. et al. (2015) 'Brain metastases in patients with EGFR -mutated or ALK -rearranged non-small-cell lung cancers', *Lung Cancer*, pp. 108–111. doi:10.1016/j.lungcan.2015.01.020.
- Reinhold, J.C. et al. (2018) 'Evaluating the Impact of Intensity Normalization on MR Image Synthesis'. Available at: <http://arxiv.org/abs/1812.04652> (Accessed: 21 October 2021).
- Rogers, W. et al. (2020) 'Radiomics: from qualitative to quantitative imaging', *The British journal of radiology*, 93(1108), p. 20190948.
- Salvestrini, V. et al. (2022) 'The role of feature-based radiomics for predicting response and radiation injury after stereotactic radiation therapy for brain metastases: A critical review by the Young Group of the Italian Association of Radiotherapy and Clinical Oncology (yAIRO)', *Translational Oncology*, p. 101275. doi:10.1016/j.tranon.2021.101275.
- Schouten, L.J. et al. (2002) 'Incidence of brain metastases in a cohort of patients with carcinoma of the breast, colon, kidney, and lung and melanoma', *Cancer*, 94(10), pp. 2698–2705.

- Sneed, P.K. et al. (2015) 'Adverse radiation effect after stereotactic radiosurgery for brain metastases: incidence, time course, and risk factors', *Journal of neurosurgery*, 123(2), pp. 373–386.
- 'Statistical normalization techniques for magnetic resonance imaging' (2014) *NeuroImage: Clinical*, 6, pp. 9–19.
- Sun, C. and Wee, W.G. (1982) 'Neighboring gray level dependence matrix for texture classification', *Computer Graphics and Image Processing*, p. 297. doi:10.1016/0146-664x(82)90093-4.
- 'Systemic therapy for brain metastases' (2018) in *Handbook of Clinical Neurology*. Elsevier, pp. 137–153.
- Thibault, G. et al. (2009) 'Texture indexes and gray level size zone matrix. Application to cell nuclei classification', in *10th International Conference on Pattern Recognition and Information Processing, PRIP 2009*, pp. 140–145.
- Thibault, G., Angulo, J. and Meyer, F. (2011) 'Advanced statistical matrices for texture characterization: Application to DNA chromatin and microtubule network classification', *2011 18th IEEE International Conference on Image Processing [Preprint]*. doi:10.1109/icip.2011.6116401.
- Tustison, N.J. et al. (2010) 'N4ITK: improved N3 bias correction', *IEEE transactions on medical imaging*, 29(6), pp. 1310–1320.
- Um, H. et al. (2019) 'Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets', *Physics in medicine and biology*, 64(16), p. 165011.
- Vellayappan, B. et al. (2018) 'Diagnosis and Management of Radiation Necrosis in Patients With Brain Metastases', *Frontiers in oncology*, 8, p. 395.
- Vogelbaum, M.A. et al. (2022) 'Treatment for Brain Metastases: ASCO-SNO-ASTRO Guideline', *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 40(5), pp. 492–516.
- Walker, A.E., Robins, M. and Weinfeld, F.D. (1985) 'Epidemiology of brain tumors: The national survey of intracranial neoplasms', *Neurology*, pp. 219–219. doi:10.1212/wnl.35.2.219.
- Walker, A.J. et al. (2014) 'Postradiation imaging changes in the CNS: how can we differentiate between treatment effect and disease progression?', *Future oncology*, 10(7), pp. 1277–1297.
- Wen, P.Y. and Loeffler, J.S. (1999) 'Management of brain metastases', *Oncology*, 13(7), pp. 941–54, 957–61; discussion 961–2, 9.
- Youden, W.J. (1950) 'Index for rating diagnostic tests', *Cancer*, 3(1), pp. 32–35.
- Zhang, Z. et al. (2018) 'A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from MR images', *European radiology*, 28(6), pp. 2255–2263.
- Zhou, M. et al. (2018) 'Radiomics in Brain Tumor: Image Assessment, Quantitative Feature Descriptors, and Machine-Learning Approaches', *American Journal of Neuroradiology*, pp. 208–216. doi:10.3174/ajnr.a5391.
- Zuiderveld, K. (1994) 'Contrast Limited Adaptive Histogram Equalization', *Graphics Gems*, pp. 474–485. doi:10.1016/b978-0-12-336156-1.50061-6.

Zwanenburg, A. et al. (2020) 'The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping', *Radiology*, 295(2), pp. 328–338.

6.10 Supplementary materials

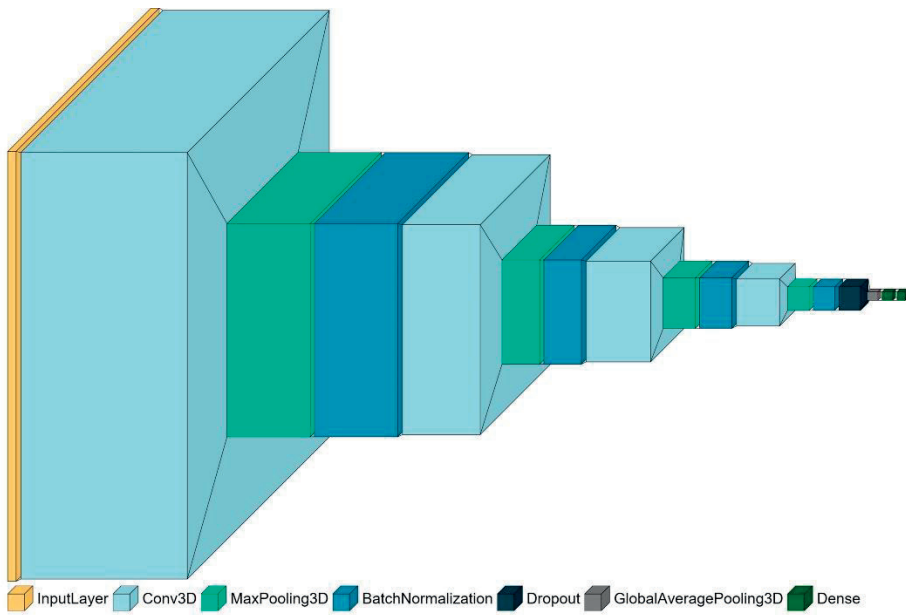


Figure 1. Architecture of Xception 3D.

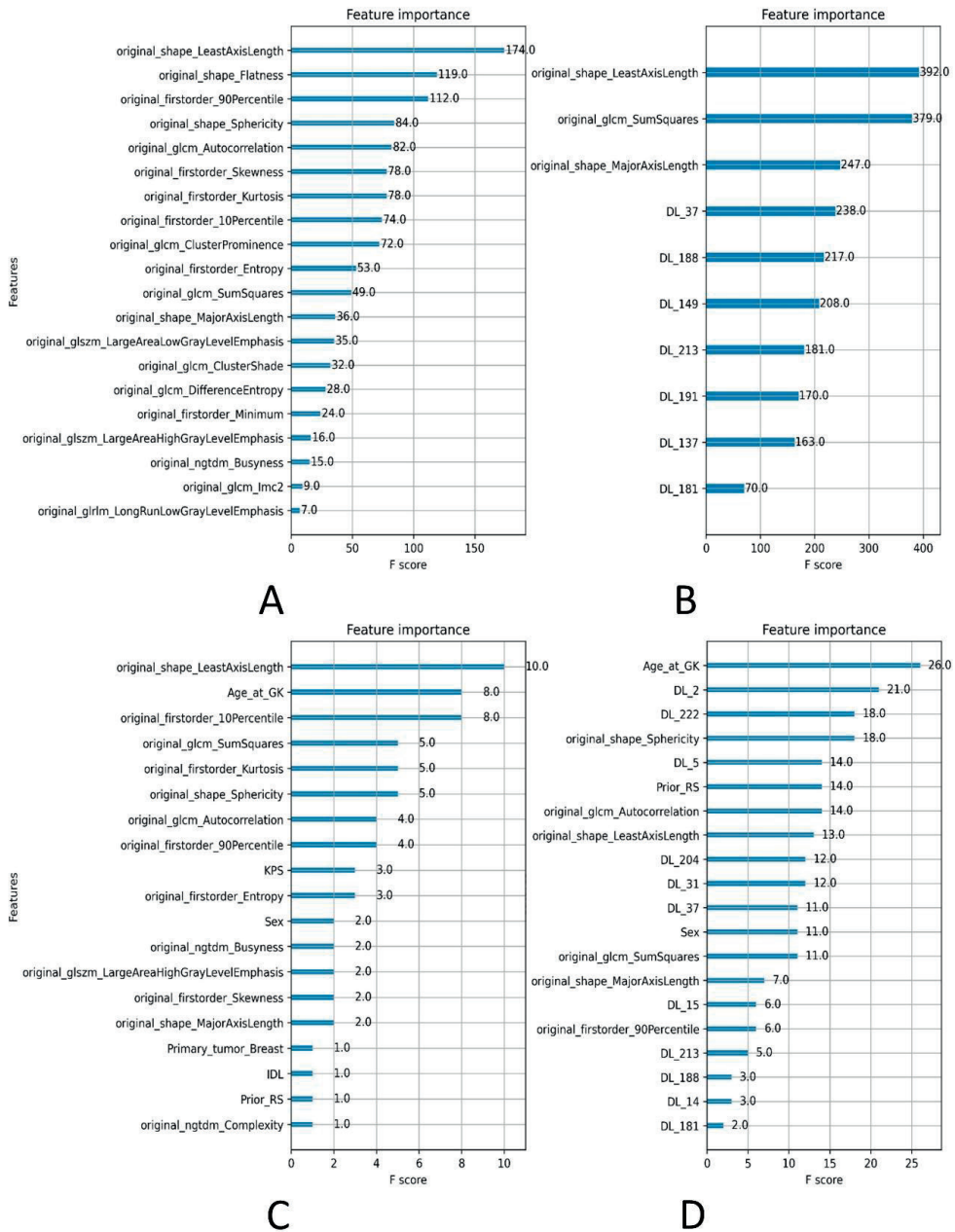


Figure 2. Feature importance lists of the ML models, respectively: (A) radiomics, (B) radiomics and deep learning, (C) radiomics and patient characteristics, and (D) radiomics, patient characteristics, deep learning.

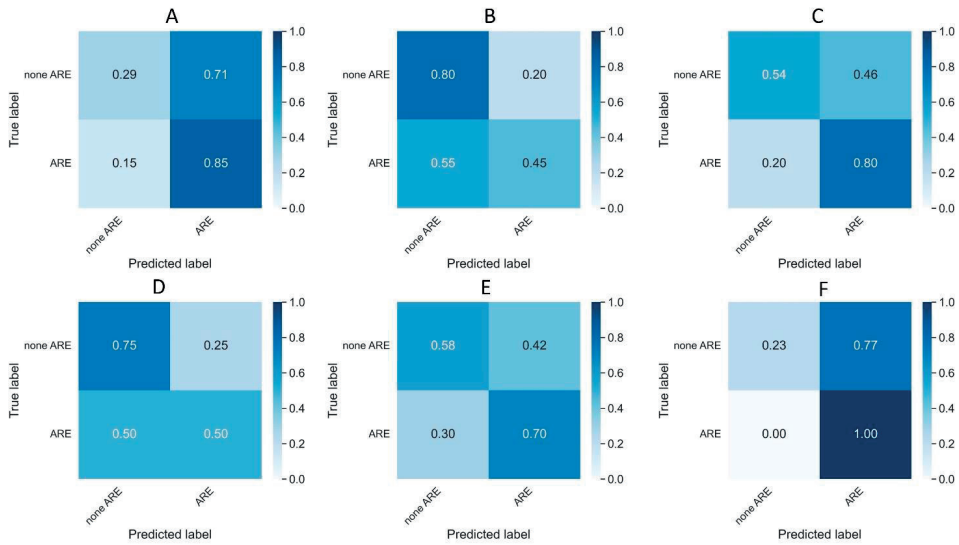


Figure 3. Top 3 missing variables in training (DESIGN) cohort.

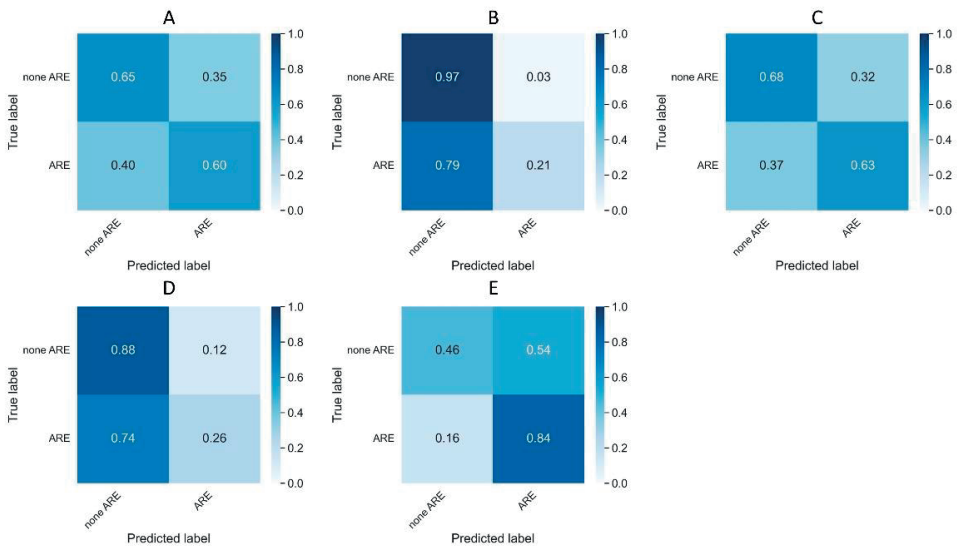


Figure 4. Normalized confusion matrices on the external validation dataset per patient for the following approaches: (A) DL, (B) radiomics, (C) radiomics and DL, (D) patient characteristics and radiomics, and (E) radiomics, DL and patient characteristics.

Table 1. Python packages used and their versions.

purpose	packages	versions
pre-processing	imutils	0.5.4
	intensity-normalization	2.0.2
	numpy	1.19.2
	opencv	4.1.0.25
	os	n/a
	pandas	0.25.0
	pydicom	2.2.2
	scikit-image	0.17.2
	scikit-learn	0.24.2
	scipy	1.5.2
	simpleITK	2.1.1
deep learning	keras	2.3.1
	tensorflow-gpu	2.1.0
feature processing and calculation	precision-medicine-toolbox	0.0.0
	missingpy	0.2.0
	pyradiomics	3.0.1
machine learning	xgboost	1.5.1
statistics	statsmodels	0.13.0
visualisation	matplotlib	3.3.4

Table 2. Overview of hyperparameters optimized through gridsearch cross-validation.

parameters/data	radiomics only	radiomics + patient characteristics	radiomics + deep learning	radiomics + patient characteristics + deep learning
gamma	0,3	0,3	0,3	0,3
learning rate	0,01	0,1	0,01	0,1
max depth	3	3	4	1
min child weight	1	1	1	5
n estimators	173	10	173	227
number of features selected	20	10	20	20

Table 3. AUC, balanced accuracy, precision, recall, and F1 metrics with CI on the training on patient and lesion levels.

Approaches	Per-lesion classification					Per-patient classification				
	AUC	Balanced accuracy	Precision	Recall	F1 score	AUC	Balanced accuracy	Precision	Recall	F1 score
DL	0.70 [0.66,0.75]	0.67 [0.63,0.71]	0.06 [0.05,0.08]	0.56 [0.48,0.64]	0.11 [0.09,0.14]	0.58 [0.53,0.64]	0.58 [0.54,0.62]	0.11 [0.09,0.13]	0.73 [0.65,0.81]	0.19 [0.16,0.23]
Rad	0.89 [0.87,0.91]	0.81 [0.78,0.84]	0.09 [0.08,0.11]	0.86 [0.80,0.01]	0.17 [0.14,0.19]	0.76 [0.72,0.80]	0.71 [0.67,0.76]	0.22 [0.18,0.26]	0.65 [0.55,0.74]	0.33 [0.27,0.38]
Rad + DL	0.92 [0.91,0.93]	0.85 [0.83,0.87]	0.10 [0.09,0.12]	0.93 [0.88,0.96]	0.18 [0.16,0.21]	0.81 [0.78,0.84]	0.75 [0.71,0.78]	0.19 [0.15,0.22]	0.84 [0.77,0.91]	0.31 [0.26,0.35]
Rad + Clin	0.88 [0.86,0.90]	0.81 [0.78,0.84]	0.09 [0.08,0.10]	0.86 [0.80,0.91]	0.16 [0.14,0.19]	0.78 [0.73,0.82]	0.70 [0.66,0.74]	0.18 [0.14,0.21]	0.73 [0.64,0.81]	0.0.29 [0.24,0.33]
Rad + DL + Clin	0.88 [0.86,0.90]	0.82 [0.79,0.85]	0.10 [0.08,0.11]	0.85 [0.79,0.90]	0.17 [0.15,0.20]	0.77 [0.73,0.81]	0.70 [0.66,0.73]	0.15 [0.12,0.18]	0.88 [0.82,0.94]	0.25 [0.21,0.29]
Agreed labels	0.88 [0.85,0.90]	0.82 [0.77,0.85]	0.09 [0.07,0.11]	0.81 [0.73,0.88]	0.16 [0.13,0.19]	0.74 [0.69,0.78]	0.60 [0.58,0.62]	0.13 [0.11,0.16]	0.97 [0.93,1.00]	0.23 [0.19,0.27]

Table 4. AUC, balanced accuracy, precision, recall, and F1 metrics with CI on the internal validation on patient and lesion levels.

