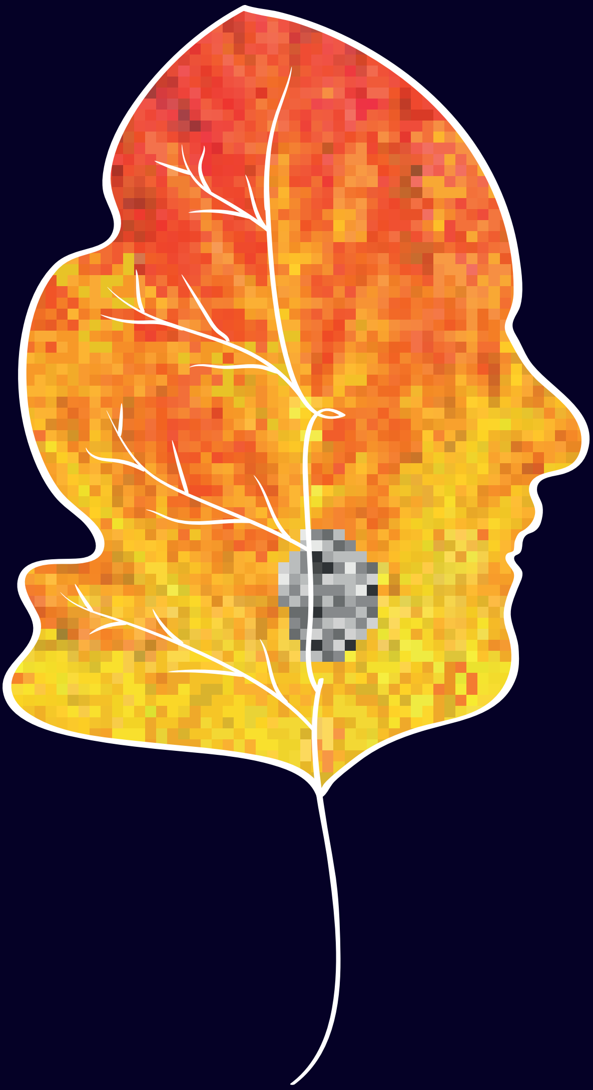
Magnetic resonance based radiomics

in oropharyngeal cancer



Paula Bos

Magnetic resonance based radiomics

in oropharyngeal cancer

Paula Bos

COLOFON

The work described in this thesis was performed at the Netherlands Cancer Institute – Antoni van Leeuwenhoek, Amsterdam, The Netherlands.

Cover art:	Paula Bos
Lay-out:	Paula Bos
Printed by:	Ridderprint| www.ridderprint.nl
ISBN:	978-94-6458-447-9

The Netherlands Cancer Institute, Maastricht University, Verwelius Foundation and Willem Meindert de Hoop Stiching kindly provided financial support for publication of this thesis.

Magnetic resonance based radiomics in oropharyngeal cancer

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 maandag 26 september 2022 om 10.00 uur

door

Paula Bos

Promotores

Prof. dr. R.G.H. Beets-Tan, Antoni van Leeuwenhoek
Prof. dr. M.W.M. van den Brekel, Antoni van Leeuwenhoek
Prof. dr. J.A. Castelijns, Antoni van Leeuwenhoek

Copromotor

Dr. B. Jasperse, Amsterdam UMC (locatie VUmc)

Beoordelingscommissie

Prof. dr. B. Kremer (voorzitter)
Prof. dr. R. de Bree, UMC Utrecht
Prof. dr. ir. A.L.A.J. Dekker
Dr. F.J.P. Hoebers
Dr. S. Steens, Radboud UMC

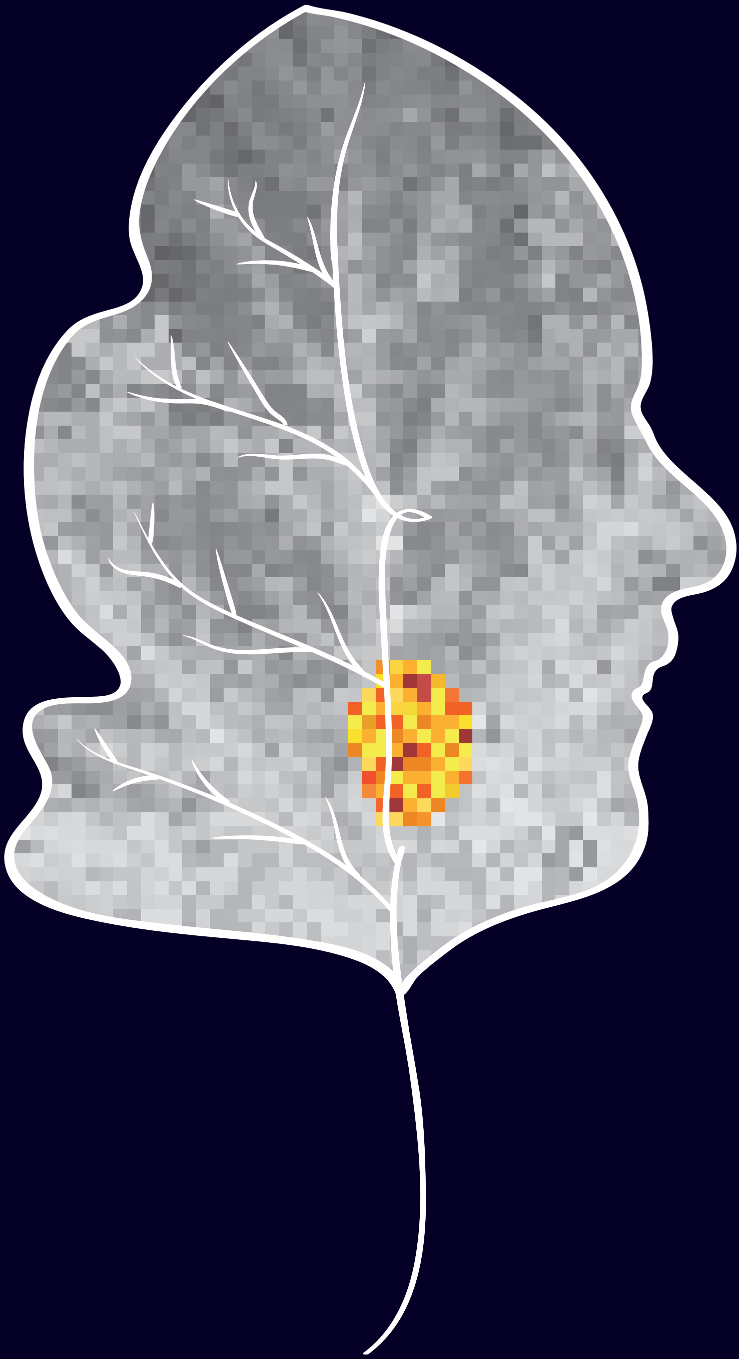


Table of contents

General introduction

1

Head and neck cancer is diagnosed in more than 900,000 patients in 2020 worldwide, representing a serious problem. Head and neck cancer occurs in various subsites of this anatomic region, each with particular characteristics with regard to tumor type, growth and likelihood of metastasizing. Because of this, and the resulting differences in treatment options and planning, these subsites can be considered as different entities. Oropharyngeal cancer is the primary topic of this thesis, and is responsible for approximately 98,412 new cases and 48,143 deaths worldwide in 2020[1]. Squamous cell carcinoma's (SCC) is the most common histopathological subtype, occurring in approximately 90% of all oropharyngeal cancer cases. Alcohol consumption, tobacco smoking and human papillomavirus (HPV) infection are the risk factors of oropharyngeal squamous cell carcinoma's (OPSCC)[1–3], which may present with pain in the throat radiating to the ear, bad breath, a neck mass and/or difficulties with chewing and/or swallowing[3–5]. As these signs and symptoms present when the tumor has metastasized or has a considerable size, OPSCC is mostly discovered in a late stage of the disease[4].

The diagnostic work-up for OPSCC consists of tumor localization, tumor size, tumor invasion and the presence of cancer cells outside the primary tumor determined by clinical and imaging evaluation. Clinical examination is performed as initial diagnosis of OPSCC patients, including fiberoptic endoscopy to determine the superficial extent of the tumor and to take biopsies. Medical imaging is acquired to assess the extent of the disease. Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) evaluate the extent and invasiveness of the primary tumor and metastases to regional lymph nodes. Ultrasound (US) is used to assess the lymph nodes in more detail and can provide pathological confirmation of suspected lymph node metastases by fine-needle aspiration. Positron Emission Tomography (PET) imaging is used to determine the presence of distant metastasis[4,6]. All this information is used to classify the tumor using the Tumor Node Metastases (TNM) classification system from the American Joint Committee for Cancer (AJCC)[7]. This international guideline provides a standardized classification to come to a uniform assessment for prognosis and treatment planning.

The large variety in tumor location and genetic behavior ask for a personalized treatment path for each patient. Radiation therapy (RT), chemotherapy (CT), surgery or a combination of these treatments are currently the most commonly used treatment strategies. With regard to chemotherapy, cisplatin based chemoradiation therapy (CRT) is recommended for advanced stage OPSCC[6]. Alternatives for cisplatin are carboplatin, cetuximab or sometimes 5-Fluorouracil. In recent years, the first trials evaluating neoadjuvant immunotherapy with curative intent as alternative treatment are described, with promising results[8–10]. Still, despite continuous improved treatment options, curation is not achieved in 25-45% of the patients[11]. To improve on this, accurate prognostic tools are needed that can further improve the identification of patients who will benefit from intensification of treatment and those for whom the treatment would do more harm than good.

Evidence of HPV infection of OPSSC, using immunohistochemistry or DNA polymerase chain reaction (PCR) techniques on biopsy material, is one of the promising prognostic tools. In the last decennia, the incidence of HPV related OPSCC is 2.5 times higher than the incidence rate of HPV-negative OPSCC[12,13] due to a decline in smoking patterns combined with changes in sexual behavior. Evidence emerged that HPV-positive and HPV-negative OPSCCs can be considered as distinct entities, each with unique clinical, histological and biological profiles[12,14–16]. Compared to non-HPV related OPSCC, HPV related OPSCC occur more frequently in younger patients[17,18], is likely to present with a relative small tumor size and a relatively high rate of nodal metastasis. The lower rate of genetic alterations, tumor dedifferentiation and non-keratinizing pathology explain the better prognosis compared to patients with HPV negative tumors. Detection of HPV infection needs invasive tissue diagnosis, that is time consuming and expensive, using p16/p53 immunohistochemistry and/or HPV DNA polymerase chain reaction (PCR). Additionally, although prognostication based on HPV status has the potential to guide treatment, it is still a single characteristic, with limited influence on personalized treatment so far.

Another prognostic tool used in clinical practice is TNM classification. This anatomic-based classification is based on tumor size (T), involvement and extent of regional lymph nodes (N), and the presence of distant metastases (M)[7]. While HPV status is adopted in the latest TNM edition (8th)[7], other factors evaluating the biological behavior of the disease remain excluded (e.g. immune features)[19]. All biologic information contains relevant parameters to take into consideration since they affect prognosis. Therefore, a multiparameter assessment evaluating clinical, pathological and biomolecular information is needed to improve therapeutic decisions[19].

Since the identification of radiological tumor volume as prognostic parameter in the 1990s[20], various radiological features and parameters have been considered and proven as a means to predict treatment outcome and prognosis. These include structural radiological features like tumor invasion in surrounding tissues, the detection and quantification of metastasis, and presence of extracapsular extension of lymph nodes[21–24], and functional imaging parameters that reflect tumor cellularity, perfusion and metabolic activity[25–28]. In general, these radiological features are interpreted based on quantification and interpretation of a single parameter in isolation without regard for detailed textural information.

Over the last decade new computer aided techniques have made it possible to quantify visually occult textural characteristics from medical images, commonly referred to as radiomics[29–31]. Radiomics is implemented by delineating a region of interest (ROI) on the radiological image by an observer. Various features that describe image intensity, shape and texture are then extracted from within this ROI. These radiomic features are then statistically analyzed to determine which features are important in prediction of an outcome variable (like treatment outcome) using classical statistics or artificial intelligence methods. The resulting statistical model can then be used in clinical practice to predict the outcome variable of interest for an individual patient.

Radiomics proved to be promising in the reflection of tumor biology[32–36] and the prediction of treatment outcomes[29–31,37–40]. With regard to the head and neck region, radiomic features have been associated with gene expression[33], histological tissue properties[32,34], and treatment outcome[29–31,37,38]. As such, texture-based analysis showed to be a useful tool in the discrimination of benign and malignant tumors[32]. Another study proved the prognostic value in the prediction of survival by a four-feature radiomic signature, indicating that radiomic features capture intra-tumour heterogeneity[29].

Current radiomics research on head and neck cancer used mostly radiomic features extracted from CT. Compared to CT, MRI has a better soft-tissue contrast and superior sensitivity in detecting small lesions or invasion of tissues surrounding the tumor. Additionally, it may provide other insights in tissue properties due to fundamental differences in image acquisition[41,42]. Only few studies have investigated prognostic radiomic features from MR images of head and neck cancer. These studies mainly focused on outcome prediction for nasopharyngeal carcinoma using radiomics or deep learning[30,41,43,44]. Studies on MR-based radiomics for OPSCC are lacking thus far, probably due to the challenging anatomy and the acquired MR signal intensities which are influenced by acquisition-related factors[43].

Therefore, the overall goal of this thesis is to investigate the potential of radiomics in oropharyngeal cancer using features extracted from diagnostic MRI. This was formulated in three main research questions.

Part I: What is the current knowledge on MR-based functional parameters in head and neck squamous cell carcinoma?

Radiological biomarkers are often used to assist the radiologist in the tumor diagnosis and treatment-decision for the patient. Volumetric biomarkers, such as tumor volume, shape and diameter, are simple tumor characteristics extracted by the radiologist. The introduction of functional imaging (dynamic contrast-enhancement MRI, diffusion weighted MRI) initiates the ability to extract parameters describing the microenvironment of the tissue. **Chapter 2** gives an extensive overview of the current level of evidence for pre-treatment MR-based perfusion and diffusion imaging parameters that are prognostic for treatment outcome in head and neck squamous cell carcinoma.

Part II: Can MR-based radiomic prediction models be used for tumor characterization and prognosis in OPSCC patients?

Prediction models in OPSCC are limited to features extracted from CT imaging, where MRI can represent other tissue properties. Therefore, in **chapter 3** a radiomics model is constructed based on pre-treatment post-contrast MRI to predict treatment outcome. Generalizability of this single-center model is evaluated using an independent external validation dataset, described in **chapter 4**. Treatment outcomes are significantly better for patients with HPV positive tumors compared to patients without HPV infected tumors. Determination of tumoral HPV status is nowadays done using invasive tissue based immunohistochemistry. **Chapter 5** addresses a non-invasive technique using MR-based radiomics to predict HPV status of the tumor in OPSCC patients.

Part III: Can MR-based radiomics for OPSCC patients be simplified using alternative or automated delineation techniques to improve clinical adoption?

3D tumor delineations by an experienced radiologist are currently needed for adequate radiomics analysis of OPSCC. These delineations take up a costly amount of time and may therefore hamper the adoption of radiomics in clinical practice. In an attempt to reduce the time needed for OPSCC delineation, **chapter 6** and **chapter 7** investigates if six different manual delineation strategies can affect performance in models predictive of LRC and HPV status, respectively. Besides time consumption, manual delineations are known to vary across observers. **Chapter 8** proposed a semi-automatic approach for tumor segmentation based on deep learning that may ameliorate time consuming manual delineations and the related interobserver variability.

REFERENCES

1. WHO. The Global Cancer Observatory, 2020.
2. Warnakulasuriya S. Global epidemiology of oral and oropharyngeal cancer. *Oral Oncol*. 2009;45(4-5):309-316. doi:10.1016/j.oraloncology.2008.06.002
3. Chi AC, Day TA, Neville BW. Oral cavity and oropharyngeal squamous cell carcinoma-an update. *CA Cancer J Clin*. 2015;65(5):401-421. doi:10.3322/caac.21293
4. McIlwain WR, Sood AJ, Nguyen SA, Day TA. Initial symptoms in patients with HPV-positive and HPV-negative oropharyngeal cancer. *JAMA Otolaryngol Head Neck Surg*. 2014;140(5):441-447. doi:10.1001/jamaoto.2014.141
5. Carpén T, Sjöblom A, Lundberg M, et al. Presenting symptoms and clinical findings in HPV-positive and HPV-negative oropharyngeal cancer patients. *Acta Otolaryngol*. 2018;138(5):513-518. doi:10.1080/00016489.2017.1405279
6. Pfister DG, Spencer S, Adelstein D, et al. Head and neck cancers, version 2. 2020, NCCN Clinical practice guidelines in oncology. *J Natl Compr Cancer Netw*. 2020;18(7):873-898. doi:10.6004/jnccn.2020.0031
7. Amin MB, Edge SB, Greene FL, et al. AJCC Cancer Staging Manual. Eight ed., Springer, 2017
8. Elbers JBW, Al-Mamgani A, Tesseslaar MET, et al. Immuno-radiotherapy with cetuximab and avelumab for advanced stage head and neck squamous cell carcinoma: Results from a phase-I trial. *Radiother Oncol*. 2020;142:79-84. doi:10.1016/j.radonc.2019.08.007
9. Stafford M, Kaczmar J. The neoadjuvant paradigm reinvigorated: a review of pre-surgical immunotherapy in HNSCC. *Cancers Head Neck*. 2020;5(4). doi:10.1186/s41199-020-00052-8
10. Hay A, Nixon IJ. Recent advances in the understanding and management of oropharyngeal cancer. *F1000Res*. 2018;7(F1000). doi:10.12688/1000research.14416.1
11. Pignon J-P, le Maître A, Maillard E, Bourhis J. Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): An update on 93 randomised trials and 17,346 patients. *Radiother Oncol*. 2009;92(1):4-14. doi:10.1016/j.radonc.2009.04.014
12. Mahal BA, Catalano PJ, Haddad RI, et al. Incidence and demographic burden of HPV-associated oropharyngeal head and neck cancers in the United States. *Cancer Epidemiol Biomarkers Prev*. 2019;28(10):1660-1667. doi:10.1158/1055-9965.EPI-19-0038
13. Henneman R, van Monsjou HS, Verhagen CVM, et al. Incidence changes of human papillomavirus in oropharyngeal squamous cell carcinoma and effects on survival in the Netherlands Cancer Institute, 1980-2009. *Anticancer Res.* 2015;35(7):4015-4022.

14. Elrefaey S, Massaro M, Chiocca S, Chiesa F, Ansarin M. HPV in oropharyngeal cancer: The basics to know in clinical practice. *Acta Otorhinolaryngol Ital*. 2014;34(5):299-309.

15. Taberna M, Mena M, Pavón MA, Alemany L, Gillison ML, Mesía R. Human papillomavirus-related oropharyngeal cancer. *Ann Oncol*. 2017;28(10):2386-2398. doi:10.1093/annonc/mdx304

16. Van Monsjou HS, Balm AJM, van den Brekel MWM, Wreesmann VB. Oropharyngeal squamous cell carcinoma: A unique disease on the rise? *Oral Oncol*. 2010;46(11):780-785. doi:10.1016/j.oraloncology.2010.08.011

17. Bajpai S, Zhang N, Lott DG. Tracking changes in age distribution of head and neck cancer in the United States from 1975 to 2016. *Clin Otolaryngol*. 2021;46(6):1205-1212. doi:10.1111/coa.13817

18. Chaturvedi AK, Engels EA, Pfeiffer RM, et al. Human papillomavirus and rising oropharyngeal cancer incidence in the United States. *J Clin Oncol*. 2011;29(32):4294-4301. doi:10.1200/JCO.2011.36.4596

19. Denaro N, Russi EG, Merlano MC. Pros and Cons of the New Edition of TNM Classification of Head and Neck Squamous Cell Carcinoma. *Oncology*. 2018;95(4):202-210. doi:10.1159/000490415

20. Castelijns JA, Golding RP, van Schaik C, Valk J, Snow GB. MR findings of cartilage invasion by laryngeal cancer: value in predicting outcome of radiation therapy. *Radiology*. 1990;174(3):669-673. doi:10.1148/radiology.174.3.2305047

21. van den Brekel MWM, Castelijns JA. Radiologic evaluation of neck metastases: The otolaryngologist’s perspective. *Semin Ultrasound CT MR.* 1999;20(3):162-174. doi:10.1016/s0887-2171(00)90017-3

22. Van den Brekel MWM, Bindels EMJ, Balm AJM. Prognostic factors in head and neck cancer. *Eur J Cancer.* 2002;38(8):1041-1043. doi:10.1016/s0959-8049 (02)00023-0

23. Castelijns JA, Becker M, Hermans R. Impact of cartilage invasion on treatment and prognosis of laryngeal cancer. *Eur Radiol.* 1996;6(2):156-169. doi:10.1007/BF00181135

24. Mukherji SK, O’Brien SM, Gerstle RJ, Weissler M, Shockley W, Castillo M. Tumor volume: an independent predictor of outcome for laryngeal cancer. *J Comput Assist Tomogr*. 1999;23(1):50-54. doi:10.1097/00004728-199901000-00011

25. Cao Y, Aryal M, Li P, et al. Predictive Values of MRI and PET Derived Quantitative Parameters for Patterns of Failure in Both p16+ and p16– High Risk Head and Neck Cancer. *Front Oncol*. 2019;9:1118. doi:10.3389/fonc.2019.01118

26. Martens RM, Noij DP, Ali M, et al. Functional imaging early during (chemo) radiotherapy for response prediction in head and neck squamous cell carcinoma; a systematic review. *Oral Oncol*. 2019;88:75-83. doi:10.1016/j.oraloncology.2018.11.005