Per-lesion classification						Per-patient classification					
Approaches	AUC	Balanced accuracy	Precision	Recall	F1 score	Approaches	AUC	Balanced accuracy	Precision	Recall	F1 score
DL	0.72 [0.66,0.78]	0.61 [0.55,0.67]	0.07 [0.04,0.09]	0.37 [0.26,0.49]	0.11 [0.07,0.16]	DL	0.63 [0.55,0.71]	0.59 [0.52,0.66]	0.12 [0.09,0.17]	0.63 [0.50,0.77]	0.21 [0.15,0.27]
Rad	0.76 [0.69,0.81]	0.70 [0.64,0.76]	0.07 [0.05,0.09]	0.67 [0.55,0.78]	0.13 [0.10,0.16]	Rad	0.76 [0.70,0.81]	0.70 [0.65,0.76]	0.07 [0.05,0.09]	0.67 [0.56,0.78]	0.13 [0.10,0.16]
Rad + DL	0.71 [0.66,0.76]	0.64 [0.58,0.70]	0.06 [0.04,0.08]	0.53 [0.41,0.64]	0.11 [0.08,0.14]	Rad + DL	0.55 [0.47,0.63]	0.51 [0.44,0.58]	0.10 [0.06,0.14]	0.84 [0.77,0.91]	0.31 [0.26,0.35]
Rad + Clin	0.77 [0.71,0.82]	0.71 [0.65,0.76]	0.07 [0.05,0.09]	0.69 [0.58,0.79]	0.13 [0.10,0.16]	Rad + Clin	0.64 [0.55,0.72]	0.60 [0.52,0.67]	0.13 [0.09,0.18]	0.55 [0.41,0.69]	0.22 [0.14,0.28]
Rad + DL + Clin	0.71 [0.65,0.76]	0.63 [0.57,0.69]	0.06 [0.04,0.08]	0.53 [0.41,0.64]	0.10 [0.07,0.13]	Rad + DL + Clin	0.59 [0.51,0.67]	0.65 [0.49,0.63]	0.11 [0.07,0.15]	0.65 [0.51,0.78]	0.19 [0.13,0.25]
Agreed labels	0.81 [0.73,0.89]	0.73 [0.64,0.81]	0.10 [0.06,0.15]	0.55 [0.38,0.71]	0.17 [0.11,0.24]	Agreed labels	0.71 [0.62,0.79]	0.61 [0.59,0.63]	0.11 [0.08,0.15]	1.00 [1.00,1.00]	0.20 [0.14,0.26]

CHAPTER 7

7

Acknowledgements

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 Hypoximmuno), ERC-2020-PoC: 957565-AUTO.DISTINCT, Authors also acknowledge financial support from EUROSTARS (DART, DECIDE), the European Union's Horizon 2020 research and innovation programme under grant agreement: ImmunoSABR n° 733008, MSCA-ITN-PREDICT n° 766276, CHAIMELEON n° 952172, EuCanImage n° 952103, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY n° UM 2017-8295) and Interreg V-A Euregio Meuse-Rhine (EURADIOMICS n° EMR4). This work was supported by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018–2

Keywords

generative modelling, PixelCNN, medical imaging, interpolation, CT

Towards texture accurate slice interpolation of medical images using PixelMiner

W. Rogers^{1*}, S. A. Keek¹, M. Beuque¹, E. Lavrova^{1,2}, S. Primakov¹, G. Wu¹, C. Yan¹, S. Sanduleanu¹, H. A. Gietema³, R. Casale^{1,4}, M. Occhipinti⁵, H. C. Woodruff^{1,3}, A. Jochems¹, P. Lambin^{1,3}

1 The D-Lab, Department of Precision Medicine, GROW – School for Oncology and Developmental Biology, Maastricht University, Maastricht, The Netherlands

2 GIGA Cyclotron Research Centre in vivo imaging, University of Liège, Liège, Belgium

3 Department of Radiology and Nuclear Medicine, GROW – School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands

4 Department of Radiology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium

5 Radiomics, Clos Chanmurlu 13, 4000 Liege, Belgium

Author contributions statements

All research was performed by W. Rogers. The main manuscript text was written by W. Rogers and S. Keek. All figures were created by W. Rogers. Source code and model architecture was created by W. Rogers, with additional source code provided by S. Primakov. All authors reviewed the manuscript and participated in a qualitative assessment trial. The project was supervised by P. Lambin and H.C. Woodruff.

7.1 Abstract

Quantitative image analysis models are used for medical imaging tasks such as registration, classification, object detection, and segmentation. For these models to be capable of making accurate predictions, they need valid and precise information. We propose PixelMiner, a convolution-based deep-learning model for interpolating computed tomography (CT) imaging slices.

PixelMiner was designed to produce texture-accurate slice interpolations by trading off pixel accuracy for texture accuracy. PixelMiner was trained on a dataset of 7,829 CT scans and validated using an external dataset. We demonstrated the model's effectiveness by using the structural similarity index (SSIM), peak signal to noise ratio (PSNR), and the root mean squared error (RMSE) of extracted texture features. Additionally, we developed and used a new metric, the mean squared mapped feature error (MSMFE). The performance of PixelMiner was compared to four other interpolation methods: (tri-)linear, (tri-)cubic, windowed sinc (WS), and nearest neighbor (NN).

PixelMiner produced texture with a significantly lowest average texture error compared to all other methods with a normalized root mean squared error (NRMSE) of 0.11 ($p < .01$), and the significantly highest reproducibility with a concordance correlation coefficient (CCC) ≥ 0.85 ($p < .01$).

PixelMiner was not only shown to better preserve features but was also validated using an ablation study by removing auto-regression from the model.

7.2 Introduction

Radiological images are an important prognostic and diagnostic tool used by clinicians in clinical decision-making. Radiological imaging is primarily used through qualitative or semi-quantitative assessment by physicians, however, there have been recent advancements in image acquisition, standardization, and analysis that have enabled the discovery of quantitative biomarkers, for example through radiomics [1]. Qualitative image analysis (QIA) provides the means not only to diagnose oncological disorders [2-4], but also to personalize treatment [5] without the use of invasive procedures.

QIA is dependent on the quality of the analyzed images. Image quality is determined by an image's contrast and spatial resolution, as well as the level of noise and the presence of artifacts. Image quality affects the ability to discern different structures and low-frequency signals within an image [6], and low-quality images may therefore have a negative effect on QIA by either the loss of information or altering the true features of an image [7]. Computed Tomography (CT) scans are acquired and reconstructed with a broad scope of imaging-specific parameters which can affect image quality, including differences in dose settings, reconstruction kernel [8], slice thickness and spacing. Different CT reconstruction protocols cause slice thicknesses and spacings to vary considerably between hospitals, creating challenges for QIA in multicenter studies. Differences in pixel size and slice spacing have for example been shown to impact the stability, robustness, and repeatability of radiomic features [9]. Reproducing features robustly contributes to better performing classification models [9] that rely on machine learning to make accurate predictions.

To overcome issues with heterogeneity, CT scans are often interpolated to a common resolution. Commonly used interpolation methods include nearest neighbors (NN), trilinear, and tricubic interpolation [10]. Medical scans are continuous latent images, where there is still uncaptured information between slices that can be predicted. Common interpolation models cause artifacts to appear on the interpolated slices, such as blurring or ghosting [11]. These artifacts have been shown to lead to errors in image registration [12]. Furthermore, during interpolation high-frequency information can be lost [13], and signal reproductions can be too low which leads to distortions [14]. Texture information is represented in a wide range of spatial frequencies which are very important for image classification [15]. Studies such as these suggest it is important to use an algorithm capable of capturing frequency signal information to generate accurate synthetic textures. PixelMiner was designed to probabilistically predict pixel intensities that most likely fit within a signal.

The aim of this research is to improve the interpolation quality of medical imaging to allow for more homogeneous datasets through preprocessing. The proposed method is

a deep learning convolutional autoregressive model that predicts a slice between two adjacent slices from a three-dimensional input including the target slice. The model outputs a single slice prediction of the target and the loss is determined by the difference between the target slice and the output slice. We hypothesize that our approach has major advantages over other methods as it is autoregressive, and as such, it is able to include information about the interpolated slice in its predictions, in contrast to other models that rely solely on surrounding slices. An ablation study was done by removing autoregression from the model to prove its efficacy. We also hypothesize that since textures can be well summarized by their component spatial frequencies [15], a model designed specifically for processing signals will achieve improved performance. PixelMiner was compared to other existing interpolation methods by comparing the sharpness, shape, and accuracy of texture generation using standard metrics and texture feature analysis.

7.2.1. Background

PixelRNN, PixelCNN, PixelCNN++, and PixelSNAIL are instances of a class of deep learning generative models used both for image generation [16, 17, 18, 19] and interpolation [20]. PixelCNN-like models differ from other generative models in that it uses an explicit density function, through an explicit specification of the distribution of the random variable [17]. Most models in machine learning and statistics are in this form [21], whereas generative adversarial networks (GAN) instead use an implicit density function, in which a generator implicitly defines a probability distribution based on a latent vector [22, 23]. PixelCNN can further be distinguished from other generative models in that it uses a tractable density function that optimizes the likelihood of the training data. Variational auto-encoders (VAE) on the other hand define an intractable density, because it makes either variational approximations, Monte Carlo approximations, or both [24]. PixelCNNs also differ greatly from other generative models because they optimize the likelihood of training based on the individual pixels [17]. PixelCNN is an autoregressive model, that predicts pixels one by one and bases the prediction based on all previous predictions [17]. The PixelCNN objective function can be expressed as the product of the probabilities of a pixel based on all previous pixels [17]:

$$p(x) = \prod_{i=1}^{n^2} p(x_i | x_1, \dots, x_{i-1})$$

where x_i represents pixels as pixel intensities, and p is the probability.

7.3 Methods

7.3.1 Data

The dataset used for training at the time of publication was a publicly available dataset from the Radiological Society of North America (RSNA) Pulmonary Embolism Detection Challenge © RSNA 2020 [25]. At training time, the dataset consisted of 7,829 chest CT scans, with pixel spacing and slice thickness of 1mm and 1.25mm, respectively. The dataset's terms and conditions allowed the dataset to be shared or redistributed in any form. There were five institutions contributing data, including AlfredHealth, Koç University Hospital, Center for Artificial Intelligence in Medicine & Imaging (AIMI) at Stanford, Unity Health Toronto, and Universidade Federal de São Paulo. All CT scans were retrospectively collected and anonymized at each institution and approved by the respective institutional review boards in accordance with relevant guidelines and regulations. All participants were provided a statement on written informed consent.

At the time of use, the dataset used for validation was the publicly available dataset from the Early Lung Cancer Action Program (ELCAP) Public Lung Image Database [26]. This database consisted of 50 low-dose CT scans, with a slice thickness of 1.25mm. The dataset's terms and conditions allowed use for non-commercial purposes only, including academic research and education. All CT scans were collected by ELCAP in collaboration with Weill Cornell Medical College and approved by its institutional review board in accordance with relevant guidelines and regulations. All participants provided a statement on written informed consent.

Additionally, the Lung Image Database Consortium image collection (LIDC-IDRI) dataset [27] initiated by the National Cancer Institute (NCI) was used to evaluate the model on the downstream task of nodule detection. The dataset comprises 1018 patients from lung cancer screenings in which the dataset was annotated for lesions in the lung by three independent radiologists. All CT scans were retrospectively collected and anonymized at each institution and approved by the respective institutional review boards in accordance with relevant guidelines and regulations. All participants were provided a statement on written informed consent.

7.3.2 Model

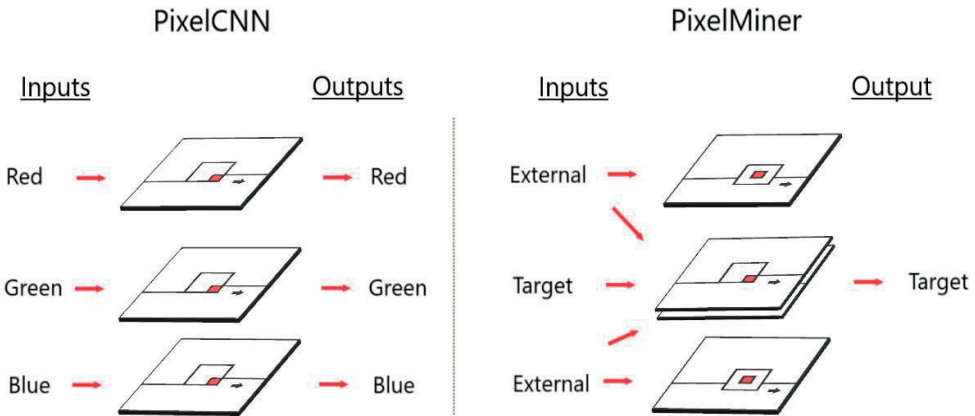


Figure 1. A comparison of how PixelCNN based models and PixelMiner function. PixelCNN has three masked causal convolutions with three inputs and three outputs. In contrast, PixelMiner uses a combination of 3d masked causal convolutions and 2d unmasked non-causal convolutions with three inputs and a single output.

The proposed model, PixelMiner, is a causal autoregressive model based on PixelCNN++ [18] with an altered objective function. PixelMiner includes many of the features of PixelCNN++, including vertically and horizontally stacked convolutions and which are gated for efficient computation and multiplicative units in the form of LSTM gates for long and short term memory. Gated convolutions are defined as:

$$y = \tanh(W_{k,f} * x) \odot \sigma(W_{k,g} * x),$$

where σ is the sigmoid non-linearity, k is the number of layer, \odot is the element-wise product and $*$ is the convolution operator.

Additionally, PixelMiner uses the logistic mixture likelihood loss [18], which takes the continuous univariate distribution as a mixture of logistic distributions to be used to calculate the probability of a given pixel intensity, defined as:

$$P(x|\pi, \mu, s) = \sum_{i=1}^K \pi_i [\sigma((x + 0.5 - \mu_i)/s_i) - \sigma((x - 0.5 - \mu_i)/s_i)]$$

where $\sigma()$ is the logistic sigmoid function. For the edge case 255, replace $x + 0.5$ with ∞ and for the edge case 0 replace $x - 0.5$ with ∞ .

Unlike PixelCNN, PixelMiner is not limited to making predictions based on only the previous pixels. The PixelMiner function can be expressed as the product of the probabilities of a pixel based on all previous pixels for the target slice and all pixels for the external slices:

$$p(t, x, b) = \prod_{i=1}^{n^2} p(x_i | t; x_1, \dots, x_{i-1}; b)$$

where t, x, b are the top, middle, and bottom slices respectively.

The model was trained using batches of three sequential slices as inputs on patches of 64 x 64 pixels to generate a single output slice. The middle input slices were used to evaluate the output slices so that the model learned to replicate the input slice given all the information from three slices. With the resolution of a single slice being 512 x 512 pixels, it is not feasible to use entire slices due to memory limitations. However, because PixelMiernr is an autoregressive model, it is capable of generating slices in patches. Therefore, patches of 64 x 64 were used for training and slice interpolation. Patches are generated by beginning at the top left corner and then expanding out based on the initial patch. Autoregressive models require enough observations to capture a signal to make optimal predictions [27]. With this in mind, PixelMiernr uses patches of 64 x 64 with incomplete 32 x 32 sub-patches. An example of this process is given in figure 2.

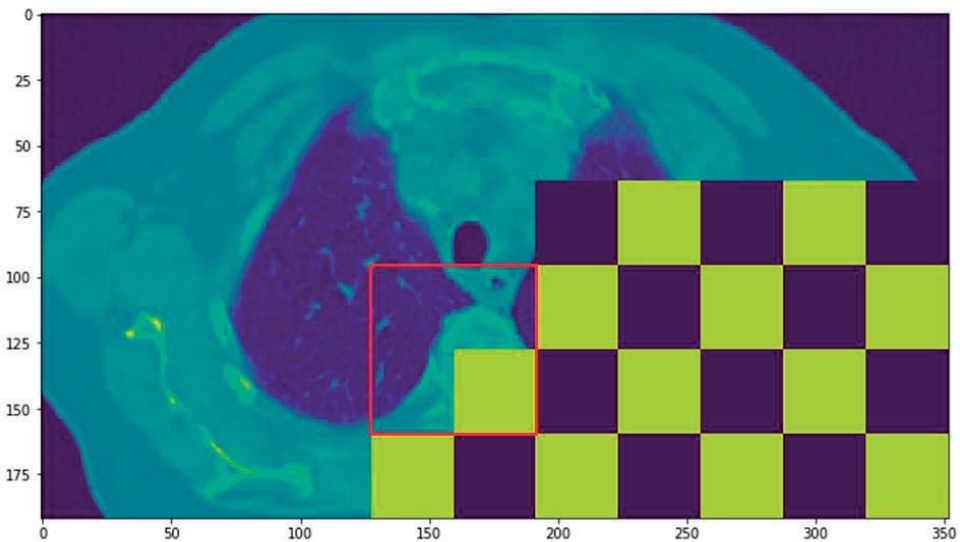


Figure 2. An example of an interpolated slice being generated in patches. The region contained in the red box indicates the next patch of 64 x 64 to be fed into the model. The yellow patch of 32 x 32 in the bottom right of the red box is the sub-patch of pixels that are to be generated next.

PixelCNN based models were designed to generate RGB images and cannot be adapted directly to perform slice interpolation due to stability issues, such as warping and artifacts. To overcome these issues PixelMiernr incorporates 3d causal convolutions and 2d non-

causal convolutions. Figure 1 provides a schematic overview of the differences between pixel predictions performed by PixelCNN and PixelMiner.

PixelMiner also uses a combination of 2D and 3D convolutions, where the 3D convolutions are used to learn the relationships between the slices. This architecture creates a pathway for non-masked non-causal filters in the outer channels using only the external slices, while using masked causal filters in the middle. The causal channel receives information from both the middle and external slices. This allows the outer slices to provide complete information to the middle and prevents the 2D convolutions' tendency to ignore the outer slices. A schematic overview of this convolutional block is provided in figure 3A.

PixelMiner further utilized residual blocks, each of which consists of six downsampling and upsampling layers which utilized long-range skip connections. Feature maps were down-sampled using 2 stride PixelMiner convolutions from 64x64 to 32x32 to 16x16, and then upsampled again using transposed convolutions. The final model was made with the input convolutional layer, n residual blocks, and an output convolutional layer using an identity convolution to combine the final output. A schematic depiction of the architecture can be found in figure 3B.

The model is able to upsample a scan using two slices. An empty slice is then inserted between the top and bottom slices. These three slices are then provided to the model, which generates an interpolated slice pixel by pixel until the entire middle slice is complete. This process is repeated until every slice of the scan is interpolated.

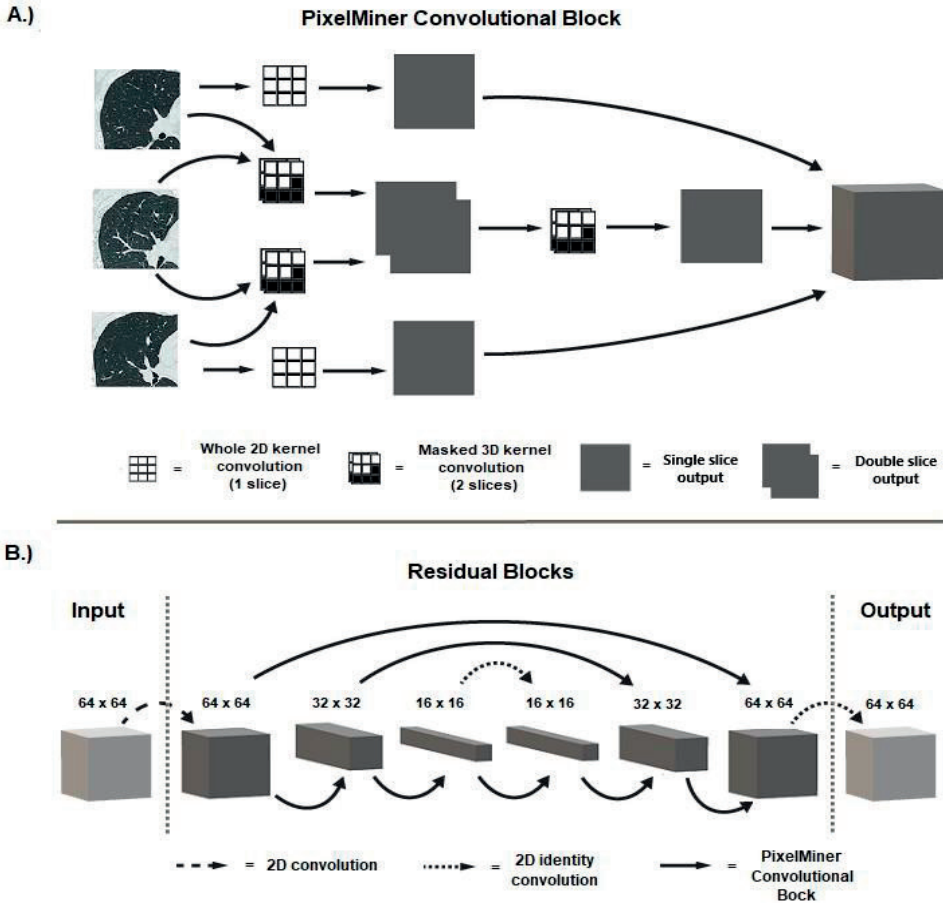


Figure 3. Diagram A.) is a schematic depiction of the PixelMiner convolutional block. Diagram B.) is the entire architecture where the input goes through a 2D convolution and the output goes through an identity convolutional. The architecture is made up of N residual blocks with two down samples using two stride convolutions and two up samples using transposed convolutions.

7.3.3 Training and Validation

Training was performed in an unsupervised fashion with the objective being the reproduction of the middle input slice without the need for labels. The model received three randomly selected slices in sequence and output a single slice, which was then used together with the middle input slice to calculate the loss.

PixelMiner was trained using three GTX 1080ti GPUs using an Adam optimizer with a starting learning rate of 8E-6 and a decay of 0.95. No learning rate scheduler was used and the learning rate was reduced manually over time. Due to hardware limitations, the batch

size was kept to 18 and two residual layers were used with 128 channels. Dropout was used at a rate of 0.25. With regard to the logistic mixture likelihood loss, 10 mixture components were used. All weights were initialized by sampling from a uniform distribution.

To validate the proposed model, the interpolated synthetic slices and ground truth slices were compared. In addition, synthetic slices were generated using four other interpolation methods: (tri-)linear [28, 29], (tri-)cubic [30, 31], window sinc (WS) [32], and NN [33] interpolation. Linear and BSpline (cubic) interpolation are methods commonly used in radiomic studies [34]. WS and NN were done using the open source project the Insight Toolkit or ITK [35] with tools designed specifically for medical imaging.

7.3.4 Evaluation

PixelMiner was evaluated using several evaluation metrics including the full reference IQA metrics peak signal-to-noise ratio (PSNR) and the Structural Similarity Index (SSIM) as well as feature extraction comparisons using gray-level co-occurrence matrix (GLCM), gray-level run length matrix (GLRLM), and gray-level size zone matrix (GLSZM). All analyses were done using 50 generated slices, one randomly selected slice from each of the 50 patients from the ECLAP dataset. In addition, an ablation study was performed to show the efficacy of autoregression over a non-autoregression model.

7.3.4 Peak Signal-to-Noise Ratio

The PSNR is the ratio between the highest power of a signal and the power of the corrupting noise. It is a widely used full reference image quality assessment metric used to determine the fidelity of an image based on the level of noise. PSNR is defined as:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y}_i)^2$$

where Y is the target and \bar{Y} is the prediction.

$$PSNR = 10 \log_{10} \left(\frac{R^2}{MSE} \right)$$

7.3.5 Structural Similarity Index Measure

The SSIM is a full-reference image quality metric for predicting the perceived quality of images. It is a perception-based model designed to consider the degradation of images based on three different properties. The formula itself contains terms for luminance, contrast, and structural information. The terms are defined as:

$$l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}$$

$$s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}$$

where $c_1 = (k_1L)^2$, $c_2 = (k_2L)^2$, and $c_3 = c_2/2$

The combined terms are defined as:

$$SSIM(x, y) = [l(x, y)^\alpha * c(x, y)^\beta * s(x, y)^\gamma]$$

where α , β , γ as parameterized weights.

With all weights set to 1, the formula can be reduced to:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where the SSIM was calculated for each generated slice.

7.3.6 Texture Features

To test the interpolation methods' ability to retain texture information GLCM, GLRLM, and GLSZM texture features were extracted. Feature extraction was performed in Python 3.7 with Pyradiomics 3.0.1 [37]. Features were extracted with pixel intensities at standard hounsfield units with a window of -1024 and 3071, and a bin width of 32. Features extraction was performed for the synthetic and ground truth slices to act as references. Pixel intensities for images were in Masks of regions of interest (ROI) were used for feature extraction to avoid subjectivity. A single image mask was created for the area around the lungs. A simple automated segmentation algorithm was created using thresholding and binary morphology operations to dilate or erode or close openings in the mask using the morphology module in the Scikit-image package [38]. Lungs were segmented using a threshold < -320 Hounsfield units. Texture based measurements features were extracted and compared to synthetic slices and ground truth using the normalized root mean square error (NRMSE) of all GLCM, GLRLM and GLSZM texture features which aggregates all the

errors of the predicted variables into a single metric and normalized using the range of the observed variables, defined as:

$$NRMSE = \frac{\sqrt{MSE}}{\sigma}$$

where σ is the standard deviation.

To further test texture information across patches of the image instead of through a single metric across the whole image, a new metric is proposed: the mean squared mapped feature error (MSMFE). For each texture feature, a feature map was generated using a sliding window with a size of 5x5 and step of 1 calculating texture features for every window. The RMSE is calculated on the reference ground truth feature map and the feature map of the evaluated scan. The MSMFE is then defined as the NRMSE of all texture feature maps.

7.3.7 Ablation

A key component of PixelMiner is auto-regression, which is an ideal method for generating images with regard to slice interpolation. To gauge the efficacy of PixelMiner, we compared the autoregressive model to a model converted to be non-autoregressive, while staying as close to the same architecture as possible. To accomplish this, vertically and horizontally stacked masked convolutions were converted to standard unmasked 2d and 3d convolutions, and gated convolutions were removed. The model was additionally modified to accept 2d inputs and to provide an output of a 1d target prediction. The logistic mixture likelihood loss was designed for predicting pixels autoregressively and was therefore changed to RMSE.

A noticeable problem with the non-autoregressive model is the difficulty of training on patches, whereas autoregression is able to continue a patch from where the last patch left off. Non-autoregression does not have this capability which leads to frame artifacts throughout the image.

7.3.7 Statistical Tests

All evaluation metrics were determined for each generated slice and compared across interpolation methods. Results were statistically tested using a binomial test to test frequency of higher ratios, and a Wilcoxon signed rank test to test pairs of ratios for competing methods on the test dataset.

Texture analysis was accompanied by the concordance correlation coefficient (CCC) which was determined for each feature. The CCC expresses the concordance between paired

ground truth feature values and generated image feature values, which quantifies the agreement and reproducibility, shown as:

$$\mathbf{E}[(\hat{Y} - X)^2] = (\bar{\mu}_y - \mu_x)^2 + \sigma_x^2 + \sigma_y^2 - 2p\sigma_y\sigma_x$$

where μ_y and μ_x are the means and σ_y^2 and σ_x^2 are the variances of the reference and tested slice features, respectively, and p is the Pearson correlation coefficient.

7.4 Results

Figure 5 provides examples of interpolated slices.

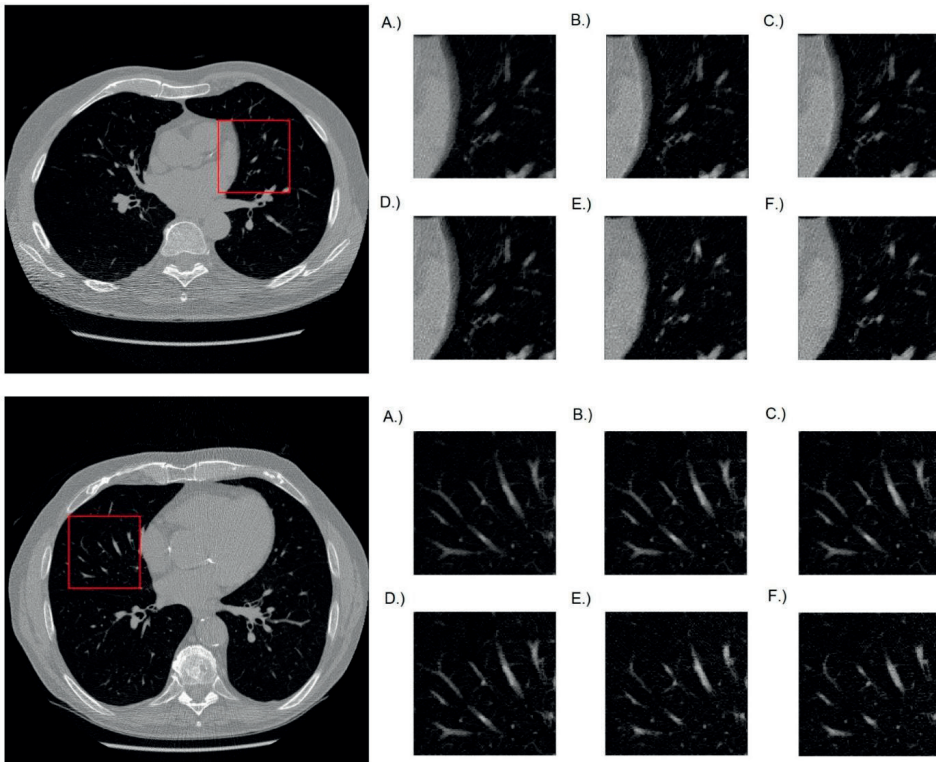


Figure 4. Five interpolation methods that are enlarged and compared side by side for two example slices. Scans to the left are the slices used and the red box indicates the region used for comparison. Methods include A.) linear, B.) windowed sinc, C.) nearest neighbor D.) BSpline and E.) PixelMiner with the corresponding F.) ground truth slice on the bottom right.

7.4.1 Image Quality Assessment Metrics

For completeness we include the common IQA metrics PSNR and SSIM. PixelMiner had a lower PSNR and SSIM value for 100% ($p < 1E-6$) of the scans compared to all other methods, and the lowest mean PSNR and SSIM values. Table 1 provides an overview of the PSNR and SSIM values for each tested interpolation method.

Table 1. Mean and standard deviation of the PSNR and SSIM ($P < 1E-7$).

	PSNR		SSIM	
	Mean	SD	Mean	SD
PixelMiner	31.466	± 3.268	0.908	± 0.048
Windowed Sinc	32.726	± 3.440	0.928	± 0.038
BSpline	32.754	± 3.441	0.928	± 0.038
Nearest Neighbor	32.779	± 3.439	0.929	± 0.037
Linear	33.158	± 3.225	0.933	± 0.034

7.4.2 Segmentation Extracted Texture Feature Outcomes

NRMSE was used as an aggregation of all GLCM, GLRLM and GLSZM texture features. PixelMiner had the lowest NRMSE at 0.109 ($p < 0.01$), and a lower NRMSE value for 62.2% ($p < 0.01$) of the scans in the validation dataset compared to the next best performing interpolation method WS. PixelMiner also had the highest amount of reproducible features (74.5% with $CCC \geq 0.85$), and the highest mean CCC at 89.8. An overview of these metrics for the tested interpolation methods can be found in Table 2.

Table 2. Average RMSE values for the tested interpolation methods, and the percent of slices each interpolation method has lower an RMSE (horizontally) and a higher RMSE (vertically) in comparison to the other methods. Also displayed for each interpolation method are the percentage of features with $CCC \geq .85$, and the mean CCC.