27. Garbajs M, Strojan P, Surlan-Popovic K. Prognostic role of diffusion weighted and dynamic contrast-enhanced MRI in loco-regionally advanced head and neck cancer treated with concomitant chemoradiotherapy. *Radiol Oncol*. 2019;53(1):39-48. doi:10.2478/raon-2019-0010
28. Wong KH, Panek R, Dunlop A, et al. Changes in multimodality functional imaging parameters early during chemoradiation predict treatment response in patients with locally advanced head and neck cancer. *Eur J Nucl Med Mol Imaging*. 2018;45(5):759-767. doi:10.1007/s00259-017-3890-2
29. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006. doi:10.1038/ncomms5006
30. Zhai TT, van Dijk LV, Huang BT, et al. Improving the prediction of overall survival for head and neck cancer patients using image biomarkers in combination with clinical parameters. *Radiother Oncol*. 2017;124(2):256-262. doi:10.1016/j.radonc.2017.07.013
31. Mes SW, van Velden FHP, Peltenburg B, et al. Outcome prediction of head and neck squamous cell carcinoma by MRI radiomic signatures. *Eur Radiol*. 2020;30(11):6311-6321. doi:10.1007/s00330-020-06962-y
32. Fruehwald-Pallamar J, Hesselink JR, Mafee MF, Holzer-Fruehwald L, Czerny C, Mayerhoefer ME. Texture-Based analysis of 100 MR examinations of head and neck tumors - Is it possible to discriminate between benign and malignant masses in a multicenter trial? *Rofo*. 2016;188(2):195-202. doi:10.1055/s-0041-106066
33. Yu K, Zhang Y, Yu Y, et al. Radiomic analysis in prediction of Human Papilloma Virus status. *Clin Transl Radiat Oncol*. 2017;7:49-54. doi:10.1016/j.ctro.2017.10.001
34. Romeo V, Cuocolo R, Ricciardi C, et al. Prediction of tumor grade and nodal status in oropharyngeal and oral cavity squamous-cell carcinoma using a radiomic approach. *Anticancer Res*. 2020;40(1):271-280. doi:10.21873/anticanres.13949
35. Zhou L, Zhang Z, Chen YC, Zhao ZY, Yin XD, Jiang HB. A deep learning-based radiomics model for differentiating benign and malignant renal tumors. *Transl Oncol*. 2019;12(2):292-300. doi:10.1016/j.tranon.2018.10.012
36. Zhang Q, Peng Y, Liu W, et al. Radiomics based on multimodal MRI for the differential diagnosis of benign and malignant breast lesions. *J Magn Reson Imaging*. 2020;52(2):596-607. doi:10.1002/jmri.27098
37. Chu CS, Lee NP, Adeoye J, Thomson P, Choi SW. Machine learning and treatment outcome prediction for oral cancer. *J Oral Pathol Med*. 2020;49(10):977-985. doi:10.1111/jop.13089
38. Yuan Y, Ren J, Shi Y, Tao X. MRI-based radiomic signature as predictive marker for patients with head and neck squamous cell carcinoma. *Eur J Radiol*. 2019;117:193-198. doi:10.1016/j.ejrad.2019.06.019

39. Staal FCR, van der Reijd DJ, Taghavi M, Lambregts DMJ, Beets-Tan RGH, Maas M. Radiomics for the prediction of treatment outcome and survival in patients with colorectal cancer: A systematic review. *Clin Colorectal Cancer*. 2021;20(1):52-71. doi:10.1016/j.clcc.2020.11.001

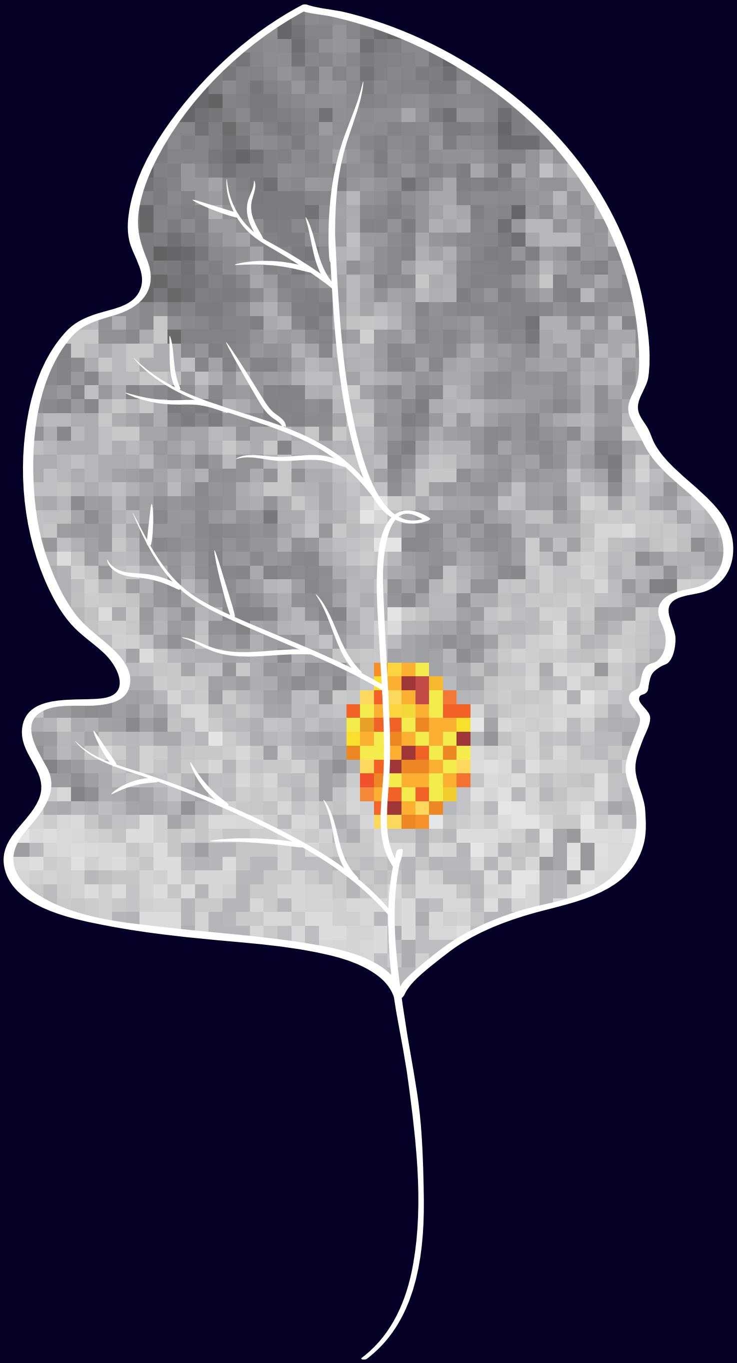
40. van Griethuysen JJM, Lambregts DMJ, Trebeschi S, et al. Radiomics performs comparable to morphologic assessment by expert radiologists for prediction of response to neoadjuvant chemoradiotherapy on baseline staging MRI in rectal cancer. *Abdom Radiol*. 2020;45(3):632-643. doi:10.1007/s00261-019-02321-8

41. Liu Z, Wang S, Dong D, et al. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. *Theranostics*. 2019;9(5):1303-1322. doi:10.7150/thno.30309

42. Castelijns JA, van den Brekel MWM. Magnetic resonance imaging evaluation of extracranial head and neck tumors. *Magn Reson Q*. 1993;9(2):113-128.

43. Jethanandani A, Lin TA, Volpe S, et al. Exploring applications of radiomics in Magnetic Resonance Imaging of head and neck Cancer: A systematic review. *Front Oncol*. 2018;8:131. doi:10.3389/fonc.2018.00131

44. Farhidzadeh H, Kim JY, Scott JG, Goldgof DB, Hall LO, Harrison LB. Classification of progression free survival with nasopharyngeal carcinoma tumors. *SPIE*. 2016;9785. doi:10.1117/12.2216976



Part I

Current knowledge of MR-based functional parameters in head and neck squamous cell carcinoma

Prognostic functional MR imaging parameters in head and neck squamous cell carcinoma: A systematic review

European Journal of Radiology; Nov 2021;144:109952
DOI: 10.1016/j.ejrad.2021.109952

Paula Bos
Hedda J. van der Hulst
Michiel W.M. van den Brekel
Winnie Schats
Bas Jasperse
Regina G.H. Beets-Tan
Jonas A. Castelijns

2

ABSTRACT

Objective*:* Functional MR imaging has demonstrated potential for predicting treatment response. This systematic review gives an extensive overview of the current level of evidence for pre-treatment MR-based perfusion and diffusion imaging parameters that are prognostic for treatment outcome in head and neck squamous cell carcinoma (HNSCC) (PROSPERO registration: CRD42020210689).

Materials and methods*:* According to the PRISMA statements, Medline, Embase and Scopus were queried for articles with a maximum date of October 19th, 2020. Studies investigating the predictive performance of pre-treatment MR-based perfusion and/or diffusion imaging parameters in HNSCC treatment response were included. All prognosticators were extracted from the primary tumor. Risk of bias was assessed using the QUIPS tool. Results were summarized in tables and forest plots.

Results*:* 31 unique studies met the inclusion criteria; among them, 11 articles described perfusion (n=529 patients) and 28 described diffusion (n=1626 patients) MR-imaging, eight studies were included in both categories. Higher K^{trans} and K_{ep} were associated with better treatment response for OS and DFS, respectively. Study findings for V_p and V_e were inconsistent or not significant. High-level controversy was observed between studies examining the MR diffusion parameters mean and median ADC.

Conclusion*:* For HNSCC patients, the accurate and consistent results of pre-treatment MR-based perfusion parameters K^{trans} and K_{ep} are potential for clinical applicability predictive of OS and DFS and treatment decision guidance. Significant heterogeneity in study designs might affect high discrepancy in study results for parameters extracted from diffusion imaging. Furthermore, recommendations for future research were summarized.

INTRODUCTION

Head and neck cancer is the sixth most common cancer, with an incidence of 5.3% of all new cancer cases worldwide[1]. Out of all head and neck cancers, most malignancies (>90%) are head and neck squamous cell carcinomas (HNSCC)[2]. Currently, radiation in combination with chemotherapy (chemoradiation (CRT)) is the standard of care for most patients with locally advanced tumors, with surgical resection and immunotherapy as alternative and/or upcoming strategies. Despite continuous improvement of the treatment options, treatment is not successful in 25–30% of the patients[3]. Thus, there is an urgent need for reliable biomarkers to predict treatment outcome, and fine-tune treatment strategies when desirable.

Over the last decades, several prognostic markers of treatment response have been studied, with the importance of HPV status gaining prominence. Furthermore, imaging markers, such as functional imaging parameters, appear to show prognostic value in determining pre-treatment treatment response. Dynamic contrast-enhanced (DCE) MRI and diffusion-weighted (DW) MRI are two common studied functional imaging modalities of which parameters are derived in HNSCC.

DCE-MRI is a contrast-based MRI technique to visualize tissue perfusion, expressed as the change in the concentration of contrast agent in the field of view. The most used pharmacokinetic model is the Tofts model[4]. Here, contrast agents are delivered by blood vessels and exchanged with the extravascular extracellular space (EES). The influx of this contrast agent between the blood vessel and the EES is defined by the K^{trans} parameter of the Tofts model, whereas K_{ep} is this reverse process, the reflux rate. Since contrast agents, like Gadolinium, are not absorbed by cells, their concentration depends on the plasma volume (V_p) and EES volume (V_e). Tissue perfusion depends on blood volume and blood flow.

DWI quantifies the diffusion of tissue water molecules in tissue volume, expressed as the apparent diffusion coefficient (ADC). Repetitions of the DWI sequence with different diffusion strengths (b-values) visualize this water displacement. The choice of b-value depends on the velocity of water diffusion, where high b-values (≥250 s/mm^2) are recommended to measure slow diffusion and low values (≤250 s/mm^2) for fast diffusion of water molecules as in flow within vessels[5,6]. Besides diffusion, ADC also includes signals caused by micro-vascularization. The intravoxel incoherent motion (IVIM)[7] model accounts for this in its bi-exponential model, resulting in the main parameters molecular diffusion, *D*, pseudo-diffusion coefficient, *D**, and vascular volume fraction (*f*).

Various studies described these functional MRI parameters for the prediction of treatment response, with encouraging results. Previous systematic reviews[8-11] already summarized the prognostic value of DCE or DW imaging separately. However, the number of studies is rapidly increasing and hence, there is a need to revise the existing reviews on the current value of both MRI parameters. Additionally, we will primarily focus on pre-treatment imaging biomarkers to stratify personalized treatment to obtain the best treatment outcome while minimizing harmful side effects. This review will give an update on current literature describing the prognostic value of pre-treatment DCE or DWI parameters extracted from the primary tumor of HNSCC to predict treatment response. This review also summarizes issues for future research.

MATERIALS AND METHODS

This systematic review (PROSPERO registration: CRD42020210689) was performed following the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) criteria[12].

Search strategy

A systematic search was conducted using Medline, Embase, and Scopus for original articles published until October 19th 2020. The search consists of a combination of the search terms “Head and neck cancer”, “MRI”, and “treatment outcome”, with their synonyms. The full literature search is described in Appendix A1. Due to the expeditious improvement of MRI techniques and quality over the last decades, only studies published in the last ten years (after 2009) were included.

Study selection

Only studies in the English language investigating the prognostic performance of pre-treatment MR-based imaging parameters on treatment response were included. Additionally, these studies had to 1) extract imaging parameters from the primary tumor; 2) examine HNSCCs patients at any age, gender, and stage; and 3) treatment response prognostication ≥ 12 months after diagnosis. Studies were excluded when 1) the study design involved reviews, guidelines, conference abstracts, posters, case reports, or technical notes; 2) the study population consisting of tumors originating from other subsites than the oral cavity, oropharynx, hypopharynx, or larynx, to prevent bias from nasopharyngeal tumors which is seen as a different entity; 3) animal studies; and 4) the examined imaging parameters were extracted from the lymph nodes, adenocarcinomas or patients with recurrence.

Relevant articles were independently selected by the first two authors (PB and

HH) using Rayyan[13]. The above-described inclusion criteria were used as criteria during the title and abstract selection. The applicability of each study was assessed during full-text screening using a standardized form (see Appendix A2, page 1. Here, a cutoff value of 20% or 80% was used for some inclusion criteria to aim for an as much as possible homogeneous population). Discrepancies between reviewers were discussed in consensus. Consensus meetings have been done after 500, 2500, and all examined titles and abstracts. Articles with a serious concern of applicability were excluded. When needed, authors of potentially relevant articles were contacted to get full-text access.

Quality assessment

The quality of the eligible articles was assessed independently by the two reviewers, using criteria of an optimized version of the QUIPS tool[14] (see Appendix A2). Initial disagreement between reviewers was resolved by discussion. If a consensus was not reached, a third reviewer (JC) participated in the discussion and had the decisive vote.

Data extraction and analysis

Data was independently extracted using a standardized extraction form by two reviewers (PB and HH). Uncertainties were resolved by discussion. Data extraction included the categories:

- Study characteristics: Study design, author, year of publication
- Patient characteristics: Number of included and analyzed patients, gender, cancer subsite, tumoral HPV status, tumor stage (TNM), AJCC stage, obtained treatment
- Imaging characteristics: Time between pre-treatment scan and start of treatment, MR pulse sequences, field strength, b-values, echo time (TE), repetition time (TR), field of view (FOV), slice thickness, acquisition matrix, acquisition time, production of ADC values and DCE parameters (i.e. pharmacokinetic model, arterial input function parameters, T1 mapping)
- Delineation characteristics: Delineation methodology (i.e. whole tumor volume, single slice delineation), used reference sequence, the number of delineation observers, (no/yes) avoidance of necrotic and cystic areas during delineation
- Outcome: Definition of outcome value, follow-up times for the total cohort and each research group separately
- Results: The statistical test used, the number of events per outcome category, the mean values of each study group, if available the value of

thresholds, p-values, odds ratio (OR), and hazard ratio (HR).

Study results are presented in summarizing tables. The standardized mean difference (SMD, Cohen's *d*[15], with its 95% confidence intervals (95% CI) were calculated for each study (see formulas in Appendix A3), as it was possible to calculate from most of the available data, and visualized in a forest plot.

If data was unavailable and could not be recalculated from given data, corresponding authors were contacted and requested to provide additional data. In case of no response, study results were still presented in tables with all available information.

Due to high heterogeneity in outcome measures, data was clustered by outcome variables according to four categories: 1) Overall survival (OS), 2) Locoregional control (LRC), 3) Disease-free survival (DFS), and 4) Alternative outcomes (AO). All outcome variables which did not fit in the first three categories were categorized as AO (e.g. distant metastases).

RESULTS

Literature search

Our search was conducted to give a broad overview of all available MR-based parameters, resulting in 5497 original articles. After careful title and abstract selection, an extensive amount of 112 heterogeneous records still remained. Therefore, records were limited to only functional MRI studies (i.e. MR diffusion and MR perfusion). Articles describing anatomical characteristics (i.e. tumor volume, tissue invasion, depth of invasion) or machine learning approaches (i.e. radiomics) were excluded (n=54). During a full-text evaluation, the applicability of each study was assessed using the QUIPS tool. Studies marked as serious concern of applicability (i.e. due to short follow-up (<12 months)) were excluded to increase homogeneity between studies, resulting in a total of 31 records for quantitative analysis. Among them, 11 and 28 studies assessed DCE[16-26] and DWI[17-20,22-24,26-36,38,39,41-46] parameters respectively. Eight publications were included in both categories. The complete in- and exclusion process is shown in Figure 1. Although limiting the scope's focus, included studies still showed high heterogeneity in patient, imaging, and treatment characteristics, restricting reproducibility. As a result, a comprehensive meta-analysis could not be performed.

Quality assessment

The quality assessment results according to the QUIPS tool are shown in Figure 2 and Appendix A4. Overall, 14 out of 31 (45%) studies were marked with an overall

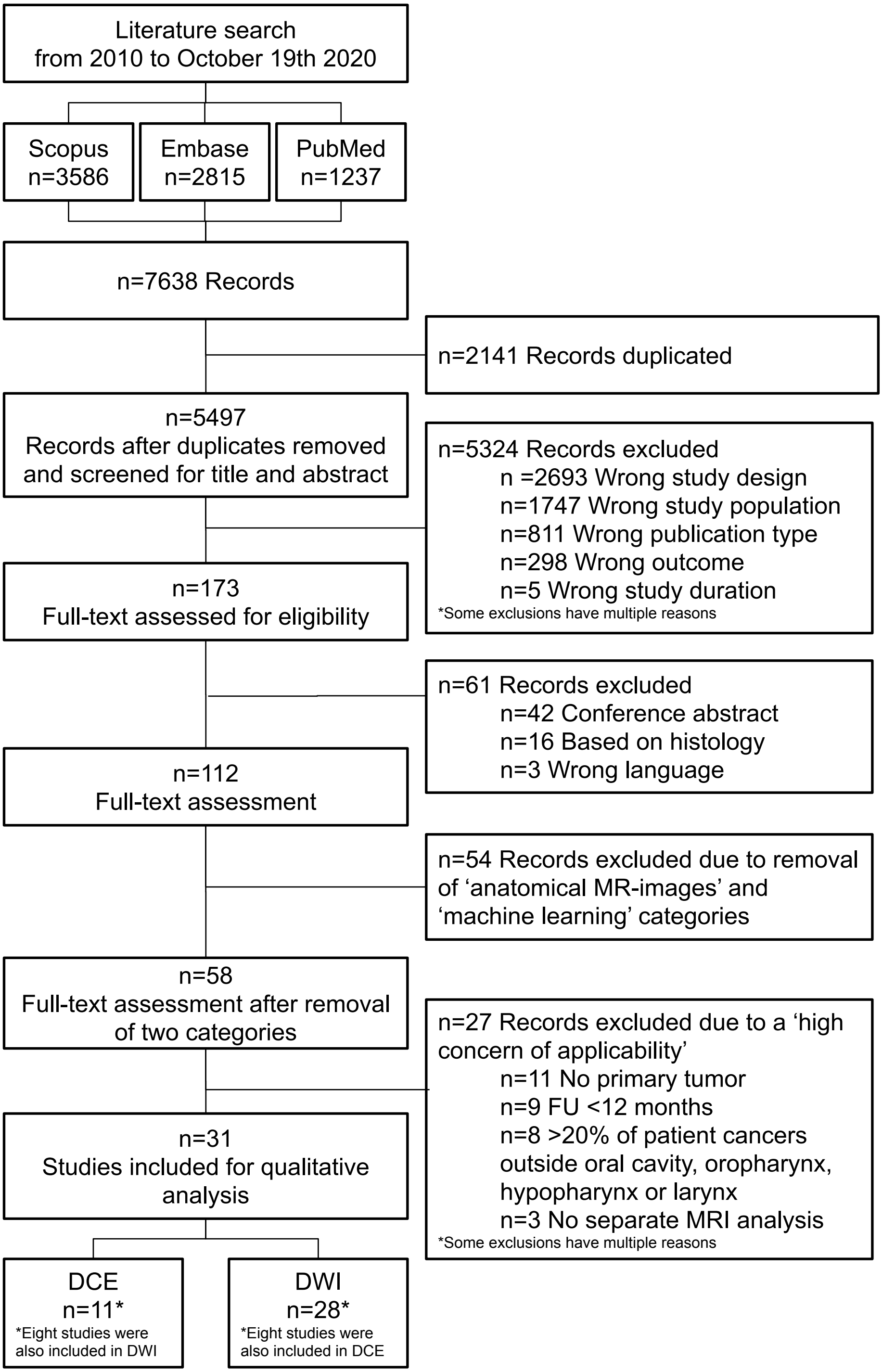

Literature search
from 2010 to October 19th 2020
Scopus n=3586
Embase n=2815
PubMed n=1237
n=7638 Records
n=2141 Records duplicated
n=5497
Records after duplicates removed and screened for title and abstract
n=5324 Records excluded
n =2693 Wrong study design
n=1747 Wrong study population
n=811 Wrong publication type
n=298 Wrong outcome
n=5 Wrong study duration
*Some exclusions have multiple reasons
n=173
Full-text assessed for eligibility
n=61 Records excluded
n=42 Conference abstract
n=16 Based on histology
n=3 Wrong language
n=112
Full-text assessment
n=54 Records excluded due to removal of 'anatomical MR-images' and 'machine learning' categories
n=58
Full-text assessment after removal of two categories
n=27 Records excluded due to a 'high concern of applicability'
n=11 No primary tumor
n=9 FU <12 months
n=8 >20% of patient cancers outside oral cavity, oropharynx, hypopharynx or larynx
n=3 No separate MRI analysis
*Some exclusions have multiple reasons
n=31
Studies included for qualitative analysis
DCE
n=11*
*Eight studies were also included in DWI
DWI
n=28*
*Eight studies were also included in DCE

Fig. 1. Flowchart of the inclusion process.

low risk of Bias (RoB), 5 (16%) studies as moderate, and 12 (39%) as high RoB. Notable areas of quality concerns included studies with patient inclusion with varying treatment approaches (Study attrition)[17,19,27,34,38,42], lacking a clear outcome definition (Outcome management)[21,22,24] a poor data presentation to assess the adequacy of the analysis (Statistical analysis and reporting)[16,28,33] or no mention and/or account for possible confounders (Study confounding)[21,25,28,39]. However, few studies applied subgroup analysis to outline the prognostic value of possible confounders, such as treatment[19], gender[34], HPV[36,41,43] or T-stage[30,34,41]. The majority of the studies show a low RoB in the domains of 'study participation' and 'prognostic factor measurement'.

Outcome prediction

DCE

Eleven DCE studies were assessed, comprising a total of 529 patients [range: 10–124], with an average age of 56.7 years, all treated with (chemo)radiation therapy. Of these, six studies were performed prospectively, and five studies had a retrospective design. Variations of the Tofts (n=5) or Kety (n=4) models were the most common models used for DCE imaging biomarkers. Imaging biomarkers were extracted from the total tumor volume in eight studies, including three studies that used the clinical available gross tumor volume (GTV) delineations. A detailed overview of the included studies is summarized in Table 1 (patient characteristics) and Appendix A5 (imaging characteristics).