	CCC		NRMSE		Percent of features with a lower error (%)				
	Mean	≥ .85	Mean	SD	Linear	Nearest Neighbor	BSpline	Windowed Sinc	PixelMiner
Linear	68.8	17.6%	0.342	± 0.267	-	18.6	17.2	14.7	13.6
Nearest Neighbor	89.0	64.6%	0.143	± 0.133	81.4	-	52.1	49.1	32.4
BSpline	87.9	64.7%	0.138	± 0.131	82.8	47.9	-	46.0	32.8
Windowed Sinc	86.7	64.7%	0.136	± 0.133	85.3	50.9	54.0	-	37.8
PixelMiner	89.8	74.5%	0.109	± 0.123	86.4	67.6	67.2	62.2	-

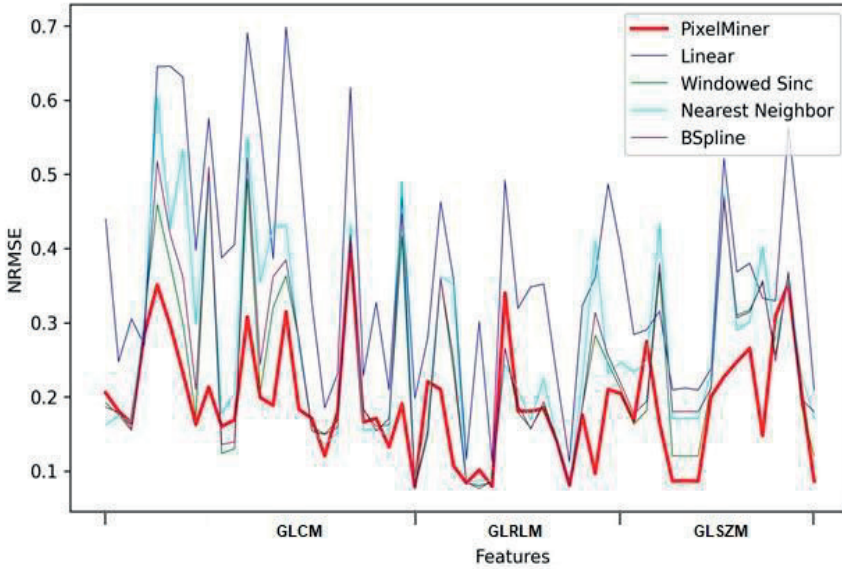


Figure 5. Five interpolation methods that are compared by NRMSE and all GLCM, GLRLM and GLSZM texture features.

7.4.3 Mean Squared Mapped Feature Error

PixelMiner had a lower average MSMFE compared to the other interpolation methods at 0.434 ($p < .01$). For 68.19% of the features, PixelMiner had the lowest MSMFE compared with the next best interpolation method ($p < .01$). Table 3 provides an overview of the MSMFE for each texture feature for the different interpolation methods.

Table 3. The lowest average MSMFE for all GLCM features with $p < .01$. Alongside it is the reported percent of texture features for which the listed interpolation method had a lower MSMFE (horizontally) and higher MSMFE (vertically) in comparison to the other interpolation methods.

	MSMFE		Percent of features with lower average error (%)				
	Mean	SD	Linear	Windowed Sinc	Nearest Neighbor	BSpline	PixelMiner
Linear	0.507	± 0.236	-	36.3	36.5	36.4	34.8
Window Sinc	0.463	± 0.217	63.7	-	44.1	40.7	31.5
Nearest Neighbor	0.462	± 0.217	63.5	55.9	-	38.6	31.2
BSpline	0.462	± 0.217	63.6	59.3	61.4	-	31.0
PixelMiner	0.434	± 0.222	65.2	68.5	68.8	69.0	-

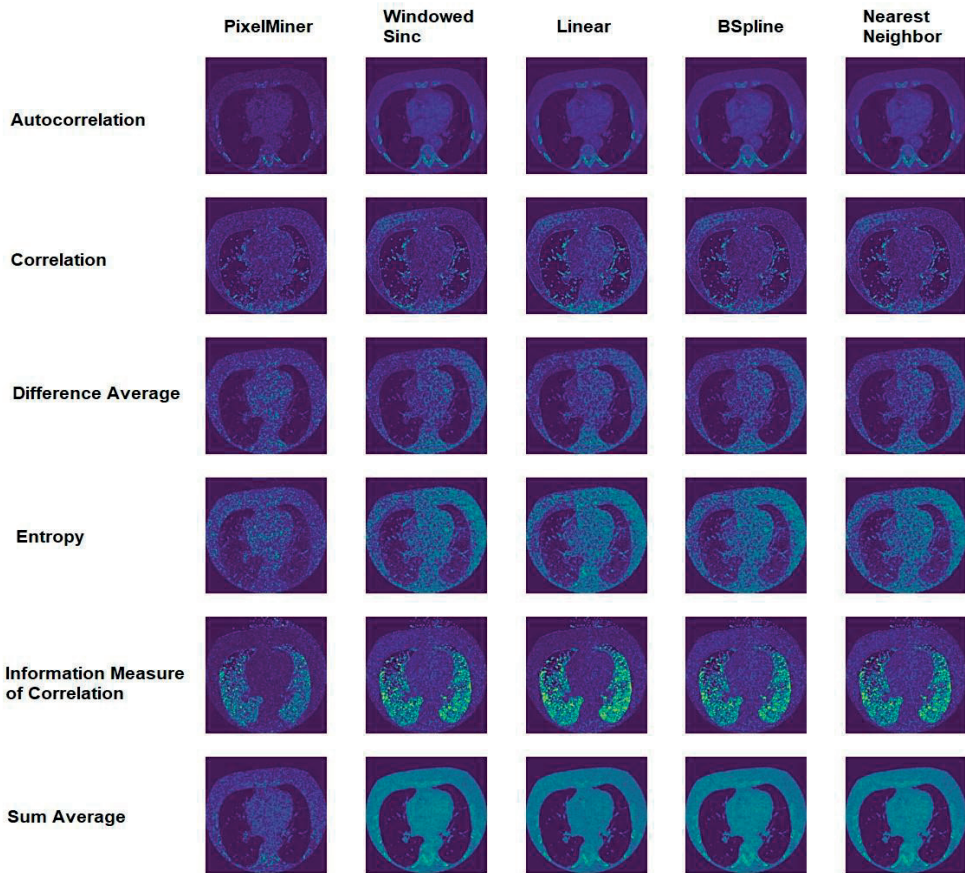


Figure 6. Five interpolation methods and six hand-picked GLCM features with there mapped out error given a matrix based on a window size of five pixels. Lower error is dark and higher error is bright.

7.4.4 Ablation

Two versions of PixelMiner was compared, an autoregressive model and a non-autoregressive model, where the non-autoregressive model was shown to be inferior based on SSIM, PSNR GLCM texture features, and MSMFE. PixelMiner with autoregression had a higher mean PSNR (31.47) and SSIM (0.91). In addition, PixelMiner had a lower GLCM texture error (0.11) and a lower MSMFE (0.434). Table 4 provides an overview of the comparison between autoregression and non-autoregression.

Table 4. Mean and standard deviation of the SSIM, PSNR, GLCM Texture, and MSMFE ($P < 1E-10$).

	SSIM		PSNR		GLCM Texture		MSMFE	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
PixelMiner without autoregression	0.684	± 0.042	20.206	± 2.400	0.504	± 0.2585	0.516	± 0.251
PixelMiner with autoregression	0.908	± 0.048	31.466	± 3.268	0.109	± 0.123	0.434	± 0.222

Non-autoregressively generated images produce noticeably more blurry compared images to autoregressively generated images. Figure 7 shows two examples of an autoregressively and non-autoregressively generated slice compared to the ground truth slice.

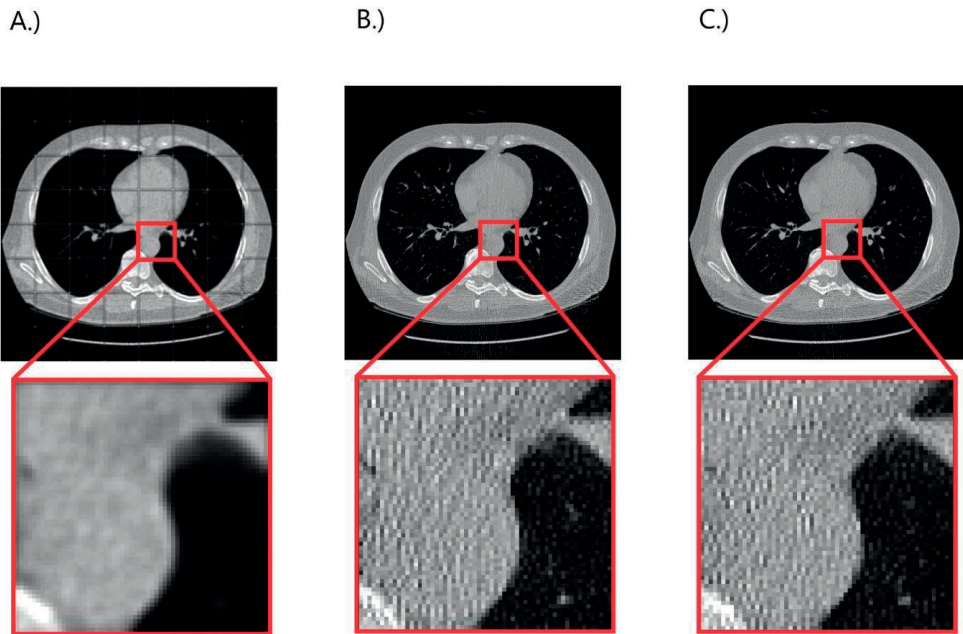


Figure 7. Image A.) is interpolation being performed non-autoregressively, B.) is the ground truth, and C.) is interpolation being performed autoregressively.

7.5 Discussion

The goal of this study was to create a model that can perform slice interpolation with a high degree of texture accuracy compared to other interpolation methods. Because pixel miner makes the tradeoff of texture accuracy over pixel accuracy the result from the SSIM and PSNR does not well reflect real-world performance in texture accuracy. SSIM and PSNR evaluate images by pixel-to-pixel comparisons in relation to mean squared error (MSE). PSNR uses MSE explicitly in its equation, whereas SSIM has been shown to be closely related to PSNR through their shared link to MSE [49]. Because PixelMiner doesn't attempt to generate images with pixel-to-pixel accuracy, the metrics fail to accurately capture the accuracy of PixelMiners ability to generate slices with improved texture accuracy. To overcome this we used texture based feature extraction techniques to better evaluate the model based on texture.

PixelMiner had a significantly lower NRMSE for compared to all other models, the lowest for 65.2% of the features compared to the next best method. Furthermore, the reproducibility of PixelMiner using a CCC ≥ 0.85 was shown to be 80% with a mean of 91.42, greater than all other models. PixelMiner does not outperform other methods for every feature type, so if the features for a particular radiomics signature are known it would be worth doing some research beforehand to determine which interpolation method would complement the model being developed first. If the signature is unknown, or a practitioner is using deep learning, the results indicate PixelMiner to be the best choice for slice interpolation. Though not quantified and reported in results, contributors to this paper were able to observe a pattern of a reduction in the bleeding effect caused by other types of interpolation methods. Averaging type mathematical operations can lead to this effect, whereas PixelMiner is a probabilistic model which makes it able to avoid these types of artifacts. Examples of this can be observed in figure 4.

PixelCNNs have been used to perform super-resolution on 2d images [20], but to the best of our knowledge it has never been used on 3d images and it's the first time it has been used to generate high fidelity images. Deep learning based super resolution has been an active area of research and includes many models such as SRGAN [39], MFTV [40], FSRCNN [41], SRResNet-V54 [42], LapSRN [43], and multiple dense residual block based GANs [44]. These types of models are designed for 3d images by doing grid based upsampling using transposed convolutions. They are unable to be compared directly with PixelMiner since they would require downsampling in 3 dimensions to be evaluated, causing the models to lose valuable information that slice interpolation is able to retain. These super-resolution models could also potentially be used in conjunction with PixelMiner, by upsampling after slices have already been interpolated. We chose to use an auto-regressive model as opposed to a GAN-based model for two reasons. First, we believe that autoregression could force a model to better learn texture information better than other methods.

Additionally, PixelMiner has an explicit density function rather than the implicit density function used in GANs, making it more suitable in medical imaging. GANs have been shown to be unreliable for use in medical imaging by sacrificing accuracy for fidelity [45].

Quantitative analysis in medical imaging relies on high-quality images that are harmonious across many types of scanners. There is a potential for quantitative analysis of medical imaging to have a profound effect on prognosis, but it requires high-quality data for training machine learning models. Medical imaging slice spacing is only one factor in the overall quality of images but it affects many different areas of quantitative analysis, such as image registration, detection, segmentation, and classification.

PixelMiner was assessed using radiomics texture features. Due to some computational constraints of the MSMFE, only GLCMs were used. Future work could allow for the use of MSMFE including additional texture features, such as gray level run length matrix (GLRLM) and gray level size zone matrix (GLSZM).

A disadvantage to using PixelMiner is that it is slower compared to other interpolation methods. Using three 1080ti graphical processing units (GPU), it should be possible to interpolate a full scan with 2mm slice spacing within 24 to 48 hours. Future improvements could be made through better parallelization of GPUs at generation, but also through more efficient models using methods similar to MobileNet [46] or EfficientNet [47]. Furthermore, advancements in graphical processing units continue to grow exponentially, and modern GPUs could considerably bring down generation times. Using the current state-of-the-art GPUs could potentially considerably speed up generation times. Without access to peer reviewed performance on GPUs and deep learning it is not possible to give accurate estimates of generation times but we estimate generation times to be roughly half using an Nvidia RTX 3090 GPU. This additionally makes it difficult to test downstream tasks such as segmentation, detection, and classification. We hypothesize that the model should have a positive effect on segmentation based on qualitative analysis showing fixes to ghosting and bleeding effects and leave this to future work once generation times can be optimized.

It is also well known that convolutional neural networks (CNN) have a dependency on large amounts of data and there is no guarantee that the data a model has been trained on will allow the model to generalize to unseen data. In addition, CNNs are known for being affected long-range dependence, which makes pixels that are farther apart more difficult to predict. This could be partially alleviated by using self-attention which showed modest improvements in the PixelSnail model [48]. Self-attention was omitted from the current version of PixelMiner due to time constraints. With the development of the attention-based model called a transformer [49], it could be possible to build an

improved version of PixelMiner similar to the transformer based image generation model ImageGPT [50]. Transformers perform faster than RNNs but also address the issues with long-range dependencies in PixelCNN and PixelRNN. Transformer based image models are a relatively new development in computer vision models, but have the potential to push the capabilities of PixelMiner much further. We were unable to find any evidence with statistical significance that PixelMiner showed improvements in downstream tasks, though we hypothesize with improvements to PixelMiner it could be possible in the near future.

The nature of PixelMiner makes the results very difficult to assess and further research needs to be done with regard to evaluation. Full reference IQA metrics are of no use for so an algorithm where the goal is to retain texture at the expense of pixel-wise accuracy. Further work is needed to not only better evaluate PixelMiner but also any other types of generative models for images in which texture accuracy may be the main focus.

7.6 Conclusion

PixelMiner provides a new state-of-the-art method for performing slice interpolation on medical imaging capable of improving imaging resolution while better preserving texture features, an important set of features in quantitative medical imaging analysis. PixelMiner performs well not only on the training dataset but also on the validated external dataset.

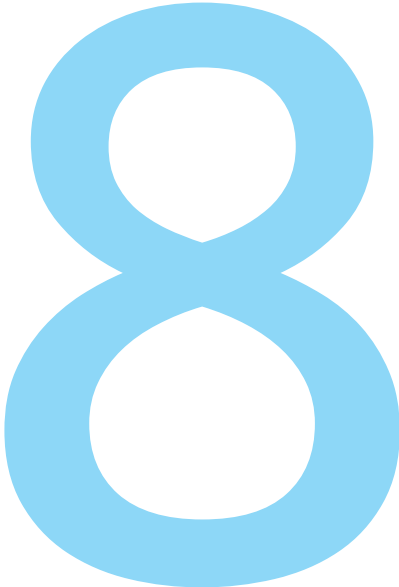
7.7 References

- Aerts, H.J.W.L., et al., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 2014. 5.
- Kurland, B.F., et al., Promise and pitfalls of quantitative imaging in oncology clinical trials. *Magnetic Resonance Imaging*, 2012. 30(9): p. 1301-1312.
- Buckler, A., et al., A Collaborative Enterprise for Multi-Stakeholder Participation in the Advancement of Quantitative Imaging. *Radiology*, 2011. 258: p. 906-14.
- Buckler, A., et al., Quantitative Imaging Test Approval and Biomarker Qualification: Interrelated but Distinct Activities. *Radiology*, 2011. 259: p. 875-84.
- Lambin, P., et al., Predicting outcomes in radiation oncology-multifactorial decision support systems. *Nature reviews. Clinical oncology*, 2012. 10.
- Goldman, L.W., Principles of CT: Radiation Dose and Image Quality. *Journal of Nuclear Medicine Technology*, 2007. 35: p. 213-225.
- Zhao, B., James, L. P., Moskowitz, C. S., Guo, P., Ginsberg, M. S., Lefkowitz, R. A., Qin, Y., Riely, G.J., Kris, M.G., Schwartz, L. H., Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non-Small Cell Lung Cancer 1. *Radiology. Radiological Society of North America (RSNA)*. 2009.
- Choe, J.A.-O., et al., Deep Learning-based Image Conversion of CT Reconstruction Kernels Improves Radiomics Reproducibility for Pulmonary Nodules or Masses. (1527-1315 (Electronic)).
- Mackin, D., et al., Correction: Harmonizing the pixel size in retrospective computed tomography radiomics studies. *PLOS ONE*, 2018. 13: p. e0191597-e0191597. 10. Whybra, P., et al., Assessing radiomic feature robustness to interpolation in 18F-FDG PET imaging. *Scientific Reports*, 2019. 9.
- Leng, J., G. Xu, and Y. Zhang, Medical image interpolation based on multi-resolution registration. *Computers & Mathematics with Applications*, 2013. 66(1): p.1-18.
- Josien, P.W.P., J.B.A. Maintz, and A.V. Max. Mutual information matching and interpolation artifacts. in *Proc.SPIE*. 1999.
- Parker, J.A., R.V. Kenyon, and D.E. Troxel, Comparison of Interpolating Methods for Image Resampling. *IEEE Transactions on Medical Imaging*, 1983. 2(1): p. 31-39.
- Crow, F.C., The aliasing problem in computer-generated shaded images. *Commun. ACM*, 1977. 20(11): p. 799-805.
- Strand, J. and T. Taxt, Local frequency features for texture classification. *Pattern Recognition*, 1994. 27(10): p. 1397-1406.
- van den Oord, A., N. Kalchbrenner, and K. Kavukcuoglu, Pixel Recurrent Neural Networks. *CoRR*, 2016. abs/1601.06759.
- van den Oord, A., et al., Conditional Image Generation with PixelCNN Decoders. *CoRR*, 2016. abs/1606.05328.
- Salimans, T., et al., PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. *International Conference on Learning Representation*, 2017.