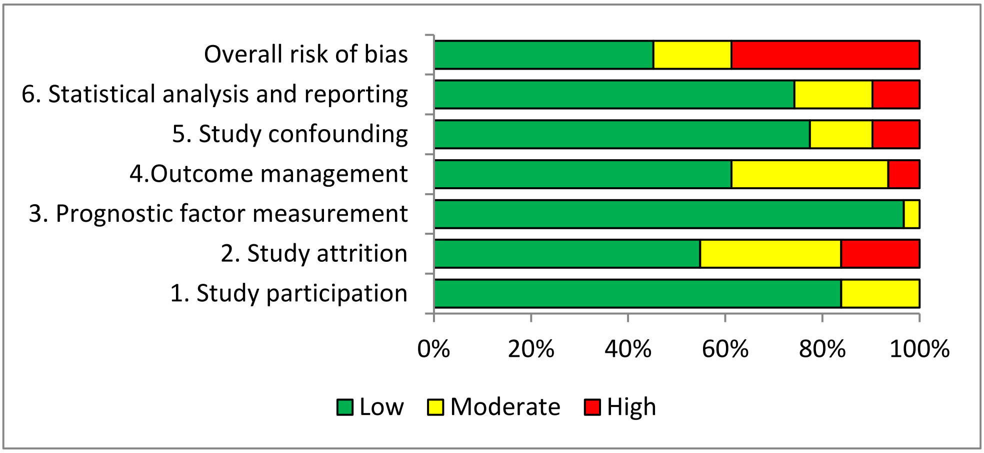


Fig. 2. Results of the QUIPS evaluation, visualizing the risk of bias for the six domains and the overall risk of bias

K^{trans}, V_p, $V_{e,}$ and K_{ep} were the most reported DCE parameters (K^{trans}, $V_{p,}$ and V_e in seven unique studies, K_{ep} in four unique studies). Mainly DFS[17-21,24] (n=6) was assessed, followed by OS[16,18,20,22,24] and LRC[17,21-23,25]. Wong et al.[26] compared DCE parameters between responders and non-responders. Table 2 gives a detailed overview of the different study results stratified per prognosticator and outcome. The following subsections explore this table in more detail. Detailed study results are summarized in Appendix A6. Forest plots for the parameters are visualized in Figure 3.

K^{trans}

For K^{trans} as a prognosticator, three studies[18,20,24] found a significantly higher K^{trans} (0.57 vs 0.22) in surviving patients compared to non-surviving patients (all $p<0.026$). Chan et al.[18] and Garbajs et al.[20] reported K^{trans} as an independent prognostic parameter in multivariate analysis ($p<0.001$, $p=0.026$ respectively). Furthermore, Baer et al.[16] found a statistical difference, but the linked value of K^{trans} was not available. The study of Martens et al.[22] could not substantiate a relationship between K^{trans} and longer OS.

For the prediction of LRC, the results of the two applicable included studies[22,23] were contradictory. Ng et al.[23] found a significantly higher K^{trans} in patients with local control (0.7±0.3 vs 0.5±0.3, $p=0.01$), whereas the study of Martens et al.[22] report lower K^{trans} values in patients with LRC (0.6±0.3 vs 0.74±0.3, $p=0.027$).

Higher K^{trans} was observed in patients with longer DFS[18-20,24]; however, this trend was only statistically significant in the studies of Chan et al.[18] ($p=0.003$) and Ng et al.[24] ($p=0.0096$). No significant difference of K^{trans} was found between responders and non-responders (domain: AO) by Wong et al.[26].

V_p

Only one study[18] described lower V_p as a significant, but not independent, predictor for longer OS and DFS, with $p=0.004$ and $p=0.001$, respectively (See Figure 3 and Table A6.2). The majority of the studies showed a possible trend where higher V_p was predictive for better DFS[18,19,24]. However, these studies were unable to substantiate this trend statistically. Wong et al.[26] reported a trend that responders had a higher V_p compared to non-responders (8.5±7.4 vs 2.7±5.6, $p=0.072$), where a lower trend was visible in the prediction of LRC by Ng. et al.[23]. Baer et al.[16] also found a near-significant difference ($p=0.068$) between V_p and OS, but results were uninterpretable due to limited information.

V_e

Study results of Martens et al.[22] described that a lower value of V_e was associated with better OS ($p=0.019$) and LRC ($p=0.015$). This is in contrast with the findings of Chan et al.[18] where higher V_e was prognostic for better survival. This higher trend,

Table 1. Baseline characteristics of perfusion studies.

Study, year	Location inclusion center	Study design	No. [N]	Inc. [N]	Age [mean]	Male [%]	Tumor subsite	Tumor stage (TNM)	Disease stage (AJCC)	Treatment	FU* [months]
Baer 2015[16]	Michigan, USA	R	24	10	58.7	80	OP, HP, OT	T1, T2, T3, T4	III - IVB	cCRT	26.2
Cao 2019[17]	Michigan, USA	R	54	54	61.0	87	OC, OP, HP, LA, OT	T1, T2, T3, T4	NA	cCRT	24.0‡
Chan 2015[18]	Taoyuan City, TW	R	149	124	52.0	94	OP, HP	T1, T2, T3, T4	III - IV	cCRT (IMRT)	28.7
Chawla 2013[19]	Pennsylvania, USA	R	32	24	57.8	81	OC, OP, LA	T1, T2, T3, T4	NA	cCRT or icCRT	23.7
Garbajs 2019[20]	Ljubljana, SI	P	20	20	58.3	95	OP, HP	T2, T3, T4	III - IVB	cCRT	27.2
Lowe 2018[21]	Manchester, UK	P	50	42	56.0†	90	OC, OP, HP, OT	T2, T3, T4	IV	cCRT (IMRT)	36.0‡
Martens 2021[22]	Amsterdam, NL	P	81	70	64.0†	69	OP, HP	T2, T3, T4	NA	c(C)RT	22.1
Ng 2013[23]	Taoyuan City, TW	P	78	58	48.5†	93	OP, HP	T1, T2, T3, T4	III - IVB	cCRT (IMRT)	19.2
Ng 2016[24]	Taoyuan City, TW	P	108	86	50.0	93	OP, HP	T1, T2, T3, T4	III - IVB	cCRT (IMRT)	28.0
Wang 2012[25]	Michigan, USA	P	14	14	56.9	86	OP, HP, LA, OT	T1, T2, T3, T4	III - IVB	cCRT	19.6‡
Wong 2018[26]	London, UK	P	35	27	61.0†	100	OP, HP, LA	T1, T2, T3, T4	III - IVB	cCRT	14.0

*For the entire cohort; †Median value; ‡Median for the non-event group.

Abbreviations: FU = Follow-up; P = Prospective; R = Retrospective; No. = Number of selected patients; Inc. = Number of analyzed patients; OC = Oral cavity; OP = Oropharynx; HP = Hypopharynx; LA = Larynx; OT = Other; cCRT = Concurrent chemoradiation therapy; icCRT = Induction chemoradiation therapy followed by concurrent chemoradiation therapy; IMRT = Intensity-modulated radiotherapy; RT = Radiation therapy; NA = Not Available.

Table 2. Overview of the study results sorted per prognosticator (perfusion and diffusion parameters) and outcome for the univariate and multivariate analysis. Each column recapitulates the number of studies describing a specific study result. The *Summary* column represents the overall summary of the studies. Overall summaries based on uniform results are marked with an asterisk (*).

DCE		**Univariate analyais**																						
		H	L	NS	NA	Summary		H	L	NS	NA	Summary		H	L	NS	NA	Summary		H	L	NS	NA	Summary
K^{trans}	OS	3	0	1	1	H	LRC	1	1	0	0	Δ	DFS	2	0	2	0	Δ	AO	0	0	1	0	NS*
V_p		0	1	3	0	NS		0	0	1	0	NS*		1	0	3	0	NS		0	0	1	0	NS*
V_e		1	1	2	0	Δ		0	1	1	0	Δ		1	0	3	0	NS		0	1	0	0	L*
K_{ep}		1	0	2	0	Δ		0	0	2	0	NS*		2	0	0	0	H*		-	-	-	-	-
		Multivariate analysis																						
		H	L	NS	NA	Summary		H	L	NS	NA	Summary		H	L	NS	NA	Summary		H	L	NS	NA	Summary
K^{trans}	OS	2	2	1	0	Δ	LRC	2	0	0	0	S*	DFS	1	1	2	0	Δ	AO	0	0	1	0	NT*
V_p		0	1	3	0	NT		0	0	1	0	NT*		0	1	3	0	NT		0	0	1	0	NT*
V_e		1	1	2	0	Δ		1	0	1	0	Δ		0	1	3	0	NT		0	0	1	0	NS*
K_{ep}		2	0	1	0	Δ		0	0	2	0	NT*		2	0	0	0	S*		-	-	-	-	-
ADC		**Univariate analysis**																						
		H	L	NS	NA	Summary		H	L	NS	NA	Summary		H	L	NS	NA	Summary		H	L	NS	NA	Summary
Mean	OS	1	4	4	0	Δ	LRC	0	1	7	0	NS	DFS	1	3	4	2	Δ	AO	-	-	-	-	-
Median		0	3	2	1	L		0	0	2	1	Δ		0	1	2	1	Δ		0	1	0	1	Δ*
		Multivariate analysis																						
		H	L	NS	NA	Summary		H	L	NS	NA	Summary		H	L	NS	NA	Summary		H	L	NS	NA	Summary
Mean	OS	0	3	6	0	NT	LRC	1	0	7	0	NT	DFS	0	3	6	1	NT	AO	-	-	-	-	-
Median		0	0	6	1	NT		0	0	2	1	NT		0	0	2	1	NT		0	0	1	1	NT

Abbreviations: OS = Overall survival; LRC = Locoregional control; DFS = Disease-free survival; AO = Alternative outcome; H = A significantly higher value was reported in the nonevent patient group compared to the event patient group, L = A significantly lower value was reported in the nonevent patient group compared to the event group; NS = No significant value was found between the nonevent and event patient group; NA = The study result were not available, despite attempts to contact the study authors; S = The significant prognostic parameter was also significant in multivariate analysis; NS = The significant prognostic parameter was not significant in multivariate analysis; NT = Multivariate analysis was not applied; Δ = Different study results were reported

PF	Outcome	Study	Patients [n]	Patients with non-event [%]	Patients with event [%]	p-value univariate	P-value multivariate	Standard mean difference [95% CI]
K^{trans}	OS	Baer 2015 [16]	10	50.0	50.0	0.002	NA	NA
		Chan 2015 [18]	124	62.1	37.9	0.025	NA	0.19 [0.09, 0.42]
		Garbajs 2019 [20]	20	75.0	25.0	0.026	0.026	-1.60 [-2.73, -0.48]
		Martens 2021 [22]	70	71.4	28.6	NS	NA	0.30 [-0.22, 0.82]
		Ng 2016 [24]	86	61.6	38.4	0.0026	NS	0.17 [0.07, 0.44]
	LRC	Martens 2021 [22]	70	57.7	24.3	0.027	NA	0.47 [-0.08, 1.02]
		Ng 2013 [23]	58	70.7	29.3	0.03	0.04	-0.67 [-1.24, -0.09]
	DFS	Chan 2015 [18]	124	75.8	24.2	0.003	NA	0.25 [0.12, 0.54]
		Chawla 2013 [19]	24	66.7	33.3	NS	NA	-0.65 [-1.51, 0.22]
		Garbajs 2019 [20]	20	75.0	25.0	0.116	NA	-1.60 [-2.73, -0.48]
		Ng 2016 [24]	86	43.0	57.0	0.0096	NS	0.19 [0.08, 0.47]
	AO	Wong 2018 [26]	35	77.1	22.9	NS	NA	-0.19 [-0.98, 0.60]
V_p	OS	Baer 2015 [16]	10	50.0	50.0	NS	NA	NA
		Chan 2015 [18]	124	62.1	37.9	0.004	NA	0.89 [0.34, 2.35]
		Garbajs 2019 [20]	20	75.0	25.0	NS	NA	0.92 [-0.13, 1.97]
		Ng 2016 [24]	86	61.6	38.4	NS	NA	0.39 [0.16, 0.96]
	LRC	Ng 2013 [23]	58	70.7	29.3	NS	NA	0.50 [-0.07, 1.07]
	DFS	Chan 2015 [18]	124	75.8	24.2	0.001	NA	0.51 [0.21, 1.27]
		Chawla 2013 [19]	24	66.7	33.3	NS	NA	-1.72 [-2.69, -0.74]
		Garbajs 2019 [20]	20	75.0	25.0	NS	NA	0.92 [-0.13, 1.97]
		Ng 2016 [24]	86	43.0	57.0	NS	NA	0.33 [0.13, 0.80]
	AO	Wong 2018 [26]	35	77.1	22.9	NS	NA	-0.82 [-1.63, -0.01]
V_e	OS	Chan 2015 [18]	124	62.1	37.9	0.038	NA	0.18 [0.06, 0.50]
		Garbajs 2019 [20]	20	75.0	25.0	NS	NA	-0.46 [-1.48, 0.56]
		Martens 2021 [22]	70	71.4	28.6	0.004	NA	0.74 [0.21, 1.28]
		Ng 2016 [24]	86	61.6	38.4	NS	NA	0.44 [0.18, 1.09]
	LRC	Martens 2021 [22]	70	75.7	24.3	0.008	NA	0.77 [0.21, 1.33]
		Ng 2013 [23]	58	70.7	29.3	NS	NA	-0.56 [-1.14, 0.01]
	DFS	Chan 2015 [18]	124	75.8	24.2	0.027	NS	0.20 [0.08, 0.55]
		Chawla 2013 [19]	24	66.7	33.3	NS	NA	1.34 [0.42, 2.27]
		Garbajs 2019 [20]	20	75.0	25.0	NS	NA	-0.46 [-1.48, 0.56]
		Ng 2016 [24]	86	43.0	57.0	NS	NA	0.32 [0.13, 0.79]
	AO	Wong 2018 [26]	35	77.1	22.9	0.003	NA	1.13 [0.30, 1.96]
K_{ep}	OS	Chan 2015 [18]	124	62.1	37.9	0.02	NA	3.78 [0.83, 17.16]
		Martens 2021 [22]	70	71.4	28.6	NS	NA	0.23 [-0.28, 0.75]
		Ng 2016 [24]	86	61.6	38.4	NS	NA	0.32 [0.12, 0.81]
	LRC	Martens 2021 [22]	70	75.7	24.3	NS	NA	0.22 [-0.32, 0.77]
		Ng 2013 [23]	58	70.7	29.3	NS	NA	-0.32 [-0.89, 0.25]
	DFS	Chan 2015 [18]	124	75.8	24.2	0.035	NA	0.24 [0.08, 0.71]
		Ng 2016 [24]	86	43.0	57.0	0.0474	NA	0.24 [0.09, 0.61]

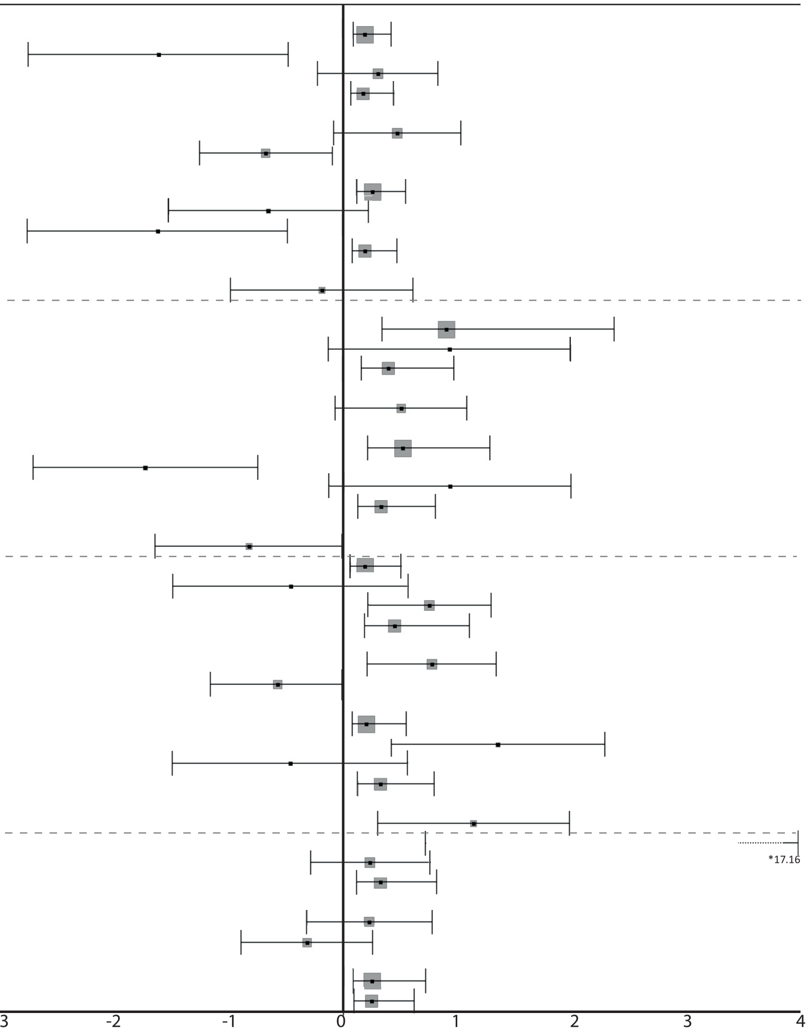


Fig. 3. Forest plot of perfusion prognosticators for the different outcome domains. OS = Overall survival; LRC = Locoregional control; DFS = Disease-Free Survival; AO = Alternative outcomes; PF = Prognostic Factor; NA = Not available.

however not significantly, was also visible in three other studies describing the relation between V_e and OS[20,24] and LRC[23] (Table A6.3). Three[18,20,24] out of four studies[18-20,24] found a relation of higher V_e and longer DFS, where only Chan et al.[18] found a significant distinction. Wong et al.[26] could substantiate that a lower V_e was prognostic for good treatment response (0.26±0.06 vs 0.32±0.06, p=0.003).

K_{ep}

Three studies[18,22,24] described the prognostic value of K_{ep} related to OS prediction. Of these, two studies[18,24] showed a correlation between a higher K_{ep} and better OS, but only Chan et al.[18] was significant (p=0.02). Concerning LRC, none of the studies[22,23] found statistical differences between the control and failure groups. Based on the results of Chan et *al.*[18] and Ng et al.[24] higher K_{ep} is an independent predictor for longer DFS (p=0.005 and p=0.001, respectively).

Other DCE parameters

Besides the above-mentioned DCE parameters, other parameters such as blood volume (BV), blood flow (BF), plasma perfusion (F_p), and the normalized area under the contrast-enhancement time curve at 60s (NAUC60) were separately analyzed. The results of these parameters are described in Appendix A7.

DWI

Detailed information of baseline characteristics with regard to DWI studies is summarized in Table 3, imaging characteristics are summarized in Appendix A8. A total of 1626 patients, with an average age of 60.0 years, encompassed the participants in the 28 included DWI studies. The majority of the patients received CRT (55%), followed by a treatment existing of only radiotherapy (25%) and surgery combined with CRT (6%). Sixteen studies enrolled patients retrospectively. DWI was acquired using an echo-planar imaging (EPI) in 89% of the studies. Other studies used periodically rotated overlapping parallel lines with enhanced reconstruction (PROPELLOR) (7%). In the study of Ravanelli et al.[43] the type of MR imaging sequence was unknown. In seven studies (25%), fat suppression was added during image acquisition. All studies used multiple b-values, with a median of two b-values (range 2–17). B-values were applied in a range of 0 to 1000 s/mm^2 (n=16)[19,20,22,27,29-31,33,34,36,39-41,43,44,46], a range of 0–800 s/mm^2 (n=7)[18,23,24,37,38,41,45] or an alternative lowest and/or highest b-value (n=7)[17,26,28,32,35,39,42].

ADC_{mean} and ADC_{median} were mostly reported in, respectively 20 and 9 unique studies. Out of all studies, DFS was reported in 15 unique studies[18-20,24,27,28,31,33,35,38-40,42,43,45], OS in 13[18,20,22,24,28,31,33-36,43,44,46], LRC in 12[17,22,23,28-33,36,37,41] and 2 studies were categorized as AO with distant metastases[33] and responders versus non-responders[26] as outcome parameter.

The following subsections summarize study findings per prognostic factor for each separate outcome domain. An overview of these study findings is described in Table 2, Figure 4, and Appendix A9.

ADC_{mean}

In four[18,22,34,44] out of the nine[18,22,24,28,34,36,43,44,46] studies that reported the prognostic value of the ADC_{mean} for OS, a significantly lower ADC_{mean} was described in survivors (all $p<0.045$). None of these studies could substantiate a significant difference in their multivariate analysis. In contrast to these results, the study of Zhang et al.[46] reported an average higher ADC_{mean} in survivors compared to non-survivors ($p=0.02$). Four studies[24,28,36,43] did not report a statistically significant difference between the groups; among them, three[24,36,43] showed a minor trend where lower ADC_{mean} was associated with better OS (Table A9.1). Results of Gupta et al.[28] were uninterpretable due to limited available information.

For the prediction of LRC, only one study[30] found that patients with local control had a significant lower ADC_{mean} in the entire cohort as well as in a sub-analysis for patients with stage T3 and T4 disease compared to patients with local failure (entire cohort: 0.74 ± 0.03 vs 1.02 ± 0.08, $p<0.001$, stage T3 and T4: 0.83 ± 0.14 vs 0.95 ± 0.04, $p=0.02$). Lower ADC_{mean} in patients with better LRC is also noted, without statistical significance (Table A9.1), in five other studies[17,23,32,36,37], where the study of Martens et al.[36] reported a nearly significant difference (12.08 ± 2.33 vs 13.17 ± 3.26, $p=0.055$).

Two studies[22,28] did not find a difference between patients with LRC and locoregional failure.

Studies investigating ADC_{mean} in relationship to DFS show high diversity in their results. Five studies[24,38,42,43,45] described a statistical relationship, among them three[24,42,45] described a lower and one higher[38] ADC_{mean} predictive of better DFS (Figure 4). Lastly, one study[43] showed a significant difference ($p=0.03$) with a non-identified value of ADC_{mean}. None of the studies show significance in multivariate analysis. The remaining studies[18,39,40] who did not report a statistical difference show a trend of lower ADC_{mean} predictive of good DFS (0.85 ± 0.27 vs 1.59 ± 0.39). Interpretation of study results from two studies[27,28] was limited due to missing values.

ADC$_{median}$

Three studies[20,34,44] described a significant value of a lower ADC$_{median}$ in surviving patients. Among them, Ren et al.[34] reported this result when the ADC was measured over the whole tumor volume and only a single slice was delineated, with p=0.016 and p=0.033, respectively. The study of Martens et al.[36] and Lu et al.[35] reported also a lower ADC$_{median}$ in survivors; however, the difference was not statistically significant (p=0.217, p=0.223). Findings concerning ADC$_{median}$ and its relationship with OS in one other studie[33] was unclear.

A trend was visible of lower ADC$_{median}$ related to longer LRC[36,41] and DFS[19,20,35], where the study of Martens et al.[36] reached nearly significance (p=0.06) in the prediction of LRC. Patients with DFS had a significant lower ADC$_{median}$ compared to patients with failure described by Lu et al.[35]. Wong et al.[26] reported a significantly lower ADC$_{median}$ in patients with good treatment response than patients who did not respond to treatment (1.02±0.19 versus 1.22±0.14, p=0.009). The study of Lambrecht et al.[33] report results inadequately to draw reliable conclusions in the prediction of LRC, DFS, and distant metastases.

Other DWI parameters

Besides the ADC$_{mean}$ and ADC$_{median}$, the parameters ADC$_{kurtosis}$, ADC$_{skewness}$, ADC$_{entropy}$, ADC$_{Standard\ Deviation}$ (SD), ADC$_{min}$, ADC$_{max}$, ADC$_{0\text{-}200}$ (ADC calculated using b-values ranging from 0 to 200 s/mm^2), ADC$_{300\text{-}1000}$ (ADC calculated using b-values ranging from 300 to 1000 s/mm^2), and a broad range of ADC$_{deciles}$ and ADC$_{percentiles}$ were reported in the included studies. The results of these parameters are described in Appendix A10.

IVIM

The prognostic value of IVIM parameters was reported by the studies of Martens et al.[22] and Lu et al.[34], totally including 86 patients. Most patients received CRT (n=78), followed by radiotherapy alone (n=6) and surgery (n=2). In both studies, ADC was measured over the whole tumor volume. Baseline and imaging characteristics of these studies can be found in Table 3 and Appendix A11.

Associations between the IVIM parameters *D*, *D*,* and *f* were reported for OS[22,35], LRC[22], and DFS[35]. Significant findings were only reported in the prediction of OS[22], where a lower *D* and higher *D** were noted in surviving patients compared to non-survivors (*D:* 0.9±0.2 versus 1.0±0.2, p=0.009, *D**: 0.19±0.07 versus 0.16±0.05, p=0.032). A trend of lower and higher *f* was reported for, respectively, OS and LRC. All other study findings did not reach statistical significance.

Table 3. Baseline characteristics of diffusion studies. Studies describing also perfusion parameters are marked with a section symbol([§]).

Study, year	Location inclusion center	Study design	No. [N]	Inc. [N]	Age [mean]	Male [%]	Tumor subsite	Tumor stage (TNM)	Disease stage (AJCC)	Treatment	DWI/ IVIM	FU* [months]
Cao 2019[17§]	Michigan, USA	R	54	54	61.0	87	OC, OP, HP, LA, OT	T1, T2, T3, T4	NA	cCRT	DWI	24.0‡
Chan 2015[18§]	Taoyuan City, TW	R	149	124	52.0	94	OP, HP	T1, T2, T3, T4	III - IV	cCRT (IMRT)	DWI	28.7
Chawla 2013[19§]	Pennsylvania, USA	R	32	24	57.8	81	OC, OP, LA	T1, T2, T3, T4	NA	cCRT or icCRT	DWI	23.7
Choi 2017[27]	Suwon, KR	R	59	44	61.0	82	OC, OP, HP, LA, OT	T1, T2, T3, T4	NA	Surgery + RT or c(C)RT	DWI	14.0
Garbajs 2019[20§]	Ljubljana, SI	R	20	20	58.3	95	OP, HP	T2, T3, T4	III - IVB	cCRT	DWI	27.2
Gupta 2019[28]	Mumbai, IN	P	20	16	58.0	85	OH, HP, LA	T2, T3	NA	(C)RT	DWI	44.0
Hatakenaka 2011[29]	Fukuoka, JP	P	32	17	64.0†	88	OC, OP, HP, LA	T1, T2, T3, T4	NA	RT	DWI	23.6‡
Hatakenaka 2011[30]	Fukuoka, JP	R	48	38	64.0†	92	OC, OP, HP, LA	NA	NA	c(C)RT	DWI	14.3‡
Hatakenaka 2014[31]	Fukuoka, JP	R	62	41	64.0†	93	OC, OP, HP, LA	T1, T2, T3, T4	I - IVB	c(C)RT	DWI	>35.0‡
King 2013[32]	Shatin, HK	P	56	37	57.0	86	OC, OP, HP, LA, OT	T1, T2, T3, T4	NA	c(C)RT	DWI	43.9‡
Lambrecht 2014[33]	Leuven, BE	P	175	161	NA	86	OC, OP, HP, LA	T1, T2, T3, T4	I - IVB	RT	DWI	>36.0‡
Li 2018[34]	Shanghai, CH	R	135	96	57.7	68	OC, OP	NA	I - IVB	Surgery ± c(C)RT	DWI	29.0
Lu 2013[35]	New York, USA	R	16	16	55.0	94	OC, OP	NA	III - IV	Surgery or c(C)RT	IVIM	20.8
Martens 2019[36]	Amsterdam, NL	R	134	89	66.4	72	OC, OP, HP, LA	T1, T2, T3, T4	II - IV	c(C)RT	DWI	25.6
Martens 2021[22§]	Amsterdam, NL	P	81	70	64.0†	69	OP, HP	T2, T3, T4	NA	c(C)RT	DWI + IVIM	22.1

Table 3 continued.

Matoba 2014[37]	Ishikawa, JP	P	40	35	66.2†	86	OC, OP, HP, LA	T1, T2, T3, T4	NA	cCRT	DWI	30.8
Nakajo 2012[38]	Kagoshima, JP	R	28	26	65.0	85	OC, OP, HP, LA, OT	T1, T2, T3, T4	NA	Surgery ± nCT or c(C)RT	DWI	16.0
Ng 2013[23§]	Taoyuan City, TW	P	78	58	48.5†	93	OP, HP	T1, T2, T3, T4	III - IVB	cCRT (IMRT)	DWI	19.2
Ng 2016[24§]	Taoyuan City, TW	P	108	86	50.0	93	OP, HP	T1, T2, T3, T4	III - IVB	cCRT (IMRT)	DWI	28.0
Noij 2015[39]	Amsterdam, NL	R	111	78	62.0	58	OC, OP, HP, LA	T1, T2, T3, T4	I - IVB	c(C)RT	DWI	18.0
Núnez 2017[40]	Vigo, SP	P	6	6	65.0	83	OP	NA	NA	cCRT (IMRT)	DWI	12.0
Peltenburg 2020[41]	Utrecht, NL	R	295	217	63.0†	75	OC, OP, HP, LA	T1, T2, T3, T4	II - IVB	c(C)RT	DWI	34.0
Preda 2016[42]	Milan, IT	R	74	57	55.0	NA	OC, OP	T1, T2, T4	NA	Surgery ± c(C)RT	DWI	21.3
Ravanelli 2020[43]	Bréscia, IT	R	68	59	66.3	73	OP	T2, T3, T4	NA	cCRT (IMRT)	DWI	>24.0
Ren 2018[44]	Shanghai, CH	R	73	73	57.9	66	OC, OP	T1, T2, T3, T4	III - IVB	Surgery ± c(C) RT	DWI	25.0
Srinivasan 2012[45]	Michigan, USA	P	24	17	57.2	94	OP, HP, LA, OT	T1, T2, T3, T4	NA	cCRT (IMRT)	DWI	>24.0
Wong 2018[26§]	London, UK	P	35	27	61.0†	100	OP, HP, LA	T1, T2, T3, T4	III - IVB	cCRT	DWI	14.0
Zhang 2018[46]	Zhejiang, CH	R	63	40	62.1	98	HP	T2, T3, T4	NA	cCRT ± Surgery	DWI	12.9

*For the entire cohort; †Median value; ‡Median for the non-event group.

Abbreviations: FU = Follow-up; P = Prospective; R = Retrospective; No. = Number of selected patients; Inc. = Number of analyzed patients; OC = Oral cavity; OP = Oropharynx; HP = Hypopharynx; LA = Larynx; OT = Other; cCRT = Concurrent chemoradiation therapy; icCRT = Induction chemoradiation therapy followed by concurrent chemoradiation therapy; IMRT = Intensity-modulated radiotherapy; RT = Radiation therapy; c(C)RT = Radiation therapy with/without concurrent chemotherapy; nCT = Neoadjuvant chemotherapy; NA = Not Available.

PF	Outcome	Study	Patients [n]	Patients with non-event [%]	Patients with event [%]	p-value univariate	P-value multivariate	Standard mean difference [95% CI]
ADC mean	OS	Chan 2015 [17]	124	73.3	26.7	0.045	NS	0.60 [0.23, 1.55]
		Li 2018 [33]	96	71.4	28.6	<0.6	NA	NA
		Martens 2019 [35]	89	69.7	30.3	NS	NS	NA
		Martens 2021 [21]	70	55.1	44.9	0.024	NA	0.50 [0.08, 0.92]
		Ng 2016 [23]	86	NA	NA	NS	NA	0.09 [-0.52, 0.70]
		Ravenelli 2020 [42]	59	60.0	40.0	NS	NA	NA
		Ren 2018 [43] (Whole tumor)	73	61.5	38.5	NA	NA	NA
		Ren 2018 [43] (Largest slice)	73	61.5	38.5	NA	NA	NA
		Zhang 2018 [45]	40	61.6	38.4	0.02	NA	NA
	LRC	Cao 2019 [16]	54	70.7	29.3	NS	NA	NA
		Hatakenaka 2011 [29] (full cohort)	38	73.2	26.8	0.005	NA	5.87 [4.37, 7.37]
		Hatakenaka 2011 [29] (stage 3/4)	17	NA	NA	0.01	0.02	5.64 [4.19, 7.10]
		King 2013 [31]	37	52.9	47.1	NS	NA	0.38 [-0.28, 1.03]
		Martens 2019 [35]	89	69.7	30.3	NS	NA	0.41 [-0.04, 0.87]
		Martens 2021 [21]	70	75.0	25.0	NS	NA	0.00 [-0.54, 0.55]
		Matoba 2014 [35]	35	62.7	37.3	NS	NA	0.20 [-0.47, 0.88]
		Ng 2013 [22]	58	70.7	29.3	NS	NA	0.00 [-0.56, 0.57]
	DFS	Chan 2015 [17]	124	NA	NA	NS	NA	0.72 [0.09, 1.36]
		Choi 2017 [26]	44	52.9	47.1	NA	NA	NA
		Nakajo 2012 [37]	26	62.1	37.9	0.009	NA	-1.49 [-2.76, -0.22]
		Ng 2016 [23]	86	70.7	29.3	0.0162	NS	1.13 [0.27, 1.98]
		Noij 2015 [38] (b0-b750)	78	75.8	24.2	NS	NA	NA
		Noij 2015 [38] (b0-b1000)	78	75.8	24.2	NS	NA	-0.59 [-1.28, 0.1]
		Núñez 2017 [39]	6	79.6	20.4	NS	NA	1.88 [-0.50, 4.28]
		Preda 2016 [41]	57	66.7	33.3	NA	NA	NA
		Ravenelli 2020 [42]	59	69.2	23.1	NA	NA	NA
		Srinivasan 2012 [44]	17	73.3	26.7	0.03	NA	1.19 [0.16, 2.22]
ADC meiaan	OS	Garbajs 2019 [19]	20	83.3	16.7	NS	NA	0.92 [-0.13, 1.97]
		Lambrecht 2014 [32]	161	75.0	25.0	NA	NA	NA
		Li 2018 [33]	96	71.4	28.6	<0.1	NA	NA
		Lu 2013 [34]	16	75.7	24.3	NA	NA	NA
		Martens 2019 [35]	89	69.7	30.0	NS	NA	0.32 [-0.10, 0.74]
		Ren 2018 [43] (Whole tumor)	73	61.5	38.5	NA	NA	NA
		Ren 2018 [43] (Largest slice)	73	61.5	38.5	NA	NA	NA
	LRC	Lambrecht 2014 [32]	161	60.9	39.1	NA	NA	NA
		Martens 2019 [35]	89	80.1	19.9	NS	NA	0.38 [-0.07, 0.84]
		Peltenburg 2020 [40]	217	56.8	43.2	NS	NA	0.04 [-0.25, 0.35]
	DFS	Chawla 2013 [18]	24	82.5	16.7	NS	NA	10.5 [7.43, 13.59]
		Garbajs 2019 [19]	20	83.3	16.7	NS	NA	0.92 [-0.13, 1.97]
		Lambrecht 2014 [32]	161	75.0	25.0	NA	NA	NA
	AO	Lambrecht 2014 [32]	161	75.0	25.0	NA	NA	NA
		Wong 2018 [25]	27	61.4	42.1	0.009	NA	1.11 [0.28, 1.94]

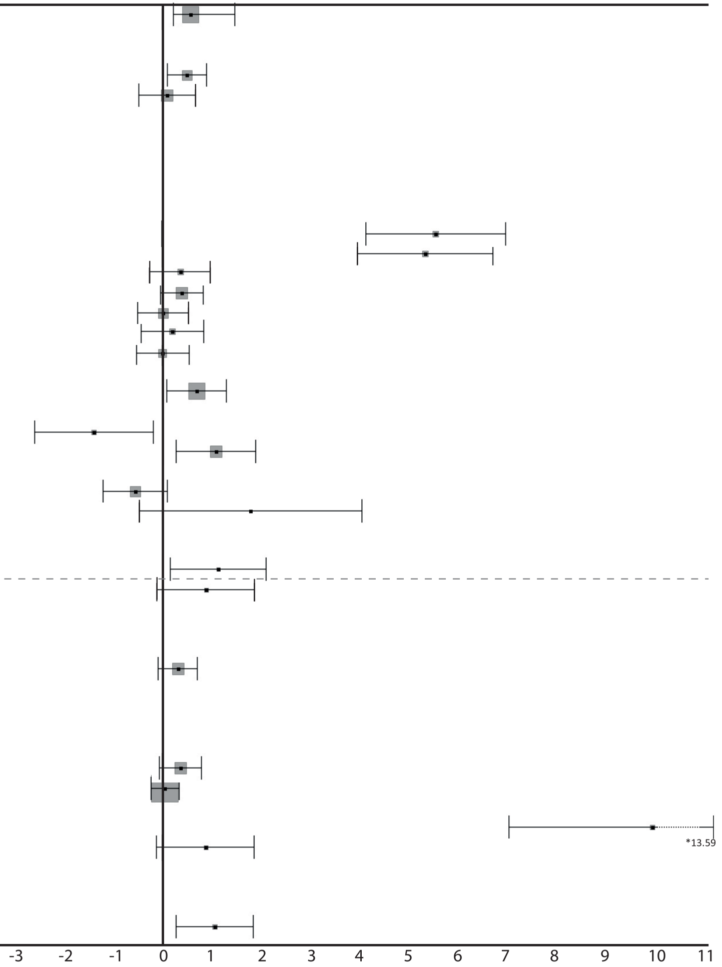


Fig. 4. Forest plot of mean and median ADC for the different outcome domains. OS = Overall survival; LRC = Locoregional control; DFS = Disease-Free Survival; AO = Alternative outcomes; PF = Prognostic Factor; NA = Not available.

DISCUSSION

This study attempts to systematically review the literature focusing on the prognostic value of pre-treatment functional MR imaging parameters extracted from the primary tumor volume in HNSCC patients of perfusion and diffusion MR scans. Overall, studies describing diffusion prognosticators were conducted more frequently compared to perfusion studies. Nevertheless, our study shows that the perfusion parameters K^{trans} and K_{ep} are promising independent prognostic factors for, respectively, OS and DFS, whereas only a trend of lower mean and median ADC was reported in survivors.

An increased blood vessel permeability optimizes tumor perfusion and contributes to the delivery of contrast agents as to the penetration of therapeutically used drugs. Additionally, better vessel permeability facilitates cell oxygenation, thus conceivably pertaining radio-chemo sensitivity. This association explains the favorability of higher influx to the extracellular extravascular space (K^{trans}) and reflux to the blood vessel (K_{ep}) in responding patients. In contrast, restricted perfusion is a consequence of the biological behavior of cancer characterized by its increasing neoangiogenesis and proliferation. Tumor cells become hypoxic and necrotic resulting in a disability of drug absorption, with a low permeability as a consequence. In line with that, Cao et al.[17] showed an association between poorly perfused and hypoxic tumors and treatment failure.

Besides tumor perfusion, tumor diffusion has shown its potential as prognostic parameter[18,20,22,24,30,34,35,38,42-46]. Tissue cellularity, the basic parameter determining diffusion, is high when diffusion is restricted, represented by a low ADC. In contrast, heterogeneous tumors (including necrotic and cystic areas) are characterized by high ADC. While not convincing, most studies[18,20,22,24,30,34,35,42,44,45] report a significant correlation between low pre-treatment ADC and good treatment response, whereas fewer studies show a significantly higher ADC[26,38] in responding patients or a lack of correlation.

The discrepancy between the results of diffusion studies was considerable. As mentioned above, tumor diffusion depends mainly on cellular density, but is also influenced by tissue heterogeneity as well as MR methodology. Substantial amount of heterogeneity was present in treatment type of included patients in studies describing diffusion parameters, where perfusion studies included only patients treated with radiotherapy.

Intratumor heterogeneity caused by cystic and necrotic tumor areas can likewise play a role in the inconsistency of study results. Due to the rapid proliferation, tumors can develop necrotic and cystic tissue areas which are highly correlated with a high ADC[7]. Inclusion of those areas in the region of interest during the delineation process will influence study results. Furthermore, as lower K^{trans} and K_{ep} values can be linked to limited tissue permeability, they could feasibly also be associated with the extent of necrotic areas within a tumor and, consequently, portraying similar biological tumor characteristics. Therefore, ADC may be more independent and accurate if necrotic and cystic areas are avoided in tumor delineation, focusing only on tumor tissue cellularity.

Fifteen[17-20,23,26,29-31,35,37-39,44,46] out of the 28 included DWI articles mentioned a conscious effort to avoid these areas in delineation. Three[36,41,45] articles, however, did explicitly choose to include those areas. The remaining ten[22,24,27,28,32-34,40,42,43] articles did not specify their delineation limitations. However, a significant difference in mean ADC for any outcome was observed in five[18,30,38,44,46] of the nine (56%) articles that did avoid cystic and necrotic tumor areas in delineation and, similarly, in one[45] of the two[36,45] (50%) that did not avoid these areas. However, a more striking difference in significance is observed for the median ADC, where all four[20,26,35,44] articles that avoided the cystic and necrotic tumor areas did produce a significant difference in treatment outcome prognostication, in contrast to none of the two[36,41] articles that did not avoid these areas. Consequently, the avoidance of cystic and necrotic areas appears to influence the significance of the median ADC for the prediction of treatment outcome after more than one year. Furthermore, a slight trend to more significant results can be observed for mean ADC when cystic and necrotic tumor areas are not delineated.

In line with previous remarks, tumor perfusion and diffusion are two linked biological processes. While random motion can be observed for individual water molecules (pure diffusion), water molecules collectively flow within the blood circulation (perfusion). Therefore, measured diffusion also depends on micro-vascularization, resulting in an overestimated ADC (this pseudo diffusion process is marked as the “apparent” in ADC)[5,6,47]. Higher minimum b-values (>100–150 s/mm^2) are recommended to eliminate this perfusion bias. Pure diffusion can also be measured using IVIM, a MR-diffusion-technique based on bi-exponential formulas, which discriminates perfusion from diffusion[5]. However, investigation of this technique for treatment prognosis is only performed in a very limited amount of studies[11].

Besides perfusion bias, ADC values are also overestimated as a result of noise. ADC values are built up from signal intensities that cannot be negative due to signal noise. Additionally, negative ADC values might occur caused by misregistration between the signal intensities of the different diffusion gradients, but these values are neglected and interpolated as zero. To limit these effects, ADC calculation based on the mean signal intensities of the region of interest is recommended, instead of a voxel-wise approach[48]. None of the included studies in this review reported on this recommended methodology.

In consonance with previous paragraphs, for a reliable ADC measurement, multiple diffusion gradients (b-values) are required. Measured ADC values depend on the used diffusion gradient due to the non-linear relation between b-values and signal intensity. This effect is supported by the findings of Noij et al.[39], where higher ADC values were found when measured using b-value 750 compared to b-value of 1000 s/mm^2. To date, a consensus of the optimal amount and combination of diffusion gradients is still pending, resulting in the high diversity of applied b-values in this study field which might be an explanation for the heterogeneity in study findings.

Finally, a lack of standardization in study methodology might cause a discrepancy between study findings. This can be confirmed by the findings of the study quality assessment, where relevant methodological steps were dismissed to obtain good reproducibility of study results. Additionally, variety in MRI vendor and acquisition protocols might affect diffusion and perfusion values. While recent studies recommend echo-planar-imaging DWI with six diffusion gradients[49,50] and spoiled gradient-echo acquisition DCE imaging[50], this is still the first step towards standardization. The impact on study results affected by each individual acquisition, methodology parameter, and statistical analysis has yet to be investigated[49-51].

Contradictory to the diffusion study results, findings in perfusion studies were more consistent. However, a discrepancy was found for K^{trans} as a prognosticator for LRC. Ng et al.[23] found a significantly higher K_{trans} in patients with LRC, whereas the study of Martens et al.[22] report lower K_{trans} prognostic for LRC. The statistical threshold, based on median values for their own data, used in the study of Ng et al.[23] might be an explanation of this inconsistency.

An extensive search was applied over the last decade. However, for some prognosticators, a limited amount of studies was available. Conclusions based on those individual studies were not yet representative but might be of potential value. For example, our findings show that NAUC60[16] and the ADC deciles (10th, 70th, and 80th) and percentiles (25th) are proven to be prognostic for OS[44] and LRC[36], showing the urgency for more research to better investigate their usefulness.

Unfortunately, the heterogeneity of studies led to an inability to draw firm conclusions in this review. To minimalize heterogeneity, clusters were created for several outcome groups. While these clusters categorize relevant treatment outcomes, each outcome might affect the prognosticator differently. Evaluating each outcome parameter individually is recommended for future research.

Moreover, non-uniformity was also visible in other categories, such as the inclusion of participants, image acquisition, delineation approach, or methodology to calculate perfusion or diffusion values. This review was limited to oral cavity, oropharyngeal, hypopharyngeal and laryngeal SCC cancer concerning patient inclusion. However, no constraints with regard to treatment approach or tumoral HPV status were applied. Regarding HPV tumor status, our study showed comparable study results between the study of Cao et al.[25], only including HPV negative tumors and other research findings[23,32,36,37] investigating the association between mean ADC and LRC (no statistical difference, but a trend was found). Concerning treatment approach, some studies[27,34,35,38,42,44,46] included patients who were (primarily) treated by surgery, among them two studies[38,42] describing significant relationships between mean ADC and DFS. These significant relationships were also reported in three other studies[24,43,45], only including patients treated with CRT. Two[24,45] of these three studies reported an association between lower mean ADC and longer DFS, similar to the study findings of Preda et al.[42]. Contradictive results were reported by Nakajo et al.[38] (higher mean ADC prognostic for better DFS), while this study includes a lower percentage of patients receiving surgery (81% vs 54%).

Non-uniformity was also present in the definition of the delineated primary tumor. Whole tumor volume delineation will be more reliable compared to delineating a single slice of the tumor area, since a larger tumor volume is included. Additionally, the inclusion of healthy tissue regions (i.e. during GTV delineation) will attenuate ADC values of the tumor tissue and restrict reflection of true tumor diffusion. Nevertheless, this review has not studied these effect on research results.