- Chen, X., et al., PixelSNAIL: An Improved Autoregressive Generative Model. 2017.
- Dahl, R., M. Norouzi, and J. Shlens, Pixel Recursive Super Resolution. 2017.
- Mohamed, S. and B. Lakshminarayanan, Learning in Implicit Generative Models. 2017.
- Goodfellow, I., et al. Generative Adversarial Nets. in *Advances in Neural Information Processing Systems*. 2014. Curran Associates, Inc.
- Goodfellow, I.J., NIPS 2016 Tutorial: Generative Adversarial Networks. CoRR, 2017. abs/1701.00160.
- Diederik, P.K. and W. Max, An Introduction to Variational Autoencoders. 2019: now. 1.
- Colak, E., et al., The RSNA Pulmonary Embolism CT Dataset. *Radiology. Artificial intelligence*, 2021. 3(2): p. e200254-e200254.
- Henschke CI, M.D., Yankelevitz DF, et al, Early Lung Cancer Action Project. *Lancet*, 1999. 354: p. 1205.
- Lehmann, E.L. and G. Casella, *Theory of point estimation*. 2nd ed. Springer texts in statistics. 1998, New York: Springer. xxvi, 589 p.
- Myers, D.E., Kriging, cokriging, radial basis functions and the role of positive definiteness. *Computers & Mathematics with Applications*, 1992. 24(12): p. 139-148.
- Stytz, M.R. and R.W. Parrott, Using kriging for 3d medical imaging. *Computerized Medical Imaging and Graphics*, 1993. 17(6): p. 421-442.
- Keys, R., Cubic convolution interpolation for digital image processing. *IEEE Trans Acoust Speech Signal Process. Acoustics, Speech and Signal Processing*, IEEE Transactions on, 1982. 29: p. 1153-1160.
- Hou, H. and H. Andrews, Cubic splines for image interpolation and digital filtering. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1978. 26(6): p. 508-517.
- Harris, F.J., On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 1978. 66(1): p. 51-83.
- Lehmann, T.M., C. Gonner, and K. Spitzer, Survey: interpolation methods in medical image processing. *IEEE Transactions on Medical Imaging*, 1999. 18(11): p. 1049-1075.
- Whybra, P., et al., Assessing radiomic feature robustness to interpolation in 18F-FDG PET imaging. *Scientific Reports*, 2019. 9(1): p. 9649.
- Johnson, H.J.a.M., Matthew M and Ibanez, Luis, *The ITK Software Guide Book 1: Introduction and Development Guidelines-Volume 1*. Kitware, Inc., 2015.
- Yu, S., et al. Applications of edge preservation ratio in image processing. in *2014 12th International Conference on Signal Processing (ICSP)*. 2014.
- van Griethuysen, J.J.M., et al., Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer research*, 2017. 77(21): p. e104-e107.
- van der Walt S, S.J., Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, the scikit-image contributors, scikit-image: image processing in Python. *PeerJ*, 2014. 2: p. e453.
- Ledig, C., et al., Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. 2017.
- Jian, X., et al., Image superresolution by midfrequency sparse representation and total variation regularization. *Journal of Electronic Imaging*, 2015. 24(1): p. 1-29.
- Dong, C., C.C. Loy, and X. Tang. Accelerating the Super-Resolution

- Convolutional Neural Network. in *Computer Vision – ECCV 2016*. 2016. Cham: Springer International Publishing.
- Rajan, D. and S. Chaudhuri, Generalized Interpolation for Super-Resolution. 2006. p. 45-72.
- Lai, W., et al. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
- Zhang, X., et al., CT super-resolution using multiple dense residual block based GAN. *Signal, Image and Video Processing*, 2021. 15(4): p. 725-733.
- Youssef, S., Pierre-Marc, J., Alain, L. 2021, GANs for Medical Image Synthesis: An Empirical Study, *ArXiv*, 2021. abs/2105.05318.
- Howard, A.G., et al., MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *ArXiv*, 2017. abs/1704.04861.
- Tan, M. and Q. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 2019.
- Vaswani, A., et al., Attention Is All You Need. 2017.
- Chen, M., et al., Generative Pretraining from Pixels. 2020.
- Horé, A. and D. Ziou. Image Quality Metrics: PSNR vs. SSIM. in *2010 20th International Conference on Pattern Recognition*. 2010.

CHAPTER 8



Discussion and future perspectives



Radiomics is an exponentially growing field of research, with an increase of almost 2247% in yearly articles published on PubMed in 2021 compared to 2016 (2183 [2021] versus 93 [2016]). Figure 1 shows a diagram of the increase of radiomics studies over the past years (PubMed, accessed 19th of April 2022). Deep learning (DL) for predictive and classification tasks in the medical field is also rapidly expanding, while DL based segmentation has become an established method for automatic delineation [1-3]. For radiomics, there have been a large amount of proof-of-concept, model development, and validation studies on different tumor histology's and disease sites, using numerous different methodologies and different regions of interest (ROI) as input. In this thesis, we investigated the potential of (deep) radiomics to identify tumor phenotypical subtypes that can subsequently be used for predictive and prognostic purposes for three different tumor subtypes: (locally) advanced head and neck squamous cell carcinoma (HNSCC), locally advanced non-small cell lung cancer (NSCLC), and brain metastases (BM) from any primary origin.

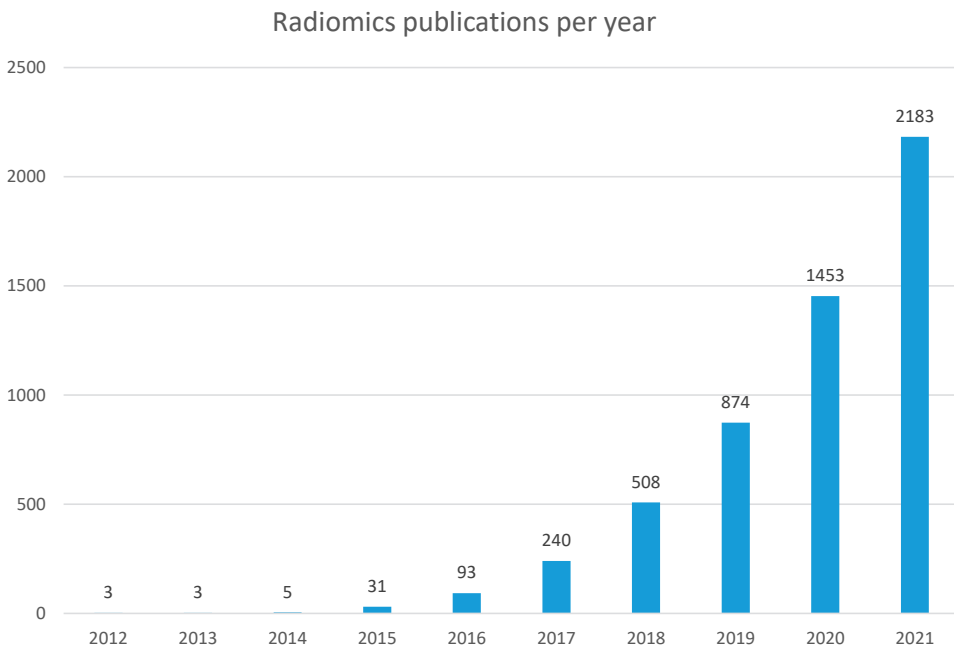


Figure 1. Radiomics publications per year.

The tumor, nodes, and metastases (TNM) staging system (American Joint Committee on Cancer (AJCC) TNM 8th edition for HNSCC, and the International Association for the Study of Lung Cancer (IASLC) TNM 8th edition for NSCLC) is currently used to classify patients according to different prognosis, and used for treatment decision-making [4, 5]. However, even within a certain disease stage or with a specific treatment, the disease course for a specific patient can vary significantly, with some patients cured and some not [6, 7]. Even

though newer editions consider for certain disease types genomic/mutational statuses to further refine the TNM staging system [8], the limited number of stages incorporated by the TNM-staging system does not accurately represent the huge variation in patients and tumor subtypes, resulting in different treatment outcomes (cure / no cure, toxicity / no toxicity). Predictive models including more information on the tumor through quantitative imaging methods such as radiomics could potentially identify these subtypes.

Our goal in this thesis was therefore to attempt to prove the following hypothesis: Quantitative information from tumor regions of advanced HNSCC, NSCLC, and BM on medical imaging acquired prior to treatment is predictive for survival, tumor recurrence, and toxicity related outcomes. The basis of this thesis is an overview of the current role of radiomics in a clinical setting and its future therein (Chapter two). Next, for several types of cancers, all at a (locally) advanced stage, we tested the possibility of using radiomics to aid in prediction and prognosis, which may help clinical decision-making. We first attempted to improve the prediction of overall survival (OS) by extracting radiomics features from the primary HNSCC tumor on baseline computed tomography (CT) images. We compared and combined this with the current gold standard of the AJCC TNM 8th edition, and known clinical (age, sex, smoking/alcohol status) and biological (HPV, hemoglobin level) predictors (Chapter three). We found that radiomics has complementary value in predicting OS, and was able to identify three significantly different survival risk-groups. Furthermore, to test if the peritumoral tissues surrounding primary HNSCC tumors contain predictive information on OS, distant metastasis (DM), and locoregional failure (LRF), we trained a radiomics signature on baseline CT using expanding rings around the gross tumor volume (GTV) (Chapter four). However, no significant predictive value of radiomics was found. For the next tumor type, stage III NSCLC, we compared and combined a radiomics signature with known predictors (age, adenocarcinoma histology, smoking status, the type of chemotherapy administered, and WHO performance status) for the development of BM on baseline contrast-enhanced computed tomography (CECT) (Chapter five). We found that, while predictive, radiomics did not outperform, nor did it complement, a ML model built on simple clinical predictors of BM development. Furthermore, for BM, we tested and compared the feasibility of handcrafted and deep radiomics to predict adverse radiation effects (ARE) after stereotactic radiotherapy (SRT) on baseline T1-weighted magnetic resonance images (MRI) (Chapter six), and found that a combination of handcrafted and deep radiomics is significantly predictive. Lastly, we explored the possibility of improving data quality by interpolating chest CT images using DL focusing on texture accuracy to predict slices (Chapter seven).

In the following sections, the challenges radiomics faces as a field and the future perspectives are discussed. First, we discuss generalizability, reproducibility, and the need

for harmonization. Next, we discuss the need for high data quantity and quality, and finally we discuss future areas for research needed for radiomics to advance as a scientific field.

8.1 Generalizability, reproducibility, and the need for harmonization

A crucial aspect of predictive modelling is the feasibility of a signature to be applied to other patient populations. For radiomics based signatures, this means models can be applied to imaging data using different settings, collected from different centers, machines, and countries. This is important, as the gold standards of most clinical predictive models are based on clinical, biological, and genomic variables that can be applied to many different populations, depending on the complexity of the methods to retrieve the necessary variables. However, for the different TNM staging systems, which are constantly updated with new guidelines, discrepancies in survival between hospitals also exist [9-13]. Nonetheless, we want to ensure that the models created in this thesis can be applied as widely on different patient populations similarly to TNM staging and other clinical predictors.

In Chapter 5, we aimed to develop a predictive model of the risk of BM development in radically treated stage III NSCLC patients using handcrafted radiomics on CECT images. Patients with NSCLC, especially with adenocarcinoma subtype, have a high risk of BM [14], which drastically lowers quality of life (QoL), while curative treatment is seldom possible. Prophylactic cranial irradiation (PCI) has shown to reduce the risk of BM by 0.33 [15], and increase the PFS in stage III NSCLC, without a significant improvement in OS. However, the treatment carries a 25-27% risk of neurocognitive impairment (mainly grade 1-2), which can influence quality of life [16, 17]. Therefore, being able to determine which patients are higher at risk of BM before treatment is important, as this may prevent administering PCI for patients who will most likely do not obtain benefit. Furthermore, a more intense follow-up with brain MRI could be used for those at high risk. We trained models on data from two different centers, and externally validated these on data from two other centers. However, when comparing the model trained on radiomics features (AUC of 0.62) to a model (AUC of 0.71) trained on known clinical predictors (lower age and adenocarcinoma histology), we found radiomics unable to outperform, or provide complementary value, to clinical variables.

One explanation for the low performance of radiomics is the large variability of acquisition and reconstruction protocols, and the differences in machines in general, which have all been proven to induce differences in feature values that negatively affect feature reproducibility and generalizability [18, 19]. This is in contrast to models trained using clinical and biological models, which in general have simpler and similar methods with set clinical protocol to obtain the variable. Genomics is an exception to this and is susceptible

to the testing setup used to acquire the gene signatures [20], which can affect the value of the genomic variable extracted, similarly to quantitative imaging. For genomic variables, a harmonization method called Combine Batches (ComBat) was introduced that can remove these differences, called batch effects, through an empirical Bayes framework [21]. This method is designed to determine and adjust for a single batch affecting the values, while keeping effects induced by biology (biological covariates) intact. This allows genomic signatures between different populations acquired with different testing setups to be compared.

Implementing a similar harmonization method for radiomics could standardize features affected by scanner effects (i.e., first order and texture features), allowing radiomics models to be used between different patient populations. Orlhac et al (2018) implemented ComBat to compare reproducibility of data from one center to two other centers, using the center from which the images were obtained as batch effect, for nine different radiomics features extracted from fluorodeoxyglucose ^{18}F (^{18}F -FDG) positron emission tomography (PET) images of breast cancer patients [22]. Before harmonization, four or six features out of nine, depending on the center used to compare reproducibility, showed significant differences in distributions, while after ComBat harmonization zero features showed significant differences. These results show ComBat allows for an increased reproducibility between centers.

However, ComBat as a method was designed for gene-expression array harmonization, where the type of array was the sole batch effect to remove while considering all the biological covariates. For medical imaging, and subsequently radiomics features, the effects of the acquisition and reconstruction parameters and the differences between brands of scanners are much larger than one. This means a single batch effect, for example center or scanner, is not the correct approach. Furthermore, the equation of ComBat requires the list of possible biological covariates to be provided that may have an effect on radiomics values. When these are not provided, ComBat cannot by itself differentiate between the effects the batch or the biological covariates has on the feature [23]. To account for this, estimating the batch effects on phantom data for the scanners and settings that will be corrected prevents biological effects on influencing the effect, which in turn allows this effect to be applied to radiomics features [24].

Although the clinical data used for the study in Chapter 5 study was collected prospectively (one phase III randomized clinical trial and a prospective series), the primary aim of both studies was not to develop a radiomics signature. Therefore, a phantom study was not performed and there were no strict guidelines to use a certain type of CT scanner with a specific scanning protocol, and settings for reconstruction and acquisition were not pre-specified. The possible application of ComBat as a result was limited, and was not tried

extensively. Nonetheless, the remaining studies in this thesis did include limited attempts at controlling the acquisition and reconstruction parameters, such as slice spacing and pixel spacing. In addition, other preprocessing methods have shown to be able to address some of the differences between datasets [25, 26].