Another disadvantage is the remark which must be made to the results related to OS. Although OS is a very reliable outcome measure, it does not solely depend on cancer death, but take all causes of death into account. Moreover, even after requesting study authors, not all results were available, resulting in the lack of a complete overview of all included studies. Conclusions are based on the available information. Another drawback of this review is the initial scope of the review was limited to only functional imaging parameters as a result of the extensive amount of records. Studies describing other approaches to predict treatment outcome (i.e. anatomical imaging parameters, radiomics or artificial intelligence) were excluded, while especially radiomics and artificial intelligence are currently highly relevant and promising[52]. Therefore it is recommended to summarize those studies in a future systematic review. A final limitation is based on the inclusion of only pre-treatment parameters in this review. Recently, upcoming studies describe prognosticators (early) during treatment[16,17,20,21,26,32,37]. These intra-treatment prognosticators can be promising for outcome prediction but were not analyzed in this review.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

This study gives an overview of the current state of literature describing the prognostic value of MR-based pre-treatment perfusion and diffusion parameters in predicting treatment response in HNSCC. The accurate and consistent results of pre-treatment MR-based perfusion parameters K^{trans} and K_{ep} are promising for the clinical applicability of these parameters to predict survival and guide treatment decision. The variable study results for parameters extracted from diffusion imaging was mainly caused by heterogeneity in study design, image acquisition, MR field strength, segmentation approach, or statistical approach.

To reduce discrepancy between studies, a consensus on imaging acquisition and study methodology is required. Based on this review, several recommendations for future research can be formulated.

1. DWI image acquisition has to be performed with multiple gradient weights. Those gradient weights contain at least a low b-value (100–250 s/mm2) and a high b-value (>250 s/mm2). ADC depends on a non-linear relationship between gradient weights and signal intensities. Therefore, the inclusion of multiple gradient weights covering the whole spectra of the ADC curve is most representative of diffusion. Selecting a minimum b-value of 100 s/mm2 excludes perfusion bias caused by micro-vascularization.

2. Diffusion and perfusion values have to be calculated from the mean signal intensities of the region of interest instead of a voxel-wise approach. In a voxel-wise approach, negative ADC values are not present as a result of noise or misregistration. Therefore, pure perfusion or diffusion values are overestimated, where this is not the case in a mean signal intensity approach.

3. Delineate the whole tumor volume instead of a single slice delineation. Tumors are characterized by their heterogeneous texture. Single slice delineations might include homogeneous tumor characteristics (i.e. necrotic area (recommended

to exclude)), which do not reflect all tumor characteristics aspects. Therefore, the more representative whole tumor delineation is recommended.

4. Necrotic and cystic tumor regions have to be avoided in tumor delineation. Perfusion or diffusion is restricted in those areas, resulting in extreme vales in measured perfusion or diffusion values. Additionally, our analysis showed a slight trend of more significant study results in studies when cystic or necrotic tissues were removed. Therefore, preventing those areas will give a more precise representation of the diffusion and perfusion value.

5. Statistically thresholds have to be independent of the data. Nowadays, a large amount of research findings depends on thresholds calculated from their own data (i.e. median). To generate comparable and generalizable findings, absolute statistically cut-off values are suggested.

SUPPLEMENTARY INFORMATION

Password: PhD_PaulaBos

REFERENCES

1. Fitzmaurice C, Abate D, Abbasi N, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol*. 2021:5(12):1749–1768. doi:10.1001/jamaoncol.2019.2996
2. Gregoire V, Lefebvre JL, Licitra L, Felip E. Squamous cell carcinoma of the head and neck: EHNS-ESMO-ESTRO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2010;21:184–186. doi:10.1093/annonc/mdq185
3. Pignon JP, le Maître A, Maillard E, Bourhis J. Meta-analysis of chemotherapy in head and neck cancer (MACH-NC): An update on 93 randomised trials and 17,346 patients. *Radiother Oncol*. 2009;92(1):4-14. doi:10.1016/j.radonc.2009.04.014
4. Sourbron SP, Buckley DL. On the scope and interpretation of the Tofts models for DCE-MRI. *Magn Reson Med*. 2011;66(3):735–745. doi:10.1002/ mrm.22861
5. Le Bihan D. What can we see with IVIM MRI? *Neuroimage*. 2019;187:56–67. doi:10.1016/j.neuroimage.2017.12.062
6. Padhani AR, Liu G, Mu-Koh D, et al. Diffusion-weighted magnetic resonance imaging as a cancer biomarker: Consensus and recommendations. *Neoplasia*. 2009;11(2):102–125. doi:10.1593/neo.81328
7. Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, Laval-Jeanted M. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. *Radiology*. 1988;168(2):497-505. doi:10.1148/radiology.168.2.3393671
8. Bernstein JM, Homer JJ, West CM. Dynamic contrast-enhanced magnetic resonance imaging biomarkers in head and neck cancer: Potential to guide treatment? A systematic review. *Oral Oncol*. 2014;50(10):963-970. doi:10.1016/j.oraloncology.2014.07.011
9. Noij DP, de Jong MC, Mulders LGM, et al. Contrast-enhanced perfusion magnetic resonance imaging for head and neck squamous cell carcinoma: A systematic review. *Oral Oncol*. 2015;51(2):124-138. doi:10.1016/j.oraloncology.2014.10.016
10. Chung SR, Choi YJ, Suh CH, Lee JH, Baek JH. Diffusion-weighted magnetic resonance imaging for predicting response to chemoradiation therapy for head and neck squamous cell carcinoma: A systematic review. *Korean J Radiol*. 2019;20(4):649-661. doi:10.3348/kjr.2018.0446
11. Noij DP, Martens RM, Marcus JT, et al. Intravoxel incoherent motion magnetic resonance imaging in head and neck cancer: A systematic review of the diagnostic and prognostic value. *Oral Oncol*. 2017;68:81–91. doi:10.1016/j.oraloncology.2017.03.016
12. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *J Clin Epidemiol*. 2021;134:178-

189. doi:10.1016/j.jclinepi.2021.03.001

13. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan–A web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210. doi:10.1186/s13643-016-0384-4
14. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ. Cochrane Handbook for Systematic Reviews of Interventions, Version 6.2. Cochrane. 2021. Available from www.training.cochrane.org/handboek.
15. Cohen JW. Statistical power analysis for the behavioural sciences, second ed., lawrence Eribaum Associates, New York, 1988.
16. Baer AH, Hoff BA, Srinivasan A, Galban CJ, Mukherji SK. Feasibility analysis of the parametric response map as an early predictor of treatment efficacy in head and neck cancer. *AJNR AM J Neuroradiol*. 2015;36(4):757–762. doi:10.3174/ajnr.A4296
17. Cao Y, Aryal M, Li P, et al. Predictive Values of MRI and PET Derived Quantitative Parameters for Patterns of Failure in Both p16+ and p16– High Risk Head and Neck Cancer. *Front Oncol*. 2019;9:1118. doi:10.3389/fonc.2019.01118
18. Chan SC, Cheng NM, Hsieh CH, et al. Multiparametric imaging using 18F-FDG PET/CT heterogeneity parameters and functional MRI techniques: Prognostic significance in patients with primary advanced oropharyngeal or hypopharyngeal squamous cell carcinoma treated with chemoradiotherapy. *Oncotarget*. 2017;8(37):62606–62621. doi:10.18632/oncotarget.15904
19. Chawla S, Kim S, Dougherty L, et al. Pre-treatment diffusion-weighted and dynamic contrast-enhanced MRI for prediction of local treatment response in squamous cell carcinomas of the head and neck. *AJR Am J Roentgenol*. 2013;200(1):35–43. doi:10.2214/AJR.12.9432
20. Garbajs M, Strojan P, Surlan-Popovic K. Prognostic role of diffusion weighted and dynamic contrast-enhanced MRI in loco-regionally advanced head and neck cancer treated with concomitant chemoradiotherapy. *Radiol Oncol*. 2019;53(1):39–48. doi:10.2478/raon-2019-0010
21. Lowe NM, Kershaw LE, Bernstein JM, et al. Pre-treatment tumour perfusion parameters and initial RECIST response do not predict long-term survival outcomes for patients with head and neck squamous cell carcinoma treated with induction chemotherapy. *PLoS ONE*. 2018;13(3):e0194841. doi:10.1371/journal.pone.0194841
22. Martens RM, Koopman T, Lavini C, et al. Multiparametric functional MRI and 18F-FDG-PET for survival prediction in patients with head and neck squamous cell carcinoma treated with (chemo)radiation. *Eur Radiol*. 2021;31(2):616–628. doi:10.1007/s00330-020-07163-3
23. Ng SH, Lin CY, Chan SC, et al. Dynamic Contrast-Enhanced MR Imaging Predicts Local Control in Oropharyngeal or Hypopharyngeal Squamous Cell Carcinoma Treated with Chemoradiotherapy. *PLoS One*. 2013;8(8):e72230. doi:10.1371/journal.pone.0072230

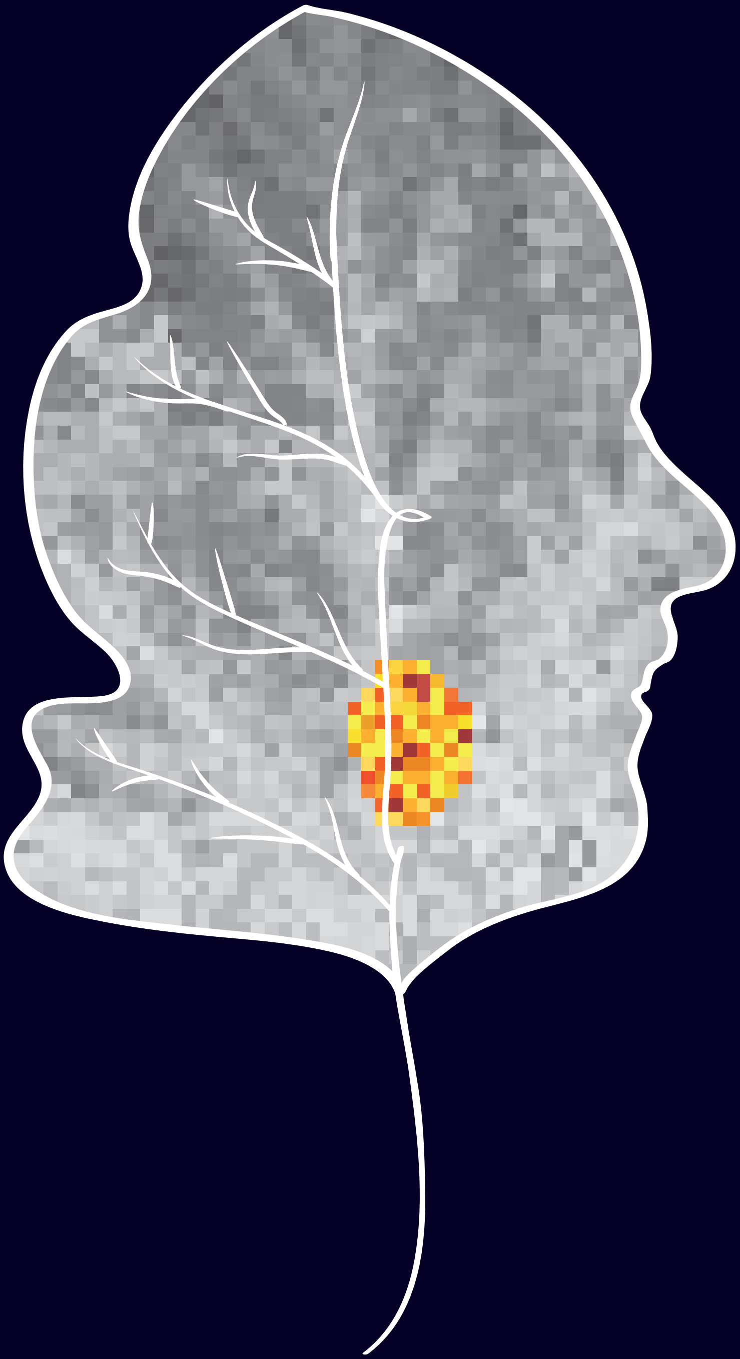
24. Ng SH, Liao CT, Lin CY, et al. Dynamic contrast-enhanced MRI, diffusion- weighted MRI and 18F-FDG PET/CT for the prediction of survival in oropharyngeal or hypopharyngeal squamous cell carcinoma treated with chemoradiation. *Eur Radiol*. 2016;26:4162–4172. doi:10.1007/s00330-016-4276-8
25. Wang P, Popovtzer A, Eisbruch A, Cao Y. An approach to identify, from DCE MRI, significant subvolumes of tumors related to outcomes in advanced head-and-neck cancer. *Med Phys*. 2012;39(8):5277–5285. doi:10.1118/1.4737022
26. Wong KH, Panek R, Dunlop A, et al. Changes in multimodality functional imaging parameters early during chemoradiation predict treatment response in patients with locally advanced head and neck cancer. *Eur J Nucl Med Mol Imaging*. 2018;45(5):759–767. doi:10.1007/s00259-017-3890-2
27. Choi JW, Lee D, Hyun SH, Han M, Kim JH, Lee SJ. Intratumoural heterogeneity measured using FDG PET and MRI is associated with tumour–stroma ratio and clinical outcome in head and neck squamous cell carcinoma. *Clin Radiol*. 2017;72(6):482–489. doi:10.1016/j.crad.2017.01.019
28. Gupta T, Chatterjee A, Rangarajan V, et al. Evaluation of quantitative imaging parameters in head and neck squamous cell carcinoma. *Q J Nucl Med Mol Imaging*. 2022;66(2):162-170. doi:10.23736/S1824-4785.19.03179-0
29. Hatakenaka M, Shioyama Y, Nakamura K, et al. Apparent diffusion coefficient calculated with relatively high b-values correlates with local failure of head and neck squamous cell carcinoma treated with radiotherapy. *AJNR Am J Neuroradiol*. 2011;32(10):1904–1910. doi:10.3174/ajnr.A2610
30. Hatakenaka M, Nakamura K, Yabuuchi H, et al. Pre-treatment apparent diffusion coefficient of the primary lesion correlates with local failure in head-and-neck cancer treated with chemoradiotherapy or radiotherapy. *Int J Radiat Oncol Biol Phys*. 2011;81(2):339–345. doi:10.1016/j.ijrobp.2010.05.051
31. Hatakenaka M, Nakamura K, Yabuuchi H, et al. Apparent diffusion coefficient is a prognostic factor of head and neck squamous cell carcinoma treated with radiotherapy. *Jpn J Radiol*. 2014;32(2):80–89. doi:10.1007/s11604-013-0272-y
32. King AD, Chow KK, Yu KH, et al. Head and neck squamous cell carcinoma: Diagnostic performance of diffusion-weighted MR imaging for the prediction of treatment response. *Radiology*. 2013;266(2):531–538. doi:10.1148/radiol.12120167
33. Lambrecht M, van Calster B, Vandecaveye V, et al. Integrating pre-treatment diffusion weighted MRI into a multivariable prognostic model for head and neck squamous cell carcinoma. *Radiother Oncol*. 2014;110(3):429–434. doi:10.1016/j.radonc.2014.01.004
34. Li X, Yuan Y, Ren J, Shi Y, Tao X. Incremental Prognostic Value of Apparent Diffusion Coefficient Histogram Analysis in Head and Neck Squamous Cell Carcinoma. *Acad Radiol*. 2018;25(11):1433-1438. doi:10.1016/j.acra.2018.02.017
35. Lu Y, Jansen JFA, Stambuk HE, et al. Comparing primary tumors and metastatic nodes

in head and neck cancer using intravoxel incoherent motion imaging: a preliminary experience. *J Comput Assist Tomogr*. 2013;37(3):346-352. doi:10.1097/RCT.0b013e318282d935

36. Martens RM, Noij DP, Koopman T, et al. Predictive value of quantitative diffusion-weighted imaging and 18-F-FDG-PET in head and neck squamous cell carcinoma treated by (chemo)radiotherapy. *Eur J Radiol*. 2019;113:39-50. doi:10.1016/j.ejrad.2019.01.031
37. Matoba M, Tuji H, Shimode Y, et al. Fractional change in apparent diffusion coefficient as an imaging biomarker for predicting treatment response in head and neck cancer treated with chemoradiotherapy. *AJNR Am J Neuroradiol*. 2014;35(2):379–385. doi:10.3174/ajnr.A3706
38. Nakajo M, Nakajo M, Kajiya Y, et al. FDG PET/CT and Diffusion-Weighted Imaging of Head and Neck Squamous Cell Carcinoma. *Clin Nucl Med*. 2012;37(5):475–480. doi:10.1097/RLU.0b013e318248524a
39. Noij DP, Pouwels PJW, Ljumanovic R, et al. Predictive value of diffusion- weighted imaging without and with including contrast-enhanced magnetic resonance imaging in image analysis of head and neck squamous cell carcinoma. *Eur J Radiol*. 2015;84(1):108–116. doi:10.1016/j.ejrad.2014.10.015
40. Núnez DA, Medina AL, Iglesias MM, et al. Multimodality functional imaging using DW-MRI and 18 F-FDG-PET/CT during radiation therapy for human papillomavirus negative head and neck squamous cell carcinoma: Meixoeiro Hospital of Vigo Experience. *World J Radiol*. 2017;9(1):17–26. doi:10.4329/wjr.v9.i1.17
41. Peltenburg B, Driessen JP, Vasmel JE, et al. Pre-treatment ADC is not a prognostic factor for local recurrences in head and neck squamous cell carcinoma when clinical T-stage is known. *Eur Radiol*. 2020;30:1228–1231. doi:10.1007/s00330-019-06426-y
42. Preda L, Conte G, Bonello L, et al. Combining standardized uptake value of FDG- PET and apparent diffusion coefficient of DW-MRI improves risk stratification in head and neck squamous cell carcinoma. *Eur Radiol*. 2016;26(12):4432–4441. doi:10.1007/s00330-016-4284-8
43. Ravanelli M, Grammatica A, Maddalo M, et al. Pre-treatment DWI with Histogram Analysis of the ADC in Predicting the Outcome of Advanced Oropharyngeal Cancer with Known Human Papillomavirus Status Treated with Chemoradiation. *AJNR Am J Neuroradiol*. 2020;41(8):1473–1479. doi:10.3174/ajnr.A6695
44. Ren JL, Yuan Y, Li XX, Shi YQ, Tao XF. Histogram analysis of apparent diffusion coefficient maps in the prognosis of patients with locally advanced head and neck squamous cell carcinoma: Comparison of different region of interest selection methods. *Eur J Radiol*. 2018;106:7–13. doi:10.1016/j.ejrad.2018.07.004
45. Srinivasan A, Chenevert TL, Dwamena BA, et al. Utility of pre-treatment mean apparent diffusion coefficient and apparent diffusion coefficient histograms in prediction of outcome to chemoradiation in head and neck squamous cell carcinoma.

J Comput Assist Tomogr. 2012;36(1):131–137. doi:10.1097/RCT.0b013e3182405435

46. Zhang SC, Zhou SH, Shang DS, Bao YY, Ruan LX, Wu TT. The diagnostic role of diffusion-weighted magnetic resonance imaging in hypopharyngeal carcinoma. *Oncol Lett*. 2018;15(4):5533–5544. doi:10.3892/ol.2018.8053
47. Lemke A, Laun FB, Simon D, Stieltjes B, Schad LR. An in vivo verification of the intravoxel incoherent motion effect in diffusion-weighted imaging of the abdomen. *Magn Reson Med*. 2010;64(6):1580–1585. doi:10.1002/mrm.22565
48. Iima M, Partridge SC, Le Bihan D. Six DWI questions you always wanted to know but were afraid to ask: clinical relevance for breast diffusion MRI. *Eur Radiol*. 2020;30(5):2561–2570. doi:10.1007/s00330-019-06648-0
49. Kolff-Gart AS, Pouwels PJW, Noij DP, et al. Diffusion-weighted imaging of the head and neck in healthy subjects: Reproducibility of ADC values in different MRI systems and repeat sessions. *AJNR Am J Neuroradiol*. 2015;36(2):384–390. doi:10.3174/ajnr.A4114
50. Shukla-Dave A, Obuchowski NA, Chenevert TL, et al. Quantitatie imaging biomarkers alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. *J Magn Reson Imaging*. 2019;49(7):e101–e121. doi:10.1002/jmri.26518
51. King AD, Thoeny HC. Functional MRI for the prediction of treatment response in head and neck squamous cell carcinoma: Potential and limitations. *Cancer Imaging*. 2016;16(1):16–23. doi:10.1186/s40644-016-0080-6
52. Mes SW, van Velden FHP, Peltenburg B, et al. Outcome prediction of head and neck squamous cell carcinoma by mri radiomic signatures. *Eur Radiol*. 2020;30(11):6311–6321. doi:10.1007/s00330-020-06962-y



Part II

MR-based radiomic prediction models for tumor characterisation and prognosis in OPSCC patients

Improved outcome prediction of oropharyngeal cancer by combining clinical and MRI features in machine learning models

European Journal of Radiology; Jun 2021;139:109701
DOI: 10.1016/j.ejrad.2021.109701

Paula Bos
Michiel W.M. van den Brekel
Zeno A.R. Gouw
Abrahim Al-Mamgani
Marjaneh Taghavi
Selam Waktola
Hugo J.W.L. Aerts
Jonas A. Castelijns
Regina G.H. Beets-Tan
Bas Jasperse

3

ABSTRACT

***Objectives**:* New markers are required to predict chemoradiation response in oropharyngeal squamous cell carcinoma (OPSCC) patients. This study evaluated the ability of magnetic resonance (MR) radiomics to predict locoregional control (LRC) and overall survival (OS) after chemoradiation and aimed to determine whether this has added value to traditional clinical outcome predictors.

***Methods**:* 177 OPSCC patients were eligible for this study. Radiomic features were extracted from the primary tumor region in T1-weighted postcontrast MRI acquired before chemoradiation. Logistic regression models were created using either clinical variables (clinical model), radiomic features (radiomic model) or clinical and radiomic features combined (combined model) to predict LRC and OS 2-years posttreatment. Model performance was evaluated using area under the curve (AUC), 95% confidence intervals were calculated using 500 iterations of bootstrap. All analyses were performed for the total population and the human papillomavirus (HPV) negative tumor subgroup.

***Results**:* A combined model predicted treatment outcome with a higher AUC (LRC: 0.745 [0.734–0.757], OS: 0.744 [0.735–0.753]) than the clinical model (LRC: 0.607 [0.594-0.620], OS: 0.708 [0.697–0.719]). Performance of the radiomic model was comparable to the combined model for LRC (AUC: 0.740 [0.729–0.750]), but not for OS prediction (AUC: 0.654 [0.646–0.662]). In HPV negative patients, the performance of all models was not sufficient with AUCs ranging from 0.587 to 0.660 for LRC and 0.559 to 0.600 for OS prediction.

***Conclusion**:* Predictive models that include clinical variables and radiomic tumor features derived from MR images of OPSCC better predict LRC after chemoradiation than models based on only clinical variables. Predictive models that include clinical variables perform better than models based on only radiomic features for the prediction of OS.

INTRODUCTION

Oropharyngeal squamous cell carcinoma (OPSCC) is a frequent tumor of the upper aero-digestive tract, with an increasing incidence in the last decades[1]. Although definitive chemo- and radiation therapy (chemoradiation (CRT)) is currently considered the standard of care for patients with locally advanced OPSCC, surgery, especially minimal invasive transoral robotic surgery (TORS), followed by CRT can be a good alternative and might enable de-intensification of the postoperative CRT, depending on the disease stage and patients' or clinicians' preference[2,3]. Although CRT has a high rate of treatment response, a considerable number of OPSCC patients have recurrent or residual disease after CRT leading to significant morbidity, mortality, and deterioration of quality of life. HPV tumor status is the most important predictor of treatment success, generally showing better treatment outcomes for HPV positive and less favorable treatment outcomes for HPV negative tumors[4]. Additional markers to predict CRT response are needed especially for HPV negative patients, allowing these patients to undergo an alternative treatment strategy (e.g. neoadjuvant chemotherapy combined with TORS or induction immunotherapy) at an early stage of the treatment trajectory.

Over the past years, image analysis techniques have been developed to extract and quantify visually occult tumor properties from computer tomography (CT) and MR images, collectively called radiomics features. These radiomic features have been associated with gene expression, histological tissue properties, survival, and treatment outcome. Previous studies on this topic have found prognostic radiomic features from CT images. For instance, intratumor heterogeneity quantified on CT images proved to be predictive of survival[5]. Compared to CT, MRI may provide other insights in tissue properties due to fundamental differences in image acquisition[6]. Few studies have investigated prognostic radiomic features from MRI images of head and neck cancer. These studies mainly focused on outcome prediction in nasopharyngeal carcinoma using radiomics or deep learning[6-10]. MRI is the preferred modality for OPSCC patients in most centers, providing a unique chance to study the ability of MRI radiomics to predict treatment outcome.

This study aimed to predict CRT treatment outcome for OPSCC using radiomic features derived from pretreatment MR images, and to determine whether these MR-based radiomic features have added value to clinical predictors of treatment outcome.

MATERIALS AND METHODS

The institutional ethics review board approved the study. Informed consent was waived for this retrospective analysis of anonymous data.

Patients

A total of 240 consecutive OPSCC patients, treated with CRT between January 2010 and December 2015 at our institute, were considered for this study. Inclusion criteria were: 1) histologically proven primary OPSCC treated with CRT, 2) minimum of 2 years of follow-up after treatment and, 3) availability of relevant clinical parameters. Exclusion criteria were unavailable pretreatment MRI examination of the primary tumor (n=38), poor image quality (n=7), and small undetectable (n=17) or double tumors (n=1). A total of 177 patients were eligible for this study.

Age, gender, smoking status (non-smoker vs. smoker), date of tumor recurrence, occurrence of lymph node metastasis, and survival within 2-years after treatment were collected for all patients. TNM-stage (7th edition), subsite and HPV status based on immunohistochemistry p16 and DNA HPV polymerase chain reaction were collected for each tumor. Clinical variables age and TNM stage were dichotomized to create groups of patients younger or older than 60 years, low and high T-stage (T1+T2, T3+T4) and positive or negative nodal disease.

Treatment

Patients were treated by chemoradiotherapy using Image-guided Intensity-modulated radiation therapy (IMRT) or volumetric modulated arc therapy (VMAT). Prescribed dose was 70 Gy to the primary tumor and the involved nodes and 46 Gy electively to the low-risk regions. The radiation was given in a daily fraction of 2 Gy, 5 times a week for 7 weeks. Set-up verification and correction of the patients was done using daily cone-beam CT. Patients received three cycles of cisplatin-based chemotherapy (100 gr/m^2), administered on day 1, 22 and 43 of their radiation treatment.

Outcome variables

The primary outcome was locoregional control (LRC), defined as the absence of a histopathological proven local recurrence and/or lymph node metastases within 2 years after initial complete response. Secondary outcome was overall survival (OS), defined as the proportion of patients surviving 2 years after treatment.

Imaging data

Pretreatment MRI was routinely performed as part of primary staging for patients with OPSCC. All MRI examinations were acquired at 1.5 Tesla (n=82 patients) or 3.0 Tesla (n=95 patients) on a Philips Medical System, see supplementary Table 1 for detailed acquisition information. The full imaging protocol included T1w, T2w, postcontrast 3D T1w and dynamic scans.

Tumor delineation

Primary tumors were manually delineated by one observer in training (PB, 1 year of head and neck experience, non-expert delineations), and, subsequently controlled and corrected by an experienced head and neck radiologist (BJ, >7 years of head and neck experience) on the postcontrast 3D T1w MRI using the freely available segmentation software 3D Slicer (version 4.8.0, www.slicer.org) (see Figure 1). Average spatial agreement was good with a mean dice similarity coefficient (DSC) of 0.83. Dice similarity coefficient was between 0.9–1.0 in 53% of the cases, between 0.8–0.9 in 22%, between 0.7-0.8 in 9%, and, below 0.7 in 16% of the cases. Larger tumor volumes showed significantly better overlap compared to small tumor volumes (p=0.001, independent t-test). Tumor volumes were delineated on every axial slice on the postcontrast 3D T1w MRI. Both observers were blinded to outcome data but were allowed to interpret other available pretreatment imaging data to optimize their delineations.

Feature extraction

Imaging features were extracted from tumor volumes using the open-source python package, Pyradiomics (version 2.2.0)[11]. All MRI examinations were normalized (centering at zero mean and one standard deviation) to obtain a homogeneous histogram of MR signal and resampled by B-spline interpolation to a pixel spacing of 1.0×1.0×1.0 mm^3. Gray values were discretized using a fixed bin width of five. Features were extracted from the image data three times: original image, with a wavelet image filter, including eight decompositions, and finally with a Laplacian of Gaussian (LoG) filter (four levels (0.5–2.0 mm).

Features with zero variance (i.e. constant features), and therefore of no discriminatory value, were removed. Features were considered stable, if they had no significant difference between non-expert and expert tumor delineations (intraclass correlation coefficient >0.75) and between magnetic field strengths (Mann-Whitney *U* test p ≥0.05). The remaining stable features were then tested for collinearity. Features that correlated with other features with a Pearson coefficient higher than 0.9 were removed. In this removal process, the feature that showed high correlation with the greatest number of other features was removed. This was repeated until only the diagonal elements of the correlation matrix exceeded the threshold of 0.9.

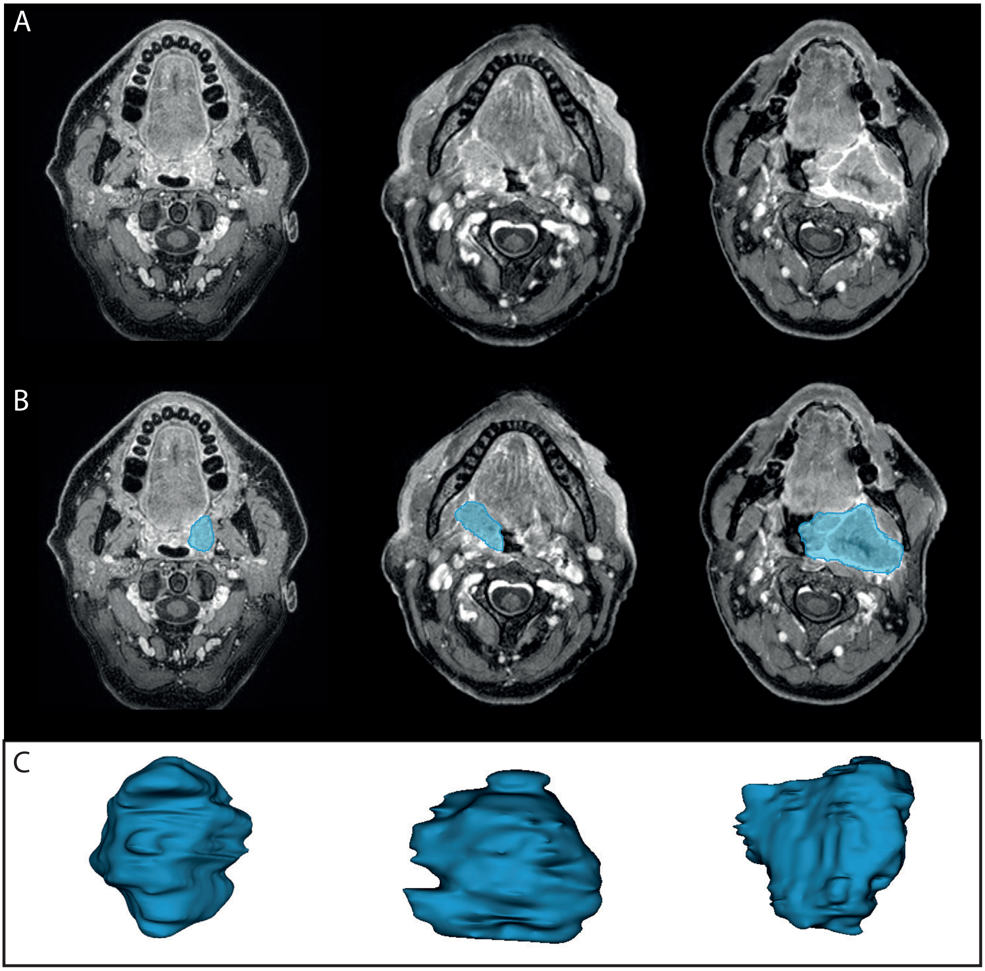


Fig. 1. Examples of tumor delineations on postcontrast 3D T1w MRIs. From top to bottom the MRI without manual delineation (A), MRI with manual delineation (B) and the reconstructed 3D tumor volumes (C) for three patients (left, middle, right).

Machine learning analysis

Analysis was performed in all eligible patients (n=177) and in a subset of patients with HPV negative tumors (n=77). Sub-analysis of patients with HPV positive tumors (n=76) was considered of limited added value, as the majority had favorable outcomes for both LRC and OS (LRC: 68/76, OS: 67/76). In 24 patients, HPV status was unavailable. Patients were randomly split into a training (70%) and test-set (30%), see Table 1, stratifying for treatment outcome and MRI field strength. HPV status was included as stratification factor for the total patient cohort.

Three models were created for each of the outcome variables (LRC and OS) using only clinical variables (clinical model), only radiomic features (radiomics model) and a combination of clinical variables and radiomic features (combined model). Features were prepared for logistic regression analysis using the following steps: 1) Standardization of features to zero mean and unit variance, and 2) Reduction of the number of features by wrapper feature selection using a sequential backward feature selection method, which removed irrelevant features by iteratively removing the feature with the weakest feature importance score[12].

In the training phase, optimal model settings of the machine learning pipeline were found utilizing 1000 iterations of Bayesian hyperparameter optimization (Python library Hyperopt version 0.2[13]), applying fourfold cross validation within the training set (see Table 1 for patient numbers). The regularization parameter and the number of selected features in wrapper feature selection were tuned during Bayesian hyperparameter optimization (supplementary Table 2). Training performance of the predictive models was evaluated using median AUC and its 95% confidence interval (95% CI) from the performance of the optimal hyperparameters in fourfold cross validation.

In the testing phase, the optimal hyperparameter combination obtained in the training step was applied to the unseen test dataset. Model test performance was evaluated by the median and 95% CI of AUC, sensitivity, specificity and accuracy obtained using 500 iterations of bootstrap (with replacement).

Table 1. Detailed information of patient numbers in training and test set for the development of a prediction model.

	Total patient cohort	**HPV negative subset**
Total number of patients	177	77
Training set (70%)	124	53
Cross validation: training (75%)	93	40
Cross validation: validation (25%)	31	13
Test set (30%)	53	24

Statistical analysis

Univariate Fishers' exact test was used to test differences in clinical features between groups with regard to outcome parameters (OS and LRC). P-values <0.05 were considered statistically significant (p=0.004 after Bonferroni correction). Statistical differences between the predictive radiomic features of the models were tested using the Wald test (p-values <0.05 were considered statistically

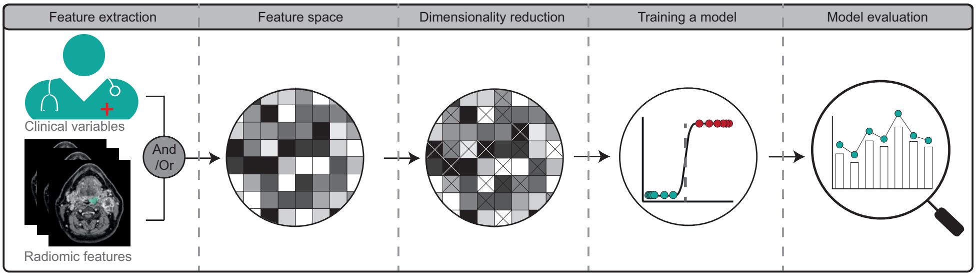


Fig. 2. Flowchart of the radiomics workflow. First clinical variables and/or radiomic features were extracted from the patient and MR image respectively. Feature space including only clinical variables, only radiomic features or the combination were created to build a clinical, radiomic or combined prediction model respectively. After dimensionality reduction, using wrapper feature selection, a logistic regression prediction model is trained and model performance is evaluated.

significant). All analyses were implemented in python 3.5 and SPSS version 25.0 (SPSS Inc.). The radiomic workflow is visualized in Figure 2.

RESULTS

Detailed patient characteristics and oncologic outcomes are summarized in Table 2. Of the total patient group, approximately half had a high T-stage (T3-T4) and 80% had node-positive disease. Considering only patients with known HPV status revealed an equal distribution between positive (n=76) and negative (n=77) HPV status. Patients with favorable outcomes for LRC and OS were more likely to have HPV positive tumors (LRC: p=0.004, OS: p=0.001).

Predictive performance of models for all patients

Out of 1184 radiomic features, 75 features were stable. Performance of the clinical, radiomics and combined model based on logistic regression for prediction of LRC and OS are summarized in Table 3.

Locoregional control

The predictive properties of the clinical model (Test AUC: 0.607, Sens: 0.57, Spec: 0.60, Acc: 0.57) are less favorable compared to the radiomic model (Test AUC: 0.740, Sens: 0.75, Spec: 0.60, Acc: 0.71) with regard to LRC. The combined model (Test AUC: 0.745, Sens: 0.73, Spec: 0.71, Acc: 0.71) shows a similar performance as the radiomic model.

Lower T-stage (r: 0.330), HPV positivity (r: 0.305), tumor not located at the posterior oropharyngeal wall (r: −0.174) and lower age (r: −0.166) were predictive

Table 2. Patient demographics. Baseline characteristics and outcome after CRT for all patients, and HPV negative tumors. Summaries are given as number of patients and % of the total group between parentheses. Median and interquartile range (IQR) are used to summarize continuous variables. Fisher exact test after Bonferroni correction *p=0.004 and p=0.001 for LRC and OS respectively. Clinical values were only significant for total patient cohort.

Patients, *n*	**Total patient cohort (n=177)**	**HPV negative tumors (n=77)**	**HPV positive tumors (n=76)**
Age (>60years)	101 (57)	52 (67)	36 (47)
Sex, *n male* (%)	111 (63)	54 (70)	42 (55)
Smoking, *n* (%)	134 (76)	72 (94)	42 (55)
HPV*			
Negative, *n* (%)	77 (44)	77 (100)	–
Positive, *n* (%)	76 (43)	–	76 (100)
Unknown, *n* (%)	24 (13)	–	–
T-stage, *n* (%)			
T1+T2	94 (53)	25 (33)	53 (70)
T3+T4	83 (47)	52 (67)	23 (30)
N-stage, *n* (%)			
N0	36 (20)	18 (23)	8 (11)
N1	26 (15)	11 (15)	12 (16)
N2	110 (62)	47 (61)	52 (68)
N3	5 (3)	1 (1)	4 (5)
Subsite of cancer			
Tonsillar tissue	99 (56)	42 (55)	46 (60)
Soft palate	18 (10)	11 (14)	2 (3)
Base of tongue	56 (32)	20 (26)	28 (37)
Posterior wall	4 (2)	4 (5)	0 (0)
Clinical endpoints LRC <2 year, *n* (%)	144 (81)	55 (71)	66 (87)
Time to LRF in months, *median* (IQR)	6 (4–17)	6 (4–13)	9 (4–18)
OS after 2 years, *n* (%)	137 (77)	50 (65)	72 (95)
OS in months for non-survivors, *median* (IQR)	12 (8–17)	14 (9–18)	12 (10–15)

HPV: Human papillomavirus; **LRC**: Locoregional control; **OS**: Overall survival; **LRF**: Locoregional failure determinants of LRC in the clinical model (supplementary Table 3). Four and five radiomic features were selected in the radiomic and combined model, respectively. Rounder and more homogeneous tumors were associated with disease control (supplementary Table 4). No clinical variables were selected in the combined model.

Overall survival

For the prediction of OS, the predictive performance of the clinical model (Test AUC: 0.708, Sens: 0.68, Spec: 0.67, Acc: 0.69) is better than the radiomic model (Test AUC: 0.654, Sens: 0.62, Spec: 0.57, Acc: 0.60). The combined model (Test AUC: 0.744, Sens: 0.71, Spec: 0.78, Acc: 0.71) had the highest performance and outperformed the two other models.

Eight, ten and twenty-two features were prognostic for overall survival, regarding respectively the clinical, radiomic and combined model (supplementary Table 3). In the clinical model, lower T-stage (r: 0.409), younger patients (r: −0.395), HPV positivity (r: 0.348), node-negative disease (r: 0.232), tumors not located in the posterior oropharyngeal wall (r: −0.147), tumors located at the base of tongue (r: 0.095) and female gender (r: −0.041) were associated with OS. Radiomic features show less complex, coarse and more homogeneous tumors in patients who are more likely to survive (supplementary Table 4).

Predictive performance of HPV negative tumors

After feature reduction, 123 features remained for the HPV negative subgroup. Table 4 summarizes predictive properties of prediction models in HPV negative tumors.

Performance of all models was generally low for LRC (Test AUCs 0.587 to 0.660) and OS (Test AUCs 0.559 to 0.600). Performance of the clinical model was lower than the model based on radiomic features for both LRC (Test AUC: 0.587 and 0.652 respectively) and OS (Test AUC: 0.559 and 0.593 respectively). Performance of the radiomic model was comparable to the combined model for both LRC (Test AUC: 0.660, Sens: 0.83, Spec: 0.43, Acc: 0.71) and OS (Test AUC: 0.600, Sens: 0.40, Spec: 0.67, Acc: 0.51).

DISCUSSION

The main finding of this study was that predictive models based on a combination of clinical variables and MR-based radiomic features have a reasonable ability to predict LRC and OS within 2 years after CRT in OPSCC. Sub analysis of HPV negative

Table 3. Performance expressed as AUC [95% CI] for the models predicting LRC and OS within 2 years after chemoradiation for all patients. Confidence intervals were calculated from 500 iterations of bootstrapping.

Model	Training AUC [CV]	Test AUC [CI bootstrap]	Sensitivity [CI bootstrap]	Specificity [CI bootstrap]	Accuracy [CI bootstrap]
LRC					
Clinical	0.637 [0.572-0.702]	0.607 [0.594-0.620]	0.57 [0.56-0.58]	0.60 [0.58-0.62]	0.57 [0.56-0.58]
Radiomic	0.783 [0.690-0.875]	0.740 [0.729-0.750]	0.75 [0.74-0.76]	0.60 [0.58-0.62]	0.71 [0.71-0.72]
Combined	0.747 [0.640-0.855]	0.745 [0.734-0.757]	0.73 [0.72-0.73]	0.71 [0.70-0.73]	0.71 [0.71-0.72]
OS					
Clinical	0.659 [0.558-0.760]	0.708 [0.697-0.719]	0.68 [0.67-0.69]	0.67 [0.65-0.68]	0.69 [0.68-0.69]
Radiomic	0.601 [0.501-0.702]	0.654 [0.646-0.662]	0.62 [0.61-0.62]	0.57 [0.56-0.59]	0.60 [0.59-0.61]
Combined	0.548 [0.519-0.577]	0.744 [0.735-0.753]	0.71 [0.70-0.72]	0.78 [0.76-0.79]	0.71 [0.71-0.72]

CV: Cross validation

Table 4. Performance expressed as AUC [95% CI] for the models predicting LRC and OS within 2 years after chemoradiation for all patients with HPV negative tumors. Confidence intervals were calculated from 500 iterations of bootstrapping.

Model	Training AUC [CV]	Test AUC [CI bootstrap]	Sensitivity [CI bootstrap]	Specificity [CI bootstrap]	Accuracy [CI bootstrap]
LRC					
Clinical	0.510 [0.442-0.579]	0.587 [0.578-0.595]	0.71 [0.70-0.72]	0.27 [0.26-0.29]	0.57 [0.56-0.58]
Radiomic	0.706 [0.510-0.901]	0.652 [0.642-0.661]	0.83 [0.83-0.84]	0.43 [0.41-0.44]	0.71 [0.71-0.72]
Combined	0.706 [0.510-0.901]	0.660 [0.650-0.670]	0.83 [0.83-0.84]	0.43 [0.41-0.44]	0.71 [0.71-0.72]
OS					
Clinical	0.606 [0.390-0.821]	0.559 [0.543-0.563]	0.47 [0.46-0.48]	0.67 [0.65-0.68]	0.54 [0.54-0.55]
Radiomic	0.501 [0.409-0.593]	0.593 [0.583-0.602]	0.60 [0.59-0.61]	0.32 [0.31-0.34]	0.51 [0.51-0.52]
Combined	0.478 [0.360-0.596]	0.600 [0.591-0.609]	0.40 [0.39-0.41]	0.67 [0.65-0.68]	0.51 [0.51-0.52]

CV: Cross validation

patients showed moderate performance in the prediction of LRC and poor performance in the prediction of OS.

Interestingly, predictive performance of models based on only clinical variables was not as good as the combined models. This implies that clinical variables and radiomics features hold independent information for outcome prediction. Radiomic features are likely to add information embedded in tumor structure for the prediction of treatment outcome not captured by clinical variables. Clinical variables may add to the radiomic features in different ways for LRC and OS. For LRC, information is added to the risk of recurrence by clinical factors that influence tumor biology, such as HPV status and age. For OS, non-tumor related information is added to risk of death, like age and comorbidities. These findings, indicates that clinical and imaging features should preferably be combined when constructing models to predict treatment outcome. This is in line with findings of Mes et al.[14] for oral cancer patients and HPV negative OPSCC.

For the prediction of LRC, the combined model consisted of only radiomic features while both clinical variables and radiomic features were included in the construction of the model. The performance of this combined model was slightly better than the radiomic model due to the addition of the radiomics variable skewness. Additional analysis (not shown) revealed that the correlation of clinical variables with selected radiomic features was low. Apparently, the combination of clinical variables and radiomics variables in the model construction sequence makes slight improvements in the eventual combined model compared to the radiomics model, in this case with the addition of skewness. This slight improvement occurred even though clinical variables do not obviously correlate with the radiomics variables**.** This is an important consideration to take into account in construction of predictive radiomics models.

For prediction of OS, the combined model consisted of a relatively large number of radiomic features and clinical variables. As mentioned previously, risk of death includes a broad range of factors that are not directly tumor related. The large number of clinical and radiomic features with generally low regression coefficients in the combined model for prediction of OS reflects this.