In Chapter 6, we aimed to develop a predictive model for ARE of patients with BM using deep and handcrafted radiomics on pre-treatment MRIs. Patients with BM are often not eligible for surgery as the number of BM and the average age of the patient makes surgery unfeasible. As an alternative, SRT can accurately treat patients with multiple BM [27]. However, this treatment carries the risk of late onset ARE which are difficult to discern from tumour progression and require a trained neuroradiologist and often advanced imaging techniques to diagnose. Predicting which patients have a higher chance of ARE before treatment can be beneficial, as it might change the choice of treatment or identify patients who will require more intense follow-up.

Both handcrafted radiomics and DL approaches on MRI data were used separately to predict ARE for patients with BM treated with SRT, and in combination with each other and with a list of known clinical predictors of ARE. External validation was performed on data from a different hospital in a different country. To adjust for the differences in acquisition and reconstruction parameters between the datasets, different pre-processing methods were tried for both methods: z-score normalization [28], contrast limited adaptive histogram equalization (CLAHE) [29], and whitestripe normalization [30]. By performing an internal validation pre-processing method, the ideal one was chosen for both radiomics and DL. The models created on this pre-processed data could significantly predict ARE on an external dataset. This indicates as an alternative to methods such as ComBat, which adjust feature values after extraction of features, methods that instead adjust intensity values could be used to improve model generalizability.

8.2 Data quantity and quality

Another reason for low performance of the model in Chapter 5 could be due to the complexity of radiomics data, which necessitates large datasets. Radiomics, and quantitative imaging in general, is data-hungry, requiring large datasets in both training and validation to achieve high performance, and to verify if the model performance is significant. The numbers of training ($N = 142$) and validation ($N = 77$) samples from the study in Chapter 5 are considered low for quantitative imaging studies, which could have negatively affected the results. However, data curation, collection, and segmentation is a time-consuming process, meaning not all imaging data may be fit for analysis. Instead, using data from routine clinical practice that gets segmented helps with acquiring the necessary data volume. However, as was shown in this Chapter, it can be challenging to collect the necessary number of patients with adequate scans available as only a subset of

scans was eligible. Setting up protocols for hospitals to standardize image acquisition and storage could help future studies with acquiring data, but would require a tremendous amount of coordination between hospitals and countries.

In Chapter 3 a predictive model for OS in stage III-IVB HNSCC patients using radiomics on CT images. The study included patients with oropharynx, larynx, hypopharynx, and oral cavity cancer, with and without HPV-infection, and these patients were treated mostly with radiotherapy. The imaging data was therefore from routine clinical practice with segmentations available. Similarly to the study in Chapter 5, this study contained data from multiple centers, and besides all the data collected having a requirement of a 3mm maximum slice spacing, and a 1mm² maximum pixel spacing, no harmonization or strict guidelines for imaging was set.

However, the number of patients included was relatively much larger, with 666 retrospectively collected patients and 143 prospectively collected patients. The results of this study are therefore in contrast with Chapter 5, as for this study we found radiomics was able to significantly predict OS and stratify the validation cohort in three distinct risk-groups, and the radiomics model contained complementary value to a model based on clinical and biological covariates. This result indicates that radiomics has a place in the clinical decision-making process, adding to the available knowledge clinicians have, not replacing it. The important difference with the model from Chapter 5 is the source of the data, as this allowed a large volume of data that in turn allows the training data to see the necessary variability between populations and within populations to make effective predictive models. To acquire this necessary volume of data, a large-scale, multinational project spanning over 5 years was required, which is indicative of the required effort.

Similarly, the model in Chapter 6, to predict ARE for patients with BM on MRI data, was built on a large volume of data. The outcome of the study was defined in two ways: ARE per lesion after treatment, and ARE for any lesion a patient has. Outcomes of both predictions would allow clinicians to use this information for treatment decision-making, opting for example for systemic therapy, and would allow for better informing of the patient of the risk of SRT. The individual lesions were all treated using SRT, and therefore had segmentations readily available needed for treatment planning. This gave a large volume of over 6000 lesions to train our model on. Radiomics was found to be predictive of ARE. In fact, a combination of radiomics features and DL features, which are features extracted from the last layer of the DL model after training it to predict ARE, was found to be most predictive. This indicates radiomics for this task can outperform patient and lesion characteristics. However, when predicting on a patient-level, including patient and lesion characteristics resulted in the highest performance. Moreover, the ML models in general outperformed the DL models for patient-level and lesion-level.

Data quality is another important factor, with especially slice thickness being reported as having a large impact on model performance [31, 32]. Acquiring images with small slice thickness (0.5-1.5mm), and therefore acquiring more images per scan, is preferable, but can be unfeasible due to extra time needed to capture these images, and because of the amount of storage which would be required for images with these resolutions. While data storage will increase over time as technology progresses, large amounts of data are available now with lower resolution that could still be a valuable source of information. Therefore, in Chapter 7 we developed a tool to interpolate these images, which increases the number of slices in a scan to increase the data quality. The DL model was trained to produce images that are texture-accurate, to preserve information that might be present in the images [33]. The model was able to create images that with a lower average texture error compared to other common interpolation methods. This indicates the model could potentially be used to lower slices spacing, and therefore increasing image quality, and to harmonize the slice spacing in a dataset using an interpolation method more accurate than conventional techniques. A next step of this study would be to test the method in a study where a predictive model is made using radiomics. Comparing the performances of models trained with and without interpolation would be a good test on the impact the method would have on predictive performance.

Besides quantity of imaging data and the settings used to acquire the images, another important aspect of data quality is the number of events available in the entire population. Many ML algorithms without human intervention wrongly assume that the number of events and controls in a binary model are balanced [34, 35]. This may result the model training to converge in such a way that a large amount of predictions will be made for the majority class, as this may lead to a high performance. Chapter 6 involves data with an extremely imbalanced outcome, as only 2-3% of the samples had an event. However, methods such as undersampling of the majority class, and adjusting the weights of the outcomes before training, were able to successfully adjust for this imbalance.

Radiomics as a field has moved on from being a novel technique with inflated expectations to a field that, while having shown successes in many different prediction and classification tasks, has encountered many pitfalls along the way. Figure 2 shows a curve of the typical expectations of any new field or discovery over time. Radiomics, after a peak of inflated expectations, has to overcome many problems, such as reproducibility, generalizability, explainability, and the lack of access to the data used for many radiomics studies to allow verification of the created models. However, as the increasing number of radiomics studies show, and the increasing amount of efforts trying to address some of these issues, radiomics is at the forefront of a phase of “enlightenment”, which could push the field to be clinically viable in the future. This is represented by the arrow in Figure 2.

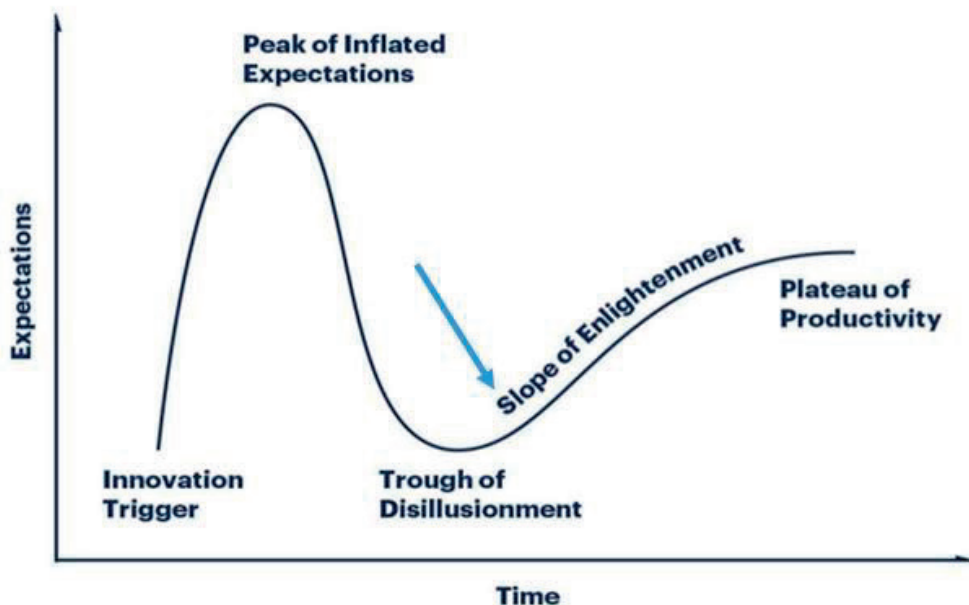


Figure 2. Graph explaining scientific fields' expectations over time. The blue arrow points to the approximate location of radiomics in this time graph.

As previously mentioned, for many quantitative imaging studies the collection and curation of data is one of the most time-consuming processes, especially if the data needs to be segmented. What is still lacking in most studies is the ability to reproduce the results by having access to the data the studies used. Using the data, and the created models, to validate other models and to recalibrate models using new data could further advance the field, especially as new ML and DL methods are continuously being developed. Having data available now and data collected in the future set up according to a set of guidelines, such as the Findable, Accessibility, Interoperability, and reusability (FAIR) principles, would therefore be ideal to enhance model reproducibility [36].

The proposed signatures that have proven predictive value should be further developed to be able to implement clinically. The next needed step therefore is to ensure the proper procedure and testing of implementation in a clinical setting takes place. As previously mentioned, models that use radiomics features are ideally implementable in routine clinical practice, without the need for extra imaging and segmentation. This would allow for models that are broadly applicable, and do not add extra burden to the clinicians. Furthermore, the models need to be implemented in such a way as to complement, not replace doctors. A method could be through digital patient/clinical decision aids, which allow for a digital implementation of the models and direct interaction. Explainability

of the models is also crucial in a clinical settings, which could for example be achieved through the use of SHapley Additive exPlanations (SHAP) modelling [37].

While we have discussed that some of the data collected in these studies come from routine clinical practice, indicating there is room for implementing a radiomics workflow, full integration of radiomics on a technical level, and for a clinician to use the information in a clinical setting would require more evidence. A logical next step to further the field of radiomics is to test radiomics in a clinical trial, either as a separate standalone trial or as inclusion as a secondary objective. Inclusion in these trials would further allow for control over the acquisition and reconstruction protocols required for quantitative imaging, and for the inclusion of phantom studies required for methods such as ComBat. It should be noted, however, that even for clinical trials following guidelines for imaging can be challenging, and should be put under a lot of scrutiny [38]. Nonetheless, inclusion in clinical trials would test the feasibility of including radiomics in a clinical workflow. For predictive studies such as presented in this thesis, *in silico* clinical trials are currently not feasible, but could be a valid replacement of *in vivo* clinical trials. As a last step, acquiring medical device certification, such as the CE marking, should be pursued. Obtaining this marking means a device or technology meets the safety and performance requirements of the European medical device regulations, and is therefore suitable to be used within the European Union.

8.3 Conclusion

The findings of this thesis have shown that quantitative imaging through radiomics and DL, extracted from clinical CT and MR imaging, can be used for a number of different predictive purposes. However, the field faces large limitations, mainly because of harmonization and generalization issues, data quality limits, and a lack of accessibility of data. Efforts to overcome these issues through methods such as ComBat, and the FAIR principles are being investigated. With this thesis, we have introduced some large-scale studies of known unmet clinical needs in **Chapter 3 and 6**, where efforts were made to collect large amounts of data and to analyze these systematically using appropriate guidelines, while the results of **Chapter 4 and 5** show that with smaller datasets radiomics may not be the optimal approach. The results indicate that radiomics has complementary value to currently used methods for prognosis and predictions, and could support clinical decision making in the future. The next step for quantitative imaging analysis to develop as a field would be to set up prospective clinical trials, where the factors inducing the aforementioned limitations can be controlled, and to test the practicality of including (deep) radiomics in a routine clinical setting.

8.4 References

1. Hesamian MH, Jia W, He X, Kennedy P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of Digital Imaging*. 2019;32(4):582-96.
2. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*. 2021;18(2):203-11.
3. Ranjbarzadeh R, Bagherian Kasgari A, Jafarzadeh Ghouschi S, Anari S, Naseri M, Bendechange M. Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. *Sci Rep-Uk*. 2021;11(1):10930.
4. Lydiatt WM, Patel SG, O'Sullivan B, Brandwein MS, Ridge JA, Migliacci JC, et al. Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. *CA Cancer J Clin*. 2017;67(2):122-37.
5. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol*. 2016;11(1):39-51.
6. Lamont EB, Christakis NA. Complexities in prognostication in advanced cancer: "to help them live their lives the way they want to". *JAMA*. 2003;290(1):98-104.
7. Christakis NA, Lamont EB. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *BMJ*. 2000;320(7233):469-72.
8. Zanoni DK, Patel SG, Shah JP. Changes in the 8th Edition of the American Joint Committee on Cancer (AJCC) Staging of Head and Neck Cancer: Rationale and Implications. *Curr Oncol Rep*. 2019;21(6):52.
9. Bostrom PJ, van Rhijn BWG, Fleshner N, Finelli A, Jewett M, Thoms J, et al. Staging and Staging Errors in Bladder Cancer. *European Urology Supplements*. 2010;9(1):2-9.
10. Takaoka MD M, Okuyama A, Mekata E, Masuda M, Otani M, Higashide S, et al. Staging discrepancies between Hospital-Based Cancer Registry and Diagnosis Procedure Combination data. *Japanese Journal of Clinical Oncology*. 2016;46(8):788-91.
11. van Nagell JR, Roddick JW, Lowin DM. The staging of cervical cancer: Inevitable discrepancies between clinical staging and pathologic findings. *American Journal of Obstetrics and Gynecology*. 1971;110(7):973-8.
12. Mercante G, Gaino F, Giannitto C, Ferrelli F, De Virgilio A, Franzese C, et al. Discrepancies between UICC and AJCC TNM classifications for oral cavity tumors in the 8th editions and following versions. *Eur Arch Otorhinolaryngol*. 2022;279(1):527-31.
13. Cienfuegos JA, Salguero J, Nuñez J, Rotellar F, Marti-Cruchaga P, Zozaya G, et al. Discrepancies between two staging systems (European-ENETS versus American-AJCC) of neuroendocrine neoplasms of the pancreas: A study of 77 cases. *Journal of Clinical Oncology*. 2015;33(3_suppl):318-.
14. Nayak L, Lee EQ, Wen PY. Epidemiology of brain metastases. *Curr Oncol Rep*. 2012;14(1):48-54.

15. Witlox WJA, Ramaekers BLT, Zindler JD, Eekers DBP, van Loon JGM, Hendriks LEL, et al. The Prevention of Brain Metastases in Non-Small Cell Lung Cancer by Prophylactic Cranial Irradiation. *Front Oncol.* 2018;8:241.
16. Hendriks LE, Brouns AJ, Amini M, Uyterlinde W, Wijsman R, Bussink J, et al. Development of symptomatic brain metastases after chemoradiotherapy for stage III non-small cell lung cancer: Does the type of chemotherapy regimen matter? *Lung cancer.* 2016;101:68-75.
17. Belderbos JSA, De Ruyscher DKM, De Jaeger K, Koppe F, Lambrecht MLF, Lievens YN, et al. Phase 3 Randomized Trial of Prophylactic Cranial Irradiation With or Without Hippocampus Avoidance in SCLC (NCT01780675). *J Thorac Oncol.* 2021;16(5):840-9.
18. Berenguer R, Pastor-Juan MdR, Canales-Vázquez J, Castro-García M, Villas MV, Mansilla Legorburo F, et al. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radiology.* 2018;288(2):407-15.
19. Kim H, Park CM, Lee M, Park SJ, Song YS, Lee JH, et al. Impact of Reconstruction Algorithms on CT Radiomic Features of Pulmonary Tumors: Analysis of Intra- and Inter-Reader Variability and Inter-Reconstruction Algorithm Variability. *Plos One.* 2016;11(10):e0164924.
20. Kupfer P, Guthke R, Pohlens D, Huber R, Koczan D, Kinne RW. Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis. *BMC Med Genomics.* 2012;5:23.
21. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2006;8(1):118-27.
22. Orhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. *J Nucl Med.* 2018;59(8):1321-8.
23. Ibrahim A, Refaee T, Leijenaar RTH, Primakov S, Hustinx R, Mottaghy FM, et al. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *Plos One.* 2021;16(5):e0251147.
24. Ibrahim A, Primakov S, Beuque M, Woodruff HC, Halilaj I, Wu G, et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods.* 2021;188:20-9.
25. Ibrahim A, Refaee T, Primakov S, Barufaldi B, Acciavatti RJ, Granzier RWY, et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers.* 2021;13(8):1848.
26. Moradmand H, Aghamiri SMR, Ghaderi R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. *Journal of Applied Clinical Medical Physics.* 2020;21(1):179-90.
27. Niranjana A, Lunsford LD. Gamma Knife Radiosurgery for 5 to 10 Brain Metastases: A Good Option for Upfront Treatment. *Oncology (Williston Park).* 2016;30(4):314-5, 7.
28. Reinhold JC, Dewey BE, Carass A, Prince JL. Evaluating the Impact of Intensity Normalization on MR Image Synthesis. *Proc SPIE Int Soc Opt Eng.* 2019;10949.
29. Zuiderveld KJ, editor Contrast Limited Adaptive Histogram Equalization. *Graphics Gems*; 1994.