The radiomic features revealed that rounder and more homogenous tumors are associated with a more favorable outcome. This relationship is probably a reflection of genetic tumor diversity/dedifferentiation resulting in more heterogeneous and irregular tumors with worse treatment response and higher rate of locoregional failure. These findings are in line with another MRI-based radiomics study of head and neck cancer showing higher homogeneity and rounder shapes for overall survival[7].

HPV is an important determinant of the biology and behavior of OPSCC, and is known to be a strong predictor of treatment outcome in OPSCC, prompting us to create separate models for HPV positive and negative tumors. As expected, most patients with HPV positive tumors had a favorable outcome, which did not permit us to create a meaningful model for this tumor type. Distribution of outcome variables for HPV negative tumors permitted the construction of a predictive model, but did not reach consistent meaningful predictive properties. This was probably due to the low number of patients (n=77) in this subgroup. The role of clinical and radiomics models in outcome prediction for HPV positive and, particularly, negative tumors therefore remains unclear.

This study has a relatively large sample size (n=177) compared to other published studies on MR radiomics in head and neck cancer (maximum 118 patients)[10]. However, these results are not generalizable to other hospitals with different scanners and scanner protocols. The next step is to replicate these findings in an external validation cohort from multiple centers[15-17].

MRI based radiomics is still difficult to implement in the clinical workup due to a lack of standardization in acquisition parameters and harmonization between MRI machines, as was shown in previous studies[14,16,17]. Until standardization of acquisition is available, standardization between centers can be reached by harmonizing pre-processing steps and correlation analysis to obtain stable features between centers. Even though stable feature reduces bias introduced by human interaction by manual delineation, some human influence cannot be ruled out completely. In the future, automated delineation techniques may be able to eliminate this unwanted bias.

This study extracted radiomic features from primary tumors based on postcontrast T1w MRI. Extracting features from other MR sequences might give a better representation of tumor biology, and may harbor information relevant to treatment outcome. For instance, the dynamic contrast-enhanced MRI parameters have shown its prognostic ability to predict OS and progression-free survival[18]. Sample size considerations and preliminary results prompted us to use only T1w 3D sequences to ensure meaningful results. Evidently, this needs to be considered in future studies.

This study shows that predictive models that include radiomic tumor features derived from MR images of OPSCC better predict LRC after chemoradiation than models based on only clinical variables. Predictive models that include clinical variables perform better than models based on only radiomic features for the prediction of OS.

Future studies on MRI based radiomics should confirm these findings in a larger patient cohort and elucidate the potential role of radiomics in outcome prediction in HPV positive and, especially, HPV negative tumors.

SUPPLEMENTARY INFORMATION

Password: PhD_PaulaBos

REFERENCES

1. van Monsjou HA, Schaapveld M, Hamming-Vrieze O, de Boer JP, van den Brekel MWM, Balm AJM. Cause-specific excess mortality in patients treated for cancer of the oral cavity and oropharynx: a population-based study. *Oral Oncol*. 2016;52:37–44. doi:10.1016/j.oraloncology.2015.10.013
2. Bonner JA, Harari PM, Giralt J, et al. Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. *N Engl J Med*. 2006;354(6):567–578. doi:10.1056/NEJMoa053422
3. Park YM, Kim HR, Cho BC, Keum KC, Cho NH, Kim SH. Transoral robotic surgery-based therapy in patients with stage III-IV oropharyngeal squamous cell carcinoma. *Oral Oncol*. 2017;75:16–21. doi:10.1016/j.oraloncology.2017.10.014
4. Fakhry C, Zhang Q, Nguyen-Tan PF, et al. Human papillomavirus and overall survival after progression of oropharyngeal squamous cell carcinoma. *J Clin Oncol*. 2014;32(30):3365–3373. doi:10.1200/JCO.2014.55.1937
5. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006. doi:10.1038/ncomms5006
6. Liu Z, Wang S, Dong D, et al. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. *Theranostics*. 2019;9(5):1303-1322. doi:10.7150/thno.30309
7. Zhai TT, van Dijk LV, Huang BT, et al. Improving the prediction of overall survival for head and neck cancer patients using image biomarkers in combination with clinical parameters. *Radiother Oncol*. 2017;124(2):256-262. doi:10.1016/j.radonc.2017.07.013
8. Qiang M, Li C, Sun Y, et al. A prognostic predictive system based on deep learning for locoregionally advanced nasopharyngeal carcinoma. *J Natl Cancer Inst*. 2020;113(5):606-615. doi:10.1093/jnci/djaa149
9. Farhidzadeh H, Kim JY, Scott JG, Goldgof DB, Hall LO, Harrison LB. Classification of progression free survival with nasopharyngeal carcinoma tumors. *SPIE*. 2016;9785. doi:10.1117/12.2216976
10. Jethanandani A, Lin TA, Volpe S, et al. Exploring applications of radiomics in Magnetic Resonance Imaging of head and neck Cancer: A systematic review. *Front Oncol*. 2018;8:131. doi:10.3389/fonc.2018.00131
11. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104–e107. doi:10.1158/0008-5472.CAN-17-0339
12. Kohavi R, John GH. Wrappers for feature subset selection. *Artif Intell*. 1997;97:273–324.
13. Bergstra J, Yamins D, Cox D. Hyperopt: a python library for optimizing the

hyperparameters of machine learning algorithms. *Proc SciPy.* 2013;13–20. doi:10.1088/1749-4699/8/1/014008

14. Mes SW, van Velden FHP, Peltenburg B, et al. Outcome prediction of head and neck squamous cell carcinoma by mri radiomic signatures. *Eur Radiol*. 2020;30(11):6311–6321. doi:10.1007/s00330-020-06962-y
15. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean J Radiol*. 2019;20(7):1124–1137. doi:10.3348/kjr.2018.0070
16. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys*. 2018;102(4):1143–1158. doi:10.1016/j.ijrobp.2018.05.053
17. Carré A, Klausner G, Edjlali M, et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Sci Rep*. 2020;10(1):12340. doi:10.1038/s41598-020-69298-z
18. Ng SH, Liao CT, Lin CY, et al. Dynamic contrast-enhanced MRI, diffusion- weighted MRI and 18F-FDG PET/CT for the prediction of survival in oropharyngeal or hypopharyngeal squamous cell carcinoma treated with chemoradiation. *Eur Radiol*. 2016;26:4162–4172. doi:10.1007/s00330-016-4276-8

External validation of an MR-based radiomic model predictive of locoregional control in oropharyngeal cancer

Submitted for publication

Paula Bos
Roland M. Martens
Pim de Graaf
Bas Jasperse
Joost J.M. van Griethuysen
Ronald Boellaard
C. Rene Leemans
Regina G.H. Beets-Tan
Mark A. van de Wiel
Michiel W.M. van den Brekel
Jonas A. Castelijns

4

ABSTRACT

Objectives: To externally validate a pre-treatment MR-based radiomics model predictive of locoregional control in oropharyngeal squamous cell carcinoma (OPSCC) and to assess the impact of differences between datasets on the predictive performance.

Methods: Radiomic features, as defined in our previously published radiomic model, were extracted from the primary tumor volumes of 157 OPSCC patients in a different institute. The developed radiomic model was validated using this cohort. Additionally, parameters influencing performance, such as patient subgroups, MRI acquisition and post-processing steps on prediction performance will be investigated. For this analysis, matched subgroups (based on human papillomavirus (HPV) status of the tumor, T-stage and tumor subsite) and a subgroup with only patients with 4mm slice thickness was studied. Also the influence of harmonization techniques (ComBat harmonization, quantile normalization) and the impact of feature stability across observers and centers was studied. Model performances were assessed by area under the curve (AUC), sensitivity and specificity.

Results: Performance of the published model (AUC/Sensitivity/Specificity: 0.74/0.75/0.60) drops when applied on the validation cohort (AUC/Sensitivity/Specificity: 0.64/0.68/0.60). The performance of the fully validation cohort improves slightly when the model is validated using a patient group with comparable HPV status of the tumor (AUC/Sensitivity/Specificity: 0.68/0.74/0.60), using patients acquired with a slice thickness of 4mm (AUC/Sensitivity/Specificity: 0.67/0.73/0.57) or when quantile harmonization was performed (AUC/Sensitivity/Specificity: 0.66/0.69/0.60).

Conclusion: The previously published model shows its generalizability and can be applied on data acquired from different vendors and protocols. Harmonization techniques as well as subgroup definition influence performance of predictive radiomics models.

INTRODUCTION

The use of imaging biomarkers to enhance diagnostic accuracy and treatment decision-making is fully in development. Radiomics is a noninvasive quantitative image analysis technique to extract large numbers of imaging biomarkers. Several studies showed the potential of radiomics in supporting the radiologist by tumor type determination[1], tumor classification[2,3] or treatment outcome and prognosis prediction[4–6]. However, while these more objective quantitative approaches are very promising and allow another layer of information extraction, traditional visual analysis is still daily routine.

One of the concerns for clinical implementation of radiomics is the generalizability and robustness[7–11]. Prediction models can only be applied in a clinical setting when they are generalizable and show reproducible performance against variations in the radiomic workflow. Such evaluations lack in monocenter studies. Several lines of evidence indicate that prediction models are mainly suitable for the trained patient population[10,11], that radiomic features are influenced by image acquisition parameters[12–14] and post-processing steps, such as tumor segmentation and data harmonization[15,16]. These factors will affect general performance in external validation studies.

To date, multi-center validation of radiomics models in head and neck cancer patients is limited to only five studies[4,5,17–19]. Among these, in only one study[5] a magnetic resonance imaging (MRI)-based radiomic signature was described, which is often the modality of choice for imaging of head and neck tumors due to the superior soft tissue contrast. Mes et al.[5] reported comparable performance during external validation for overall survival (AUC: both 0.69) and relapse-free survival (AUC: 0.63 vs 0.70). In contrast to this multicenter validation study, mostly monocenter MR-based studies were performed[3,20,21].

Recently, an MR-based radiomic model predictive of locoregional control (LRC) in oropharyngeal squamous cell carcinoma (OPSCC) patients was published (Train AUC: 0.0.783, Test AUC: 0.740)[6]. The current research aims to investigate the generalizability of this published model by validating the results using an independent external dataset. Additionally, the influence of differences in the factors 1) patient population, 2) MRI acquisition and 3) post-processing steps on prediction performance will be investigated.

MATERIALS AND METHODS

Study population

The validation cohort consist of a subset of patients collected for earlier published research[22,23]. In more detail, patients with primary histological proven OPSCC were consecutively collected at the Amsterdam UMC. Written informed-consent was obtained from all patients. Patients were treated with (chemo)-radiotherapy ((C)RT) between 2012 and 2018 and had an available pretreatment MRI. Subjects with insufficient image quality were excluded from analysis. Treatment consisted of pre-determined radiotherapy (7 weeks, 70 Gy in 35 fractions) with/without concomitant-chemotherapy (3-weekly 100mg/m2 cisplatin), or weekly cetuximab (400mg/m2 loading-dose followed by seven weekly infusions of 250mg/m^2). HPV-status was determined by p16-immunostaining followed by DNA-PCR on p16-immuno-positive cases. The clinical variables age, gender and smoking status were collected from patient records. Additionally, tumor variables, such as tumor subsite, TNM-stage and HPV status, were collected. Locoregional control (LRC) was defined as the absence of a histopathological proven local recurrence and/or lymph node metastases within 2 years after the end of treatment.

Image acquisition

Pretreatment contrast-enhanced T1-weighted magnetic resonance images (MRI) were acquired on a 1.5 Tesla Signa HDxt MR scanner (GE Medical Systems) (n=54) or 3.0 Tesla Ingenuity MR scanner (Philips Medical Systems) (n=72). MR examinations were performed using a slice thickness ranging from 4 to 7mm and a pixel spacing of 0.40-0.56mm. A flip angle of 90°, echo time of 8.6-16.0mm and a repetition time of 400-820ms were used during MR acquisition. Table 1 summarizes the acquisition parameters used at both centers.

Tumor delineation

The VELOCITY-software was used to manually delineate primary tumors on T1-weighted images by two independent head and neck radiologists with more than 10 (RB) and 30 (JC) years' experience. The 3D whole tumor volumes included necrotic and cystic areas; image artefacts were excluded. Observers were able to use other available MR imaging sequences and clinical information, with the exception of treatment outcome. Discrepancies in tumor segmentations between the observers were solved in a consensus meeting.

Feature extraction

Radiomic features were extracted from the primary tumor volumes using PyRadiomics (version 2.2.0)[24], with the same methodology as reported by Bos et al.[6]. In summary, MR images were normalized, resampled to 1mm^3 isotropic voxels and discretized into bins with a fixed width of five. Features extracted from the Amsterdam UMC were used to 1) validate the published radiomic model[6] and 2) to evaluate factors influencing prediction performance.

Table 1. An overview of the parameters used during the acquisition of MR images for the original dataset, used for building the model, and the validation dataset.

	Training & test dataset	**Validation dataset**
	The Netherlands Cancer Institute (n=177)	**Amsterdam UMC (n=157)**
Manufacturer		
Signa HDxt (GE Medical Systems)	-	54 (34%)
Ingenuity (Philips Medical Systems)	-	72 (46%)
Achieva (Philips Medical Systems)	177 (100%)	-
Unknown	-	31 (20%)
Magnetic field strength [Tesla]		
1.5	82 (46%)	80 (51%)
3.0	95 (54%)	77 (49%)
Acquisition	3D	2D
Slice thickness [mm]	0.8 - 1.0	4.0 - 7.0
Pixel spacing [mm]	0.2 – 1.0	0.4 – 0.6
Repetition time [ms]	4.3 – 10.0	4.0 – 8.2
Echo time [ms]	1.7 – 4.6	8.6 – 16.0
Flip Angle [°]	10	90
Fat suppression	Yes	None

1. Validating the prediction model

A previously developed logistic regression model predictive of LRC[6], based on monocenter data (n=177 OPSCC patients) of the Netherlands Cancer Institute, was validated using external data from the Amsterdam UMC. In summary, the previously developed model[6] is based on radiomic features extracted from pre-treatment T1-weighted postcontrast MR examinations (normalized (centering at zero mean and one standard deviation), discretized and interpolated to isotropic voxels of 1.0 mm^3). After dimensionality reduction, only 77 'stable' features remain for model training using 1000 iterations of Bayesian hyperparameter optimization and fourfold cross validation. The optimal hyperparameters were validated using the unseen test set. The final model included four features (*ClusterShade, Mean, Kurtosis and Sphericity*). Performance was measured using area under the curve

(AUC). AUCs for the predictive performance were 0.78 for the training subset and 0.74 for the test subset.

2. Evaluation of factors influencing prediction performance

Performance obtained by external validation might differ from performance received after internal testing. As mentioned, diversity in performance might be the result of differences in 1) patient demographics, 2) MRI acquisition, and, 3) post-processing. The impact of each of these categories will be evaluated.

2.1 Patient demographics

Matched subgroups (based on clinical variables that differed significantly between the datasets) were randomly created from the validation cohort, which meets the patient demographics of the Netherlands Cancer Institute. Performance of each subgroup was evaluated. Analysis was repeated ten times to correct for selection bias.

2.2 Image acquisition

To compare for differences in image acquisition, model validation was performed on a subset of patients which were acquired using a slice thickness of 4mm. This thickness corresponds best to the slice thickness used at the Netherlands Cancer Institute.

2.3 Post-processing steps

A final factor that might impact prediction performance is post-processing steps. While the radiomic pipeline is becoming more and more standardized[16], there are subtle differences within each step of the pipeline. This experiment evaluates the influence of delineation differences, feature stability and data harmonization on prediction performance.

2.3.1 Agreement on tumor delineation

To examine differences in tumor delineation, ten randomly selected patients of the Amsterdam UMC were delineated by an observer from the Netherlands Cancer Institute. Spatial overlap between the delineations of an observer from the Netherlands Cancer Institute and an observer from the Amsterdam UMC was calculated using the Dice similarity coefficient (DSC)[25] and Hausdorff distance (HD)[26]. A DSC above 0.6 is considered appropriate, where an HD value close to 0mm represents good spatial overlap. Additionally, intraobserver correlation coefficients (ICC) between features extracted from both observers were calculated, considering an ICC above 70% as appropriate.

2.3.2 Data harmonization

An important step during external validation is the correction of data variations across centers by removing batch effects. For this study, the two approaches Combining Batches (ComBat) harmonization and quantile normalization, were evaluated. Both approaches harmonize in the feature domain, where features derived from the validation dataset were harmonized towards the feature domain of the Netherlands Cancer Institute. ComBat harmonization[27] performs location-scaling using Bayes estimations to transform each radiomic feature to a comparable data distribution, resulting in a similar mean and variance in both datasets. ComBat harmonization was performed using the Python package neuroCombat[28]. Quantile normalization[29] discretized each radiomic feature into bins with equal frequencies as the reference data (the Netherlands Cancer Institute), where quantiles of the data distribution were used to determine the points for the bins. Code for quantile normalization was written in R.

2.3.3 Feature stability

Since radiomic features are calculated from the image itself, variations in acquisition parameters affect feature values. Therefore, disparity in feature values extracted from both centers were tested using the Mann-Whitney U test. Radiomic features were considered to be stable, when the p-value was above 0.05.

Statistics

Differences between patient demographics of both cohorts were assessed using Fishers' exact test (binary variables) and independent Student t-test (continuous variables). Performance of the prediction models was depicted using median AUC, sensitivity, specificity and accuracy using 500 iterations of bootstrap (with replacement). Histograms were plotted to visualize data distributions of both centers to examine harmonization performances. Significant differences between AUC performance was calculated using McNeil test[30].

RESULTS

Study population

A total of 157 OPSCC patients were selected for the validation population with a median age of 61 years [IQR: 56-67 years]. In general, most of the patients were male (71%), smokers (80%), had HPV negative tumors (69%), a high T-stage (T3+T4, 67%) with positive nodal disease (82%).

Patient characteristics of the Netherlands Cancer Institute and the Amsterdam UMC are summarized in Table 2. The Amsterdam UMC included more patients with HPV negative tumors (44% vs 69%, p=0.001) and higher T-stages (47% vs 67%, p<0.001). Additionally, significant differences were observed with regard to cancer subsite (p<0.033), as cohort tumors at the posterior wall were more common at the Amsterdam UMC (p<0.001). For both centers, LRC was achieved in a comparable relative number of patients (80% vs 81%).

Table 2. Patient demographics of the dataset used for building the prediction model (the Netherlands Cancer Institute) and the dataset used for validating this model (the Amsterdam UMC) are summarized. Numbers in brackets represents percentages. Differences between clinical variables were assessed using the Fishers' exact test (^{a}) or independent Student t-test (^{b}).

Cohort	Training & test dataset	Validation dataset	p-value
	The Netherlands Cancer Institute (n=177)	Amsterdam UMC (n=157)	
Age (>60 years)	101 (57)	95 (61)	0.578^{a}
Age, y [IQR]	61 [56-66]	61 [56-67]	0.713^{b}
Sex, n male (%)	111 (63)	112 (71)	0.104^{a}
Smoking, n (%)	134 (76)	125 (80)	0.432^{a}
HPV, n (%)			0.001^{a}
Negative	74 (44)	108 (69)	
Positive	76 (43)	49 (31)	
Unknown	24 (31)	0 (0)	
T-stage, n (%)			<0.001^{a}
T1 + T2	94 (53)	52 (33)	
T3 + T4	83 (47)	105 (67)	
N-stage 0, n (%)	141 (80)	128 (82)	0.681^{a}
Subsite of cancer, n (%)			
Tonsillar tissue	99 (56)	66 (42)	0.012^{a}
Soft palate	18 (10)	6 (3)	0.033^{a}
Base of tongue	56 (32)	66 (42)	0.053^{a}
Posterior wall	4 (2)	19 (12)	<0.001^{a}
Clinical endpoints			
LRC < 2 year, n (%)	144 (81)	149 (80)	
Time to LRF in months, median (IQR)	6 (4-17)	6 (4-11)	

1. Validating the prediction model

The performance of the monocentric trained and tested model predictive of LRC[6] had an AUC of 0.74, sensitivity/specificity of 0.75/0.60 and accuracy of 0.71. This performance drops when externally validated to an AUC of 0.64, sensitivity/specificity of 0.68/0.60 and accuracy of 0.66.

2. Evaluation of the impact on performance obtained from the validation dataset

2.1 Patient demographics

Patient subgroups were created based on T-stage, tumor subsite and HPV status and their combinations. Validation performances are summarized in Table 3. Performance of a single clinical variable correction was highest when a subset was matched on HPV status (AUC/Sensitivity/Specificity/Accuracy: 0.68/0.73/0.67/0.71), followed by tumor subsite (AUC/Sensitivity/Specificity/Accuracy: 0.65/0.69/0.63/0.69) and T-stage (AUC/Sensitivity/Specificity/Accuracy: 0.62/0.70/0.50/0.66). Correcting for two clinical variables show the highest performance when patient groups were corrected on both HPV status and cancer subsite (AUC/Sensitivity/Specificity/Accuracy: 0.66/0.71/0.60/0.69). Matching on all three clinical variables drops AUC to 0.61, increases sensitivity of 0.79, decreases specificity of 0.50 and increases accuracy to 0.74. The different performances are illustrated in Figure 1A.

2.2 Image acquisition

As described in Table 2, slice thickness ranges from 0.8 to 1.0mm and 4 to 7mm for patients from the Netherlands Cancer Institute and the Amsterdam UMC, respectively. Validation of the model using only patients acquired with a slice thickness of 4mm (n=111) results in a slight improvement of performance (AUC/Sensitivity/Specificity/Accuracy: 0.67/0.73/0.57/0.71) (see Figure 1B).

2.3 Post-processing steps

2.3.1 Agreement on tumor delineation

Substantial agreement between the observers (each from another center) was shown when primary tumors were delineated independent, with an average DSC of 0.69 ± 0.11 and HD of 10.6 ± 3.0mm (see Figure 2). Most radiomic features (73.2%) were considered as stable. Out of the four predictive features of the published model[6], two features (*Cluster shade* (Wavelet, LLL), *Sphericity*) were considered as stable, whereas the other features (*Mean* (Wavelet LLH), *Kurtosis* (Laplacian of Gaussian, 2.0mm)) were affected by the observer.