30. Nyul LG, Udupa JK, Xuan Z. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging*. 2000;19(2):143-50.
31. Park S-H, Lim H, Bae BK, Hahm MH, Chong GO, Jeong SY, et al. Robustness of magnetic resonance radiomic features to pixel size resampling and interpolation in patients with cervical cancer. *Cancer Imaging*. 2021;21(1):19.
32. Escudero Sanchez L, Rundo L, Gill AB, Hoare M, Mendes Serrao E, Sala E. Robustness of radiomic features in CT images with different slice thickness, comparing liver tumour and muscle. *Sci Rep-Uk*. 2021;11(1):8262.
33. Strand J, Taxt T. Local frequency features for texture classification. *Pattern Recognition*. 1994;27(10):1397-406.
34. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. *Intell Data Anal*. 2002;6(5):429-49.
35. Krawczyk B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*. 2016;5(4):221-32.
36. Wilkinson MD, Dumontier M, Aalbersberg JJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3(1):160018.
37. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 4768-77.
38. de Jong EEC, Hendriks LEL, van Elmpst W, Gietema HA, Hofman PAM, De Ruyscher DKM, et al. What you see is (not) what you get: tools for a non-radiologist to evaluate image quality in lung cancer. *Lung cancer*. 2018;123:112-5.

CHAPTER 9

9

Impact statement

Quantitative image analysis through artificial intelligence (AI) has made great leaps in the past years, both through the development of machine learning (ML) methods using semantic and handcrafted (radiomics) features extracted from images, and through direct application of deep learning (DL) algorithms on images. While the development of these models requires a lot of time, coordination, and resources, the final product could be a tool that seamlessly integrates in the current clinical routine without adding workload. Ideally, when a patient enters the clinic and receives a scan, an automatic algorithm would (possibly after automatically delineating any regions of interest) analyze the image and combine this with other patient information in a prediction of any relevant clinical outcome. This is therefore not a tool that would replace a clinician, but rather augment them by condensing the available imaging, clinical, biological, and other information in a single understandable metric. In this thesis, we have contributed to this goal by attempting to solve relevant unmet clinical needs with routine imaging data. Several types of cancers were investigated for the predictive value of radiomics and DL in predicting future events and their complementary value to existing predictors, using large, multi-center datasets to create generalizable models.

All the studies included in this thesis are published in peer-reviewed open-access journals (Cancers, Frontiers in Oncology, Therapeutic advancements in oncology, British journal of radiology, PLOS ONE), and contribute to the field of science, specifically to precision medicine and cancer therapy because of a number of varying reasons. The study in chapter 2 provides an introduction to radiomics and the overview of the current radiomics as a scientific field. The studies in chapter 3 and 4 showed both a positive and negative result in predicting survival and tumor recurrence outcomes for patients with head and neck squamous cell carcinoma. This contrast in results highlights the downsides of radiomics, among others the need for high quality and high volume data to make effective models, and the existing problems in generalizability and reproducibility, and may serve as guidelines for future research. The study in chapter 5 shows that radiomics, while being able to predict development of brain metastases for patients with non-small cell lung cancer, could not outperform models based on clinical predictors. This emphasizes the relevance of existing data that radiomics does not seek to replace, but rather complement existing predictors built on clinical findings. The study in chapter 6 was a study on a large scale dataset of patients with brain metastasis to predict risk of radiation necrosis. In this study we thoroughly investigated pre-processing of MRI data, and the complementary value DL and radiomics have for prediction studies. Lastly, the study in chapter 7 was a study into slice spacing, which is one of the largest causes of quality discrepancy between images. A DL model to interpolate CT images can potentially address this by increasing the number of slices. In these studies, we have shown the potential of radiomics to be complementary to other clinical predictors, but also the need for future research mainly in the generalizability of the features from different scanners or different imaging protocols.

We further think the next step for radiomics would be the inclusion in multi-center clinical trials, where control over the imaging parameters and inclusion of phantom scans could properly test the feasibility and generalizability of the developed models.

The results presented further have a number of (potential) societal impacts. We believe that radiomics, through the identification of tumor and patient subtypes, can play an important role in the realization of personalized medicine, instead of the current broader staging systems used based on clinical information. The risk stratification models presented would further allow for better informing of patients of their chance of survival or risk of side effects, and would allow for better selection of patients that are eligible for clinical trials. In addition, radiomics can be integrated clinically to perform, in an automated fashion, highly specific tasks done by a clinician now. This would lower clinicians' workloads, would save on time and money by using a machine that could do tasks in a fraction of the time a human could, and reduce the variability between doctors and clinics in performing these tasks. Lastly, radiomics would not replace clinicians by doing these tasks, but instead augment them, transforming the current clinician in an "AI-enhanced" clinician which would be better equipped to face the increasing workload in hospitals.

CHAPTER 10

10

Summary

Most patients with head and neck squamous cell carcinoma (HNSCC) and non-small cell lung cancer (NSCLC) show no early symptoms, and are therefore commonly diagnosed at an advanced stage of the cancer, which drastically lowers the chance of survival and the quality of life (QoL). Similarly, symptomatic brain metastases (BM), which are tumors that have spread from a different part of the body to the brain, are also associated with a poor prognosis and drastically decreased QoL and poor prognosis. Advanced stages of HNSCC, NSCLC, and BM usually have no curative treatment options available.

Current risk-stratification of patients with these types of cancer happens through the tumor-nodes-metastasis (TNM) staging system. This system describes the primary tumor location, size, and invasiveness (T-status), the presence of and extent the cancer has spread to local lymph nodes (N-status), and finally the presence of cancer metastases to distant parts of the body (M-status). This combination of T-, N-, and M-status results in an overall cancer staging, ranging from I to IVA/B. In addition to size and location, the genetic and mutational status of the tumor are commonly used for deciding treatment, and sometimes also for staging. Furthermore, information regarding patient characteristics and preferences is also taken into consideration by determining patient performance score, comorbidities, the expected QoL, and risk of complications for the treatment choices available. Lastly, patient specific clinical and biological factors are regarded, such as smoking and alcohol consumption, hemoglobin level, sex, and age.

Advanced stages of NSCLC and HNSCC, as well as BM, have several unmet clinical needs. For advanced (stage III-IVB) HNSCC patients stratification in overall survival (OS) risk-groups, even with the implementation of the 8th edition staging, remains difficult. For stage III NSCLC, even though effective treatments exist to lower the risk of BM, due to the possible side-effects of these treatments determining which patients are at a high risk of BM is needed, which is currently clinically not feasible. Last, for BM determining which patients are at risk of adverse radiation effects (ARE) such as radiation necrosis (RN) before delivering stereotactic radiotherapy (SRT) is important, as this information may be used for risk stratification, for informing the patient, or to opt for different a treatment.

What these patients have in common is the use of medical imaging, either for diagnosis, staging, or for treatment planning purpose. Common medical imaging modalities include positron emission tomography (PET), computed tomography (CT), and magnetic resonance imaging (MRI). The imaging modality used depends mainly on the location of the tumor and the aim of the procedure (finding distant metastases, providing local details etc.). Quantitative image analysis through radiomics and deep learning of these medical images may allow for the identification of phenotypical subtypes of tumors that could be investigated for their correlation to certain clinical outcomes, and subsequently improve prognosis.

In this thesis, we investigated this through a number of studies involving radiomics and deep learning. We first attempted to improve the prediction of OS by extracting radiomics features from advanced HNSCC tumors on baseline computed tomography (CT) images (chapter 3). We compared and combined this with the current gold standard of the American Joint Committee on Cancer (AJCC) TNM 8th edition, and known clinical (age, sex, smoking/alcohol status) and biological (HPV, hemoglobin level) predictors of OS. We found that radiomics has complementary value in predicting OS, and was able to identify three significantly different survival risk-groups. Furthermore, to test if the peritumoral tissues surrounding primary HNSCC tumors contain predictive information on OS, distant metastasis (DM), and locoregional failure (LRF), we trained a radiomics signature on baseline CT using expanding rings around the gross tumor volume (GTV) (chapter 4). However, no significant predictive value of radiomics was found. For the next tumor type, stage III NSCLC, we compared and combined a radiomics signature with known predictors (age, adenocarcinoma histology, smoking status, the type of chemotherapy administered, and WHO performance status) for the development of BM on baseline contrast-enhanced computed tomography (CECT) (chapter 5). We found that, while predictive, radiomics did not outperform or complement a ML model built on simple clinical predictors of BM development. Lastly, for BM, we tested and compared the feasibility of handcrafted and deep radiomics to predict ARE after stereotactic radiotherapy (SRT) on baseline T1-weighted MRI, and found that a combination of handcrafted and deep radiomics is significantly predictive (chapter 6).

We therefore conclude that quantitative imaging through radiomics and DL, extracted from clinical CT and MR imaging, can be used for a number of different predictive purposes. With this thesis, we have introduced some large-scale studies of known unmet clinical for HNSCC and BM, where efforts were made to collect large amounts of data and to analyze these systematically using appropriate guidelines, and were able to improve prognosis significantly. In contrast, for studies with smaller datasets radiomics may not be the optimal approach, as the studies for NSCLC and peritumoral HNSCC were not able to produce significant results. These results indicate that radiomics has complementary value to currently used methods for prognosis and predictions, and could support clinical decision making in the future.

CHAPTER 11

11

Samenvatting

Samenvatting

De meeste patiënten met hoofd-hals plaveiselcelcarcinoom (HNSCC) of niet-kleincellig longcarcinoom (NSCLC) presenteren zich niet met vroegtijdige symptomen, waardoor de ziekte vaak in een laat stadium wordt ontdekt. Patiënten met deze typen kanker hebben hierdoor een lage kans op overleving, en een verslechterde kwaliteit van leven (QoL). Daarnaast zijn symptomatische hersenmetastasen (BM), tumoren die naar de hersenen vanaf een andere primaire tumor locatie zijn uitgezaaid, ook geassocieerd met een slechte prognose en een zeer lage QoL. Patiënten met deze typen kanker hebben vaak geen kans op genezing.

Risicostratificatie gebeurt nu klinisch aan de hand van het tumor-lymfeklier-metastase (TNM) classificatiesysteem. Dit systeem duidt aan of en hoe in hoeverre de tumor plaatselijk is uitgebreid (T-stadium), of en in hoeverre de tumor is uitgezaaid naar de lymfeklieren (N-stadium), en of er uitzaaiingen zijn gevonden die zich via het bloed naar de rest van het lichaam hebben verspreid (M-stadium). De combinatie van het T-, N-, en M-stadium resulteert in een TNM-stadiëring van de tumor van stadium I tot stadium IVA/B. Naast de grootte en locatie van de tumor(en), wordt het genetische en mutatie profiel van de tumor ook in overweging genomen bij de keuze van behandeling, en mogelijk tijdens de stadiëring. Daarnaast worden de voorkeuren van de patiënt, en informatie over het functioneren van de patiënt, zoals co-morbiditeit, performance-scores, de verwachte QoL, en het risico op complicaties, meegenomen bij keuze voor behandeling. Tenslotte zijn klinische en biologische factoren zoals leeftijd, geslacht, hemoglobine waarde, en rook- en alcoholconsumptie van belang.

Late stadia van HNSCC en NSCLC, en BM, hebben verscheidene onvervulde klinische behoeften. Voor patiënten met late stadia (III-IVB) HNSCC is stratificatie naar een algemeen overlevingskans (OS) risicogroep, zelfs met de invoering van de 8^{ste} editie van het American Joint Committee on Cancer (AJCC) TNM-stadiëring, moeilijk. Voor stadium III NSCLC patiënten bestaan behandelingsopties die effectief het risico op metastasering naar de hersenen verlagen. Maar omdat deze behandelingen gepaard gaan met mogelijke ernstige bijwerkingen, is een methode die patiënten met een verhoogd risico op metastasering naar de hersenen kan identificeren van groot belang. Zo kan er geselecteerd worden op patiënten met een hoog risico voor metastasering naar de hersenen. Tenslotte is het voor patiënten met bestaande BM moeilijk om te bepalen wie een verhoogd risico heeft op stralingsnecrose (RN) na behandeling met stereotactische radiotherapie (SRT). Patiënten met een verhoogd risico voor ARE zouden met deze informatie kunnen worden geïnformeerd over de mogelijke bijwerkingen van SRT, of worden geadviseerd om een andere behandeling te ondergaan.

Wat deze patiëntgroepen gemeen hebben is het gebruik van medische beeldvorming voor diagnose, stadiëring, of het plannen van de behandeling. Veelgebruikte modaliteiten zijn positronemissietomografie (PET), computertomografie (CT), en magnetic resonance imaging (MRI, ook wel aangeduid als kernspintomografie). De modaliteit die gebruikt wordt hangt af van de locatie van de tumor, en het doel van de beeldvorming (bijvoorbeeld stadiëring, lokaliseren van metastasen, of gebruik voor behandelingsdoeleinden). Kwantitatieve analyse van medische beelden door middel van radiomics en deep learning (DL) zou het mogelijk kunnen maken om fenotypische subtypen van tumoren te kunnen onderscheiden en correleren met bepaalde klinische uitkomsten, en zo de prognose te verbeteren.

In dit proefschrift hebben wij dit onderzocht aan de hand van een aantal studies naar radiomics en DL. We hebben als eerste getracht de voorspelling van OS te verbeteren met radiomics kenmerken van late stadia HNSCC tumoren op pre-behandeling CT beelden (hoofdstuk 3). We vergeleken en combineerden deze kenmerken met de gouden standaard voor risico stratificatie (8ste AJCC editie TNM-stadiëring) en bekende klinische (leeftijd, geslacht, rook en alcohol consumptie) en biologische (humaan papillomavirus infectie, hemoglobine niveau) voorspellers. De uitkomst van deze studie was dat radiomics kenmerken complementaire waarde hebben voor het voorspellen van OS, en drie significant verschillende risicogroepen kunnen onderscheiden. We hebben verder onderzocht of de wefelsels die de tumor direct omringen voorspellende waarde hebben voor OS, locoregionaal tumor falen, en uitzaaiingen van de primaire tumor naar andere gebieden in het lichaam (hoofdstuk 4). We vonden echter dat radiomics kenmerken die geëxtraheerd zijn van deze wefelsels op CT-beelden geen voorspellende waarde hadden. Voor de volgende tumor soort, stadium III NSCLC, werd onderzocht of radiomics kenmerken op CT-beelden, na toediening van een contrastvloeistof, voorspellende waarden hebben voor het risico op BM (hoofdstuk 5). Deze kenmerken werden vergeleken en gecombineerd met bekende risicofactoren (leeftijd, adenocarcinoom histologie, rook status, het type chemotherapie, en Wereldgezondheidsorganisatie (WHO) performance status) voor de ontwikkeling van BM (hoofdstuk 6). We vonden dat, hoewel radiomics kenmerken voorspellende waarde hebben, ze niet een model gebaseerd op simpele klinische voorspellers voor het risico op BM konden overtreffen, en geen toegevoegde waarde hadden voor dit model. Tenslotte hebben we voor patiënten met BM getest of met radiomics kenmerken en DL op T1-gewogen MRI beelden adverse straling effecten (ARE) kunnen voorspellen na behandeling met SRT, en vonden dat een combinatie van radiomics en DL kenmerken significant voorspellend was.

Wij concluderen daarmee dat kwantitatieve beeldvorming op klinische CT en MRI beelden een rol kunnen spelen voor verschillende klinische doeleinden. In dit proefschrift hebben wij aan de hand van enkele grootschalige studies naar onvervulde klinische behoeften

voor HNSCC en BM aangetoond dat de prognose verbeterd zou kunnen worden met behulp van radiomics en DL kenmerken. In tegenstelling daarmee concluderen wij met de onderzoeken naar weefsels rond HNSCC en naar NSCLC dat voor studies met kleinere datasets radiomics een minder ideale methode is voor risico stratificatie. Deze resultaten laten zien dat radiomics complementaire waarde bezit, en klinische besluitvorming in de toekomst zou kunnen ondersteunen.

CHAPTER 12

12

Acknowledgements / Dankwoord



First, I would like to thank my promotor Prof. Dr. Philippe Lambin. Your vision and drive as a scientist, and as a person, are an inspiration. Whenever there was a moment of doubt during a project, you were able to immediately assess the situation and knew what to focus on. I am also very thankful for the great opportunities you have given me to work at other centers across the world, with a half-year exchange in San Francisco as one of the most exciting moments of my career.

To my co-promotor, Dr. Henry C. Woodruff, I would like to thank you for having given me the opportunity to do my PhD in the first place, and for all the support you gave as my supervisor. I will remember all the nice moments outside of work that were spent with the rest of the lab.

I would like to express me sincerest gratitude to my second co-promotor Dr. Lizza Hendriks. While I started working with you at a later stage of my PhD, I feel without your input and drive I would never have been able to be at the stage I am at currently. Thank you for accepting the role of co-promotor, and for all the time you have spent with me discussing projects and helping write this thesis.

Prof. Dr. Vera Schrauwen-Hinderling, Prof. Dr. Bram van Ginneken, Prof. Dr. Lois Holloway, and Dr. Monique Hochstenbag; thank you for being part of my promotion team and taking the time to review my thesis. Dr. Steve Braunstein, Dr. Olivier Morin, Dr. Martin Vallieres, thank you for amazing opportunity to work with you in San Francisco at the UCSF, and for the great time I had there. Dr. Frank Hoebbers and Dr. Frederik Wesseling, thank you as well for the guidance and the good times working on the BD2Decide project.

To all my friends and colleagues in Maastricht: thank you for the wonderful years, and all the support in- and outside of the office. I fell in love with the city during my stay, and I think this is largely due to all the moments I spent there with you. Manon, I made a true friend in Maastricht by meeting you. Thank you for the countless hours spent together at work and especially outside of it. I have greatly enjoyed all the things we experienced during our stay, be they novel ones like diving in Croatia or the many (mostly good) movies we have seen, together with Alex. I hope this time will continue in north, and wish you and Alex a wonderful future. Sebastian, you were one of the first persons I met in Maastricht, and what started as a shared interest in fitness turned into a wonderful friendship. I wish you the best in your further career as a doctor, and hope I will have the opportunity to see many times more when visiting the south. Sergey, I very clearly remember meeting you on your first day at the office, and I feel we became friends almost as soon as we met. You make any task, be it during work or outside of it, seem easy. Many thanks to you and Kate as well for all the great times in Maastricht, and the help you provided with my PhD. Abdalla, the final member of our office. Thank you as well for the great time in Maastricht,

and for always having your house open for us to visit. Jose, thank you fun time we had. I was sad to see you leave for Australia, but I know you are having a great time there. Janita and Evelyn, thank you as well for the great time in Maastricht, and with all the help with my PhD, especially when I just moved from Amsterdam. I am glad we will see each other plenty still in the field of radiotherapy. Lastly, Will, thank you for being an amazing friend and for all the fond memories. I will always be thankful to you for hosting my 30th birthday party during Covid times. Visiting Maastricht will not be the same without you.

Ik zou ook mijn huidige opleider willen bedanken, Dr. Christoph Schneider. Hoewel je geen directe rol bij mijn promotie hebt gespeeld, heb ik het gevoeld dat je sinds mijn start op het AVL toch een onderdeel ervan bent geweest, en mij veel steun hebt gegeven. Ik zou je daarnaast enorm willen bedanken voor de kans die je mij hebt gegeven als klinisch fysicus in opleiding, een baan waar ik al een lange tijd naar toe werk. Verder wil ik ook mijn nieuwe collega's in Amsterdam bedanken, in het bijzonder Joost en de andere collega klifio's. Ik hoop dat we samen een fijne tijd zullen hebben, zowel tijdens als na de opleiding.

Aan al mijn vrienden in Amsterdam: ik heb jullie helaas wat minder gezien tijdens mijn verblijf in Maastricht. Ik ben blij terug te zijn in Amsterdam, en hoop dat we weer net zulke tijden als vroeger zullen meemaken. Specifiek wil ik Sten en Axel bedanken, twee van mijn oudste vrienden. Voor mij is een teken van een goede vriend dat je ermee kan praten na een lange tijd elkaar niet gezien te hebben alsof het gisteren was. Dit geldt zeker voor jullie.

Mam, Pap, Pieter en Anne-Marie: bedankt voor alle steun de laatste jaren. Ik heb een enorm leuke tijd in Maastricht gehad, maar was blij dat ik altijd in het weekend thuis kon komen, helemaal als Pieter en Anne-Marie er ook waren. Dat ik nota bene in Amsterdam ben aangenomen was een verrassing, maar ik ben blij dat ik nu weer dichtbij woon en er vaker kan zijn.

Ten slotte aan alle andere familieleden, vrienden, en kennissen: bedankt voor alle steun door de jaren heen, en bedankt voor het lezen van mijn proefschrift.

CHAPTER 13

13

Curriculum vitae

Simon was born at the BovenIJ ziekenhuis in Amsterdam the 12th of December, 1991. After finishing high-school at Gymnasium level at the Keizer Karel college in Amstelveen, he started the bachelor Medische Natuurwetenschappen, and subsequently the master Medical Natural sciences, at the Vrije Universiteit in Amsterdam. Following internships at the radiology and radiotherapy departments of the Vrije Universiteit Medisch Centrum, he graduated in 2016. During his graduation ceremony he also received a certificate from the NVKF, stating he is eligible for the training to become a clinical physicist.

Choosing to gain valuable experience in the field of radiotherapy and in academics first, Simon started his PhD in June 2016 at Maastricht in the team of Prof. Dr. Philippe Lambin, which later moved to the Precision Medicine department of Maastricht University. The theme of the research was to create and validate prognostic models for advanced tumor stages on medical imaging using machine- and deep learning algorithms. He was involved with every step of the research process, including but not limited to data collection and curation, tumor delineation, image pre-processing, feature extraction, the development, testing, and validation of data-driven models, the generation of results and the writing of the corresponding articles.

The body of research of the PhD included several projects categorized by different tumor sites. For head- and neck-cancer, he worked with Dr. Frank Hoebbers and Dr. Frederik Wesseling to develop and validate a machine learning model to stratify patients with locally advanced tumors into risk-groups based on their overall survival. This project was a part of the multi-center "Big Data and models for personalized Head and Neck Cancer support" EU-project, which included many trips to other European centers for collaboration and was finished with a presentation of the results of the project for a panel of experts in Luxembourg. The second tumor site was central nervous systems, specifically metastases to the brain. For this project Simon travelled to University of California - San Francisco for 5 months to collect and curate a large dataset of patients treated with Gamma Knife radiotherapy. Together with Dr. Olivier Morin, Dr. Martin Vallières, Dr. Steve Braunstein, and his colleague Manon Beuque he developed and validated a model to predict the risk of radiation necrosis. The last tumor site was lung cancer, for which he worked together with his co-promotor Lizza Hendriks to curate and delineate a dataset of CT-images of stage III lung cancer patients, and subsequently developed a model to predict the risk of metastasis to the brain.

Other tasks during his PhD include the mentoring of 1st year Biomedical sciences bachelor students, the co-supervision of honors students from 2020-2021, the supervision of bachelor and master internships at the Precision Medicine department, and the designing and supervising of the radiomics exercise of the AI4Imaging workshop in 2018 and 2019. Throughout his PhD, Simon has worked under a group of distinguished supervisors: his

promotor Prof. Dr. Philippe Lambin, and co-promotors Dr. Henry Woodruff and Dr. Lizza Hendriks.

Since June 2022, Simon has been in training as a clinical physicist, with a specialization in radiotherapy physics, at the Antoni van Leeuwenhoek hospital in Amsterdam.

CHAPTER 13

14

List of publications



14.1 Publications in this thesis

Rogers, W., **S. A. Keek**, M. Beuque, E. Lavrova, S. Primakov, G. Wu, C. Yan, S. Sanduleanu, H. A. Gietema, R. Casale, M. Occhipinti, H. C. Woodruff, A. Jochems and P. Lambin (2023). "Towards texture accurate slice interpolation of medical images using PixelMiner." Computers in Biology and Medicine **161**: 106701.

Keek, S. A., E. Kayan, A. Chatterjee, J. S. A. Belderbos, G. Bootsma, B. van den Borne, A.-M. C. Dingemans, H. A. Gietema, H. J. M. Groen, J. Herder, C. Pitz, J. Praag, D. De Ruyscher, J. Schoenmaekers, H. J. M. Smit, J. Stigt, M. Westenend, H. Zeng, H. C. Woodruff, P. Lambin and L. Hendriks (2022). "Investigation of the added value of CT-based radiomics in predicting the development of brain metastases in patients with radically treated stage III NSCLC." Therapeutic Advances in Medical Oncology **14**: 17588359221116605.

Keek, S. A., M. Beuque, S. Primakov, H. C. Woodruff, A. Chatterjee, J. E. van Timmeren, M. Vallières, L. E. L. Hendriks, J. Kraft, N. Andratschke, S. E. Braunstein, O. Morin and P. Lambin (2022). "Predicting Adverse Radiation Effects in Brain Tumors After Stereotactic Radiotherapy With Deep Learning and Handcrafted Radiomics." Frontiers in Oncology **12**.

Keek, S. A., F. W. R. Wesseling, H. C. Woodruff, J. E. van Timmeren, I. H. Nauta, T. K. Hoffmann, S. Cavalieri, G. Calareso, S. Primakov, R. T. H. Leijenaar, L. Licitra, M. Ravanelli, K. Scheckenbach, T. Poli, D. Lanfranco, M. R. Vergeer, C. R. Leemans, R. H. Brakenhoff, F. J. P. Hoebbers and P. Lambin (2021) "A Prospectively Validated Prognostic Model for Patients with Locally Advanced Squamous Cell Carcinoma of the Head and Neck Based on Radiomics of Computed Tomography Images." Cancers **13** DOI: 10.3390/cancers13133271.

Keek, S., S. Sanduleanu, F. Wesseling, R. de Roest, M. van den Brekel, M. van der Heijden, C. Vens, C. Giuseppina, L. Licitra, K. Scheckenbach, M. Vergeer, C. R. Leemans, R. H. Brakenhoff, I. Nauta, S. Cavalieri, H. C. Woodruff, T. Poli, R. Leijenaar, F. Hoebbers and P. Lambin (2020). "Computed tomography-derived radiomic signature of head and neck squamous cell carcinoma (peri)tumoral tissue for the prediction of locoregional recurrence and distant metastasis after concurrent chemo-radiotherapy." PLOS ONE **15**(5): e0232639.

Keek, S. A., R. T. H. Leijenaar, A. Jochems and H. C. Woodruff (2018). "A review on radiomics and the future of theranostics for patient selection in precision medicine." The British Journal of Radiology **91**(1091): 20170926.

14.2 Publications outside of this thesis

Primakov, S. P., A. Ibrahim, J. E. van Timmeren, G. Wu, **S. A. Keek**, M. Beuque, R. W. Y. Granzier, E. Lavrova, M. Scrivener, S. Sanduleanu, E. Kayan, I. Halilaj, A. Lenaers, J. Wu, R. Monshouwer, X. Geets, H. A. Gietema, L. E. L. Hendriks, O. Morin, A. Jochems, H. C. Woodruff and P. Lambin (2022). "Automated detection and segmentation of non-small cell lung cancer computed tomography images." *Nature Communications* **13**(1): 3423.

Granzier, R. W. Y., A. Ibrahim, S. Primakov, **S. A. Keek**, I. Halilaj, A. Zwanenburg, S. M. E. Engelen, M. B. I. Lobbes, P. Lambin, H. C. Woodruff and M. L. Smidt (2022). "Test-Retest Data for the Assessment of Breast MRI Radiomic Feature Repeatability." *Journal of Magnetic Resonance Imaging* **56**(2): 592-604.

Cavalieri, S., L. De Cecco, R. H. Brakenhoff, M. S. Serafini, S. Canevari, S. Rossi, D. Lanfranco, F. J. P. Hoebbers, F. W. R. Wesseling, **S. Keek**, K. Scheckenbach, D. Mattavelli, T. Hoffmann, L. López Pérez, G. Fico, M. Bologna, I. Nauta, C. R. Leemans, A. Trama, T. Klausch, J. H. Berkhof, V. Tountopoulos, R. Shefi, L. Mainardi, F. Mercalli, T. Poli, L. Licitra and B. D. D. C. the (2021). "Development of a multiomics database for personalized prognostic forecasting in head and neck cancer: The Big Data to Decide EU Project." *Head & Neck* **43**(2): 601-612.

Sanduleanu, S., **S. Keek**, L. Hoezen and P. Lambin (2021). *Biomarkers for Hypoxia, HPVness, and Proliferation from Imaging Perspective*. Critical Issues in Head and Neck Oncology: Key Concepts from the Seventh THNO Meeting, Springer International Publishing Cham.

Rogers, W., S. Thulasi Seetha, T. A. G. Refaee, R. I. Y. Lieverse, R. W. Y. Granzier, A. Ibrahim, **S. A. Keek**, S. Sanduleanu, S. P. Primakov, M. P. L. Beuque, D. Marcus, A. M. A. van der Wiel, F. Zerka, C. J. G. Oberije, J. E. van Timmeren, H. C. Woodruff and P. Lambin (2020). "Radiomics: from qualitative to quantitative imaging." *The British Journal of Radiology* **93**(1108): 20190948.

O'Farrell, A. C., M. A. Jarzabek, A. U. Lindner, S. Carberry, E. Conroy, I. S. Miller, K. Connor, L. Shiels, E. R. Zanella, F. Lucantoni, A. Lafferty, K. White, M. Meyer Villamandos, P. Dicker, W. M. Gallagher, **S. A. Keek**, S. Sanduleanu, P. Lambin, H. C. Woodruff, A. Bertotti, L. Trusolino, A. T. Byrne and J. H. M. Prehn (2020) "Implementing Systems Modelling and Molecular Imaging to Predict the Efficacy of BCL-2 Inhibition in Colorectal Cancer Patient-Derived Xenograft Models." *Cancers* **12** DOI: 10.3390/cancers12102978.

Bogowicz, M., A. Jochems, T. M. Deist, S. Tanadini-Lang, S. H. Huang, B. Chan, J. N. Waldron, S. Bratman, B. O'Sullivan, O. Riesterer, G. Studer, J. Unkelbach, S. Barakat, R. H. Brakenhoff, I. Nauta, S. E. Gazzani, G. Calareso, K. Scheckenbach, F. Hoebbers, F. W. R. Wesseling, **S. Keek**,

S. Sanduleanu, R. T. H. Leijenaar, M. R. Vergeer, C. R. Leemans, C. H. J. Terhaard, M. W. M. van den Brekel, O. Hamming-Vrieze, M. A. van der Heijden, H. M. Elhalawani, C. D. Fuller, M. Guckenberger and P. Lambin (2020). "Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer." Scientific Reports **10**(1): 4542.

Ibrahim, A., M. Vallières, H. Woodruff, S. Primakov, M. Beheshti, **S. Keek**, T. Refaee, S. Sanduleanu, S. Walsh, O. Morin, P. Lambin, R. Hustinx and F. M. Mottaghy (2019). "Radiomics Analysis for Clinical Decision Support in Nuclear Medicine." Seminars in Nuclear Medicine **49**(5): 438-449.

Walsh, S., E. E. de Jong, J. E. van Timmeren, A. Ibrahim, I. Compter, J. Peerlings, S. Sanduleanu, T. Refaee, **S. Keek** and R. T. Larue (2019). "Decision support systems in oncology." JCO clinical cancer informatics **3**: 1-9.

14.3 Abstracts

Hendriks, L., **S. A. Keek**, A. Chatterjee, J. Belderbos, G. Bootsma, B. van den Borne, A. M. C. Dingemans, H. Gietema, H. J. M. Groen, G. Herder, C. Pitz, J. Praag, D. De Ruyscher, J. Schoenmaekers, H. J. M. Smit, J. Stigt, M. Westenend, H. Zeng, H. Woodruff and P. Lambin (2022). "127P Does radiomics have added value in predicting the development of brain metastases in patients with radically treated stage III non-small cell lung cancer (NSCLC)?" Annals of Oncology **33**: S91.

Primakov, S., A. Ibrahim, J. van Timmeren, G. Wu, **S. Keek**, M. Beuque, R. Granzier, M. Scrivener, S. Sanduleanu and E. Kayan (2021). "OC-0557 AI-based NSCLC detection and segmentation: Faster and more prognostic than manual segmentation." Radiotherapy and Oncology **161**: S441-S443.

Keek, S., F. Wesseling, H. Woodruff, J. van Timmeren, I. Nauta, T. Hoffmann, S. Cavalieri, G. Calareso, S. Primakov and R. Leijenaar (2021). "OC-0642 A radiomics based prognostic model for patients with head and neck squamous cell carcinoma." Radiotherapy and Oncology **161**: S509-S510.