2.3.2 Data harmonization

The predictive features of the published model[6], without and after harmonization

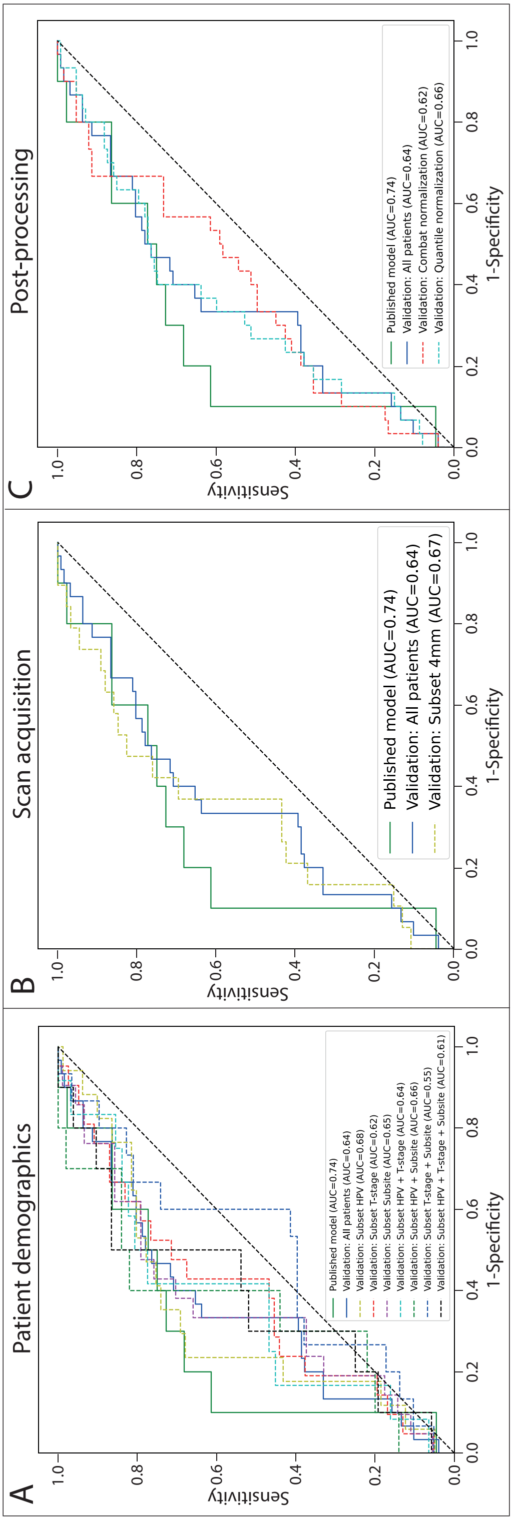


Fig. 1. ROC-curves for the cohort of the patients of the published model (trained and testing cohort (green line)) and validation cohort (blue line). Prediction performance for each subset (dashed lines) of patients (A), scan acquisition (B) and post-processing steps (C) are illustrated. A line closer to the upper left corner represents a better performance of the model in predicting locoregional control of the patients.

are shown in Figure 3. Data distribution after quantile normalization meets the original data better compared to ComBat harmonization. Prediction performance (see Table 3) without harmonization (AUC/Sensitivity/Specificity/Accuracy: 0.64/0.68/0.60/0.66) decreased compared to prediction performance with ComBat harmonization (AUC/Sensitivity/Specificity/Accuracy: 0.62/0.09/1.00/0.26). An improvement was shown when data was harmonized using quantile normalization (AUC/Sensitivity/Specificity/Accuracy: 0.66/0.69/0.60/0.67).

2.3.3 Feature stability

Out of 1,184 features, 83 (7.0%) features were stable against the selected center when no harmonization was applied. Looking at the original radiomic signature[6], '*Sphericity*' is considered as the only stable feature across centers.

Training performance of a monocenter radiomic model, taking only stable features into account decreased significantly (Train AUC: 0.77 vs. 0.58, p=0.02). A

Table 3. Prediction performance for LRC obtained for the different experiments based on 1) patient demographics, 2) scan acquisition or 3) post-processing steps. During experiments considering patient demographics, patient groups were matched for the next clinical parameters HPV status of the tumor (p<0.001), T-stage (p<0.001) and cancer subsite (p<0.053), which were significant different between the two centers. Performance is calculated using area under the curve (AUC), sensitivity, specificity and accuracy. The 95% confidence intervals were measured for experiments based on matching patient groups, where ten randomly selected samples were used for analysis.

Performance		Patients [n]	Validation AUC	Sensitivity	Specificity	Accuracy
Patient intrinsic para-meters	All patients	157	0.64	0.68	0.60	0.66
	Patient subsets					
	HPV	98	0.68 [0.64-0.74]	0.73	0.67	0.71
	T-stage	98	0.62 [0.59-0.65]	0.70	0.50	0.66
	Subsite	112	0.65 [0.61-0.69]	0.69	0.63	0.69
	HPV + T-stage	70	0.64 [0.58-0.74]	0.81	0.50	0.77
	HPV + Subsite	66	0.66 [0.53-0.73]	0.71	0.60	0.69
	T-stage + Subsite	75	0.55 [0.44-0.64]	0.72	0.40	0.66
	HPV + T-stage + Subsite	33	0.61 [0.51-0.66]	0.79	0.50	0.74
Scan acquisi-tion	4mm slices	132	0.67	0.73	0.57	0.71
Post-proces-sing	Harmonization					
	ComBat harmonization	157	0.62	0.09	1.00	0.26
	Quantile normalization	157	0.66	0.69	0.60	0.67

similar trend, but without significance, was shown for test (Test AUC: 0.74 vs 0.53, p=0.10) and validation (Validation AUC: 0.64 vs. 0.51, p=0.11) performance. Figure 4 represents the train, test and validation performances of a model based on all features and based on stable features.

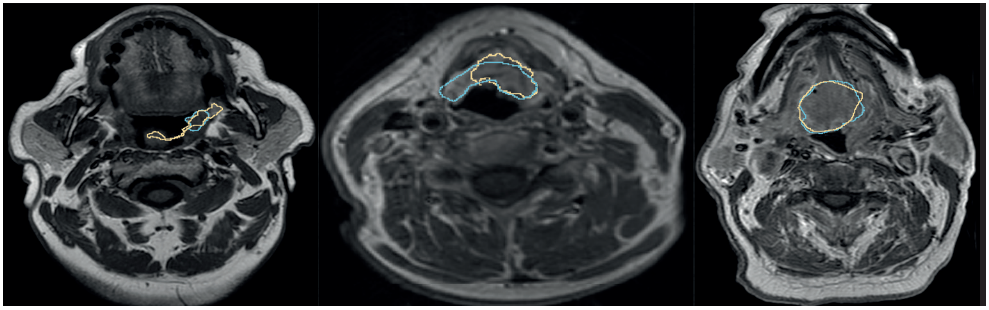


Fig. 2. Spatial overlap between two observers from both institutes visualized, calculated with the dice similarity coefficient (DSC). The patient at the left has a DSC of 0.42, the middle patient a DSC of 0.61, and the patient at the right a DSC of 0.82.

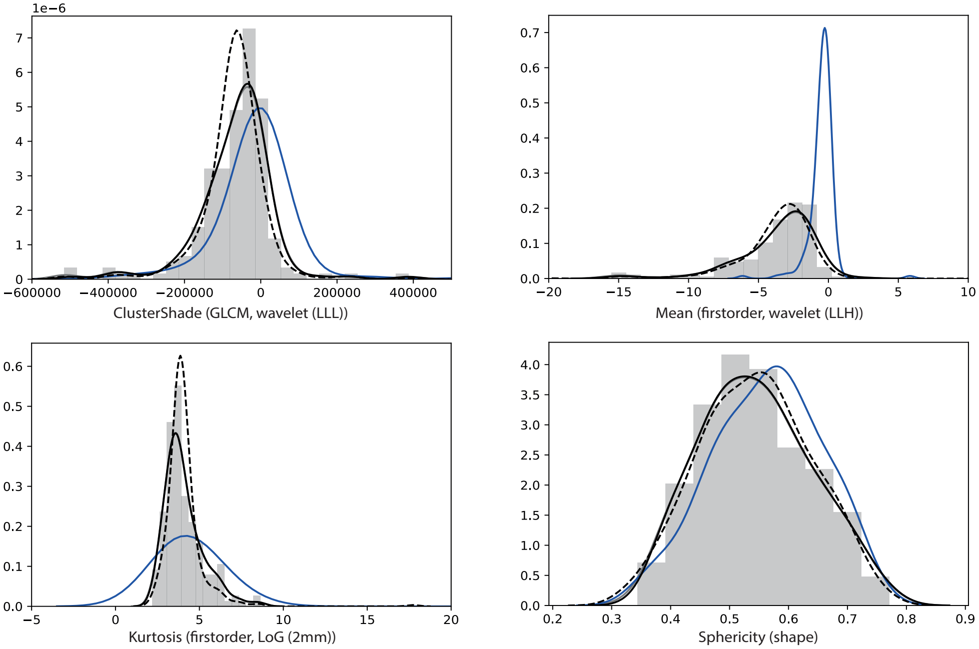


Fig. 3. Histograms of the original radiomic signature, including four features predictive of LRC. The data distribution of the original center (the Netherlands Cancer Institute (grey histogram)) is visualized combined with the data distributions of data from the Amsterdam UMC before (blue line) and after harmonization (Combat harmonization) (dashed black line) or Quantile harmonization (black line).

DISCUSSION

In this study an external validation of a published monocenter pre-treatment MR-based radiomic model predictive of LRC in OPSCC[6] was performed. The main finding is that this model is generalizable and can be applied on data acquired with different vendors and protocols.

Only a slight drop in model performance was observed when the model was validated on an external dataset. A slight drop of prediction performance during external validation was also reported in previous literature[18,19], validating a CT-based radiomic model predictive of HPV (AUC test/validation: 0.83/0.76)[18] or nodal failure (AUC: 0.79/0.71)[19]. Apparently, prediction models developed on internal patient data learned the relation between predictors and the outcome parameter. The systematic differences between patient cohorts and centers makes it harder, but not impossible, to maintain this relation and to classify patients within the correct outcome group, resulting in a slight drop of performance[10].

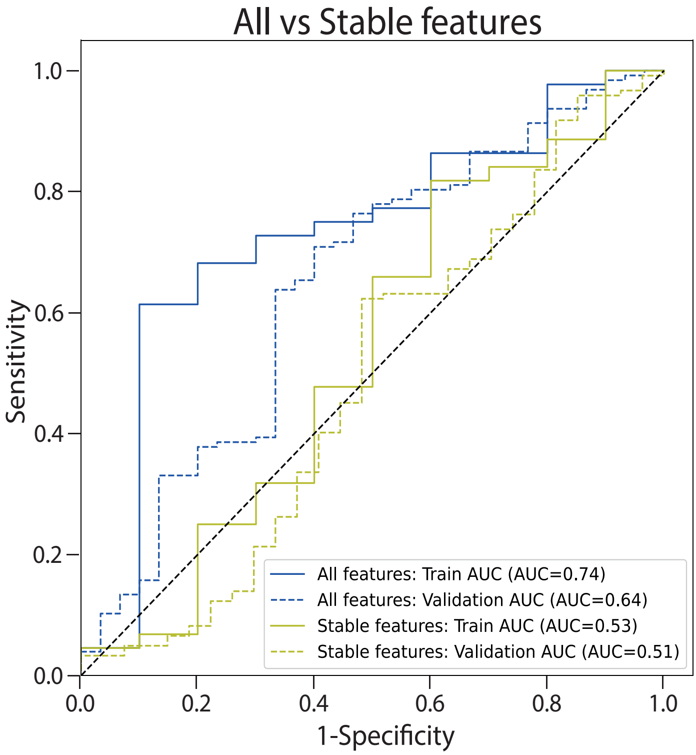


Fig. 4. ROC-curves of a model predictive of locoregional control based on all radiomic features (blue lines) and a model based on stable features (yellow line). The performance of the training and testing cohort (solid line) and validation cohort (dashed line) are visualized. A line closer to the upper left corner represents a better performance of the model.

Another explanation for the drop in performance can be the variation of tumor delineation. Necrotic and cystic areas were excluded during tumor delineation of the center used for model training, whereas the tumor delineations of the validation cohort included these tumor regions. Considering the good spatial overlap in tumor volumes between both observers, the drop has to be caused by the hypointense appearance of necrotic and cystic regions on contrast-enhanced T1-weighted MR images. Since radiomic feature values depend on the voxel intensities, inclusion of these regions affect mainly feature values relying on histogram features[10,12,31,32] such as mean and kurtosis. It is though not surprising that especially these predictors selected in the trained model[6] were unstable in relation to the observer, and therefore, likely relate to the small decrease of prediction performance.

Poorly reproducible features across centers were discarded[12,31], and only 7% (n=83) of all features were considered as being stable. Correcting these features for collinearity, redundancies and removing irrelevant features, resulted in the limited number of only seven features to feed into the model. A low number of features as input reduces model complexity and thereby improves the ability to find optimal model settings, resulting in a better reproducibility[33]. This study shows that a single center model with only seven 'stable' features performed significantly worse compared to the original trained model[6], which was trained on 77 'stable' features. Presumably, the elimination of poorly reproducible features might also remove features associated with LRC, causing a decrease of discriminative power.

Interestingly, validation of performance increases when the model is applied on a patient subset that matches the demographics of the trained patient cohort. Improvement was especially apparent when HPV status of the tumor was matched. Model training was based on an equal distribution of patients with HPV positive and negative tumors[6], whereas in the validation cohort the majority of patients had HPV negative tumors (69%) resulting in less discriminative power. HPV positivity is proven to be associated with better outcome[2,34,35], while this parameter was not included in the radiomic signature due to its clinical behavior. It can be assumed that when HPV is marked as predictor in a model, the prediction of the model is more robust.

Prediction performance also improved when the model was validated on patients acquired on 4mm MR slices. Before extracting feature values, MR images were interpolated to 1mm^3 voxels. Larger transformations were needed when the distance between two acquired slices was larger, introducing a higher uncertainty and more interpolation bias. This bias decreases when a tumor volume is distributed over a larger amount of slices[36]. Sub analysis showed that tumor volumes were larger when acquired on 4mm slices. Large tumor volumes acquired on slices with small distances is the ideal combination to obtain realistic feature values representing biological behavior.

Radiomic features relate on hardware and acquisition protocol are called "center-effects". Equalization of data distributions can be applied to correct for these "center-effects"[37–40]. This study shows that quantile normalization is more robust than ComBat harmonization. This can be explained by the transformation after ComBat which does not completely match the reference distribution of the training cohort, particularly for the features 'Cluster Shade' and 'Kurtosis' (see Figure 3). The higher data kurtosis in the Amsterdam UMC result in dissimilar distributions, a requirement to apply ComBat harmonization. Considering this, discrimination power is reduced and consequently the model is not capable to predict LRC. Due to this lack of sensitivity, unreliable classifications were made by the model, a big concern for clinical application. Our findings are not in line with the report of another study showing that ComBat outperforms other harmonization methods, such as voxel size or singular value decomposition[41], histogram normalization, pixel resampling or Butterworth filtering[39]. However, a, good comparison of these methods is difficult given the used CT-acquired parameters and the lack of quantile harmonization in these studies. This study is novel in the evaluation of quantile normalization in the radiomic field. While ComBat harmonization has recently been adopted in the radiomic research field, consensus concerning harmonization is still in its infancy, requiring further investigation in the large scala of harmonization techniques.

After the study of Mes et al.[5], this is the second study externally validating a MR-based single-center radiomic model in OPSCC. However, this study evaluates a contrast-enhanced T1-weighted MR prediction model, where Mes et al.[5] focused on T1-weighted MRI. Another strength of the study is that all tumors were delineated independently by two observers at the same institute. Additionally, ten patients were delineated by two observers at both institutes. This made evaluation of variability across observers possible.

This study has several limitations. Firstly, it is important to realize that improvement of prediction performance using quantile normalization is only marginal. This study does not allow optimization of quantile normalization by correcting for clinical covariates. Matching the training and validation datasets before applying harmonization provide a (non-linear) transformation representative for differences across centers, excluding any clinical varieties. This methodology removes also the concern of the creation of subsets with comparable patient demographics, something which is not feasible in a real world situation. Thereby, it is important to keep in mind that the calculated transformation is only valid for centers with comparable scanners. Outcome prediction of a prospective patient from a new center with different scanners requires a new transformation, where a minimum of 50 patients have to be involved to calculate this new quantile transformation. The second limitation of the study is that this study is limited by the individual evaluation of factors influencing prediction performance. Combinations of these factors might improve the model performance (i.e., when a model is build based on stable features across centers determined after data harmonization). A third limitation of the study is the lack of a broader evaluation of the acquisition parameters which might influence prediction performance, like flip angle, echo train length or other functional MRI parameters. Fourthly, radiomic features were extracted from post contrast T1-weighted MRI in both cohorts; however, the acquisition parameters differed across the centers. Besides, a fifth limitation is the variety within the delineated tumor contours with regard to the inclusion/exclusion of necrotic and cystic areas was shown. To date, no consensus is reached yet in the literature. Necrotic tissues have been suggested to be indicative for poor treatment response, but also result in extreme radiomic feature values due to its hypo- and hyperintense aspect. Recently, a study[42] investigated the influence of excluding necrotic tissue in tumor delineation on radiomic analysis based on PET images. At least 65% of the radiomic features show significant differences between both groups, but no statistically significant difference was shown in prediction performance (measured by AUC). However, consistency within the methodology of tumor delineation is recommended to optimize analysis. A final remark has to be made on the comparable number of patients included in both cohorts (n=177 vs n=157). Ideally, the external patient cohort should be 25-40% of the training sample[43]. A model trained and tested on larger datasets becomes more robust against heterogeneity across patient selection, image acquisition and post-processing steps, which might fade out some of the described factors influencing model performance.

CONCLUSION

This study shows that our previously published radiomic model predictive of LRC in OPSCC patients is generalizable across centers and can be applied on data acquired from different vendors and protocols. Prediction performance increased when adequate (quantile) harmonization was applied, patient groups were matched for comparable demographics and the acquisition protocol was adapted towards the protocol used during model training.

REFERENCES

1. Fruehwald-Pallamar J, Hesselink JR, Mafee MF, Holzer-Fruehwald L, Czerny C, Mayerhoefer ME. Texture-Based analysis of 100 MR examinations of head and neck tumors - Is it possible to discriminate between benign and malignant masses in a multicenter trial? *Rofo*. 2016;188(2):195-202. doi:10.1055/s-0041-106066
2. Bos P, van den Brekel MWM, Gouw ZAR, et al. Clinical variables and magnetic resonance imaging-based radiomics predict human papillomavirus status of oropharyngeal cancer. *Head Neck*. 2021;43(2):485–495. doi:10.1002/hed.26505
3. Yu Y, He Z, Ouyang J, et al. Magnetic resonance imaging radiomics predicts preoperative axillary lymph node metastasis to support surgical decisions and is associated with tumor microenvironment in invasive breast cancer: A machine learning, multicenter study. *EBioMedicine*. 2021;69:103460. doi:10.1016/j.ebiom.2021.103460
4. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006. doi:10.1038/ncomms5006
5. Mes SW, van Velden FHP, Peltenburg B, et al. Outcome prediction of head and neck squamous cell carcinoma by mri radiomic signatures. *Eur Radiol*. 2020;30(11):6311–6321. doi:10.1007/s00330-020-06962-y
6. Bos P, van den Brekel MWM, Gouw ZAR, et al. Improved outcome prediction of oropharyngeal cancer by combining clinical and MRI features in machine learning models. *Eur J Radiol*. 2021;139:109701. doi:10.1016/j.ejrad.2021.109701
7. Guha A, Connor S, Anjari M, et al. Radiomic analysis for response assessment in advanced head and neck cancers, a distant dream or an inevitable reality? A systematic review of the current level of evidence. *Br J Radiol*. 2020;93(1106):20190496. doi:10.1259/bjr.20190496
8. Granzier RWY, Ibrahim A, Primakov SP, et al. Mri-based radiomics analysis for the pretreatment prediction of pathologic complete tumor response to neoadjuvant systemic therapy in breast cancer patients: A multicenter study. *Cancers*. 2021;13(10):2447. doi:10.3390/cancers13102447
9. Song J, Yin Y, Wang H, Chang Z, Liu Z, Cui L. A review of original articles published in the emerging field of radiomics. *Eur J Radiol*. 2020;127:108991. doi:10.1016/j.ejrad.2020.108991
10. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean J Radiol*. 2019;20(7):1124–1137. doi:10.3348/kjr.2018.0070
11. Jethanandani A, Lin TA, Volpe S, et al. Exploring applications of radiomics in Magnetic Resonance Imaging of head and neck Cancer: A systematic review. *Front Oncol*. 2018;8:131. doi:10.3389/fonc.2018.00131

12. Rai R, Holloway LC, Brink C, et al. Multicenter evaluation of MRI-based radiomic features: A phantom study. *Med Phys*. 2020;47(7):3054-3063. doi:10.1002/mp.14173
13. Mayerhoefer ME, Szomolanyi P, Jirak D, Materka A, Trattnig S. Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: An application-oriented study. *Med Phys*. 2009;36(4):1236–1243. doi:10.1118/1.3081408
14. Wahid KA, He R, McDonald BA, et al. Intensity standardization methods in magnetic resonance imaging of head and neck cancer. *Phys Imag Radiat Oncol*. 2021;20:88-93. doi:10.1016/j.phro.2021.11.001
15. van Velden FHP, Kramer GM, Frings V, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. *Mol Imaging Biol*. 2016;18(5):788–795. doi:10.1007/s11307-016-0940-2
16. 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(5):e0251147. doi:10.1371/journal.pone.0251147
17. Leijenaar RTH, Carvalho S, Hoebers FJP, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncol*. 2015;54(9):1423–1429. doi:10.3109/0284186X.2015.1061214
18. Leijenaar RTH, Bogowicz M, Jochems A, 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. doi:10.1259/bjr.20170498
19. Zhai TT, Wesseling F, Langendijk JA, et al. External validation of nodal failure prediction models including radiomics in head and neck cancer. *Oral Oncol*. 2021;112:105083. doi:10.1016/j.oraloncology.2020.105083
20. Romeo V, Cuocolo R, Ricciardi C, et al. Prediction of tumor grade and nodal status in oropharyngeal and oral cavity squamous-cell carcinoma using a radiomic approach. *Anticancer Res*. 2020;40(1):271-280. doi:10.21873/anticanres.13949
21. Yuan Y, Ren J, Shi Y, Tao X. MRI-based radiomic signature as predictive marker for patients with head and neck squamous cell carcinoma. *Eur J Radiol*. 2019;117:193-198. doi:10.1016/j.ejrad.2019.06.019
22. Martens RM, Noij DP, Koopman T, et al. Predictive value of quantitative diffusion-weighted imaging and 18-F-FDG-PET in head and neck squamous cell carcinoma treated by (chemo)radiotherapy. *Eur J Radiol*. 2019;113:39-50. doi:10.1016/j.ejrad.2019.01.031
23. Martens RM, Koopman T, Lavini C, et al. Multiparametric functional MRI and 18F-FDG-PET for survival prediction in patients with head and neck squamous cell carcinoma treated with (chemo)radiation. *Eur Radiol*. 2021;31(2):616–628. doi:10.1007/s00330-020-07163-3

24. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104–e107. doi:10.1158/0008-5472.CAN-17-0339
25. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945;26(3):297–302. doi:10.2307/1932409
26. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. *IEEE Trans Pattern analysis and machine intelligence*. 1993;15(9):850-863. doi:10.1109/CVPR.1992.223209
27. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–127. doi:10.1093/biostatistics/kxj037
28. Fortin JP, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*. 2019;167:104-120. doi:10.1016/j.neuroimage.2017.11.024
29. Warnat P, Eils R, Brors B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*. 2005;6:265. doi:10.1186/1471-2105-6-265
30. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36. doi:10.1148/radiology.143.1.7063747
31. Baeßler B, Weiss K, Dos Santos DP. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Invest Radiol*. 2019;54(4):221–228. doi:10.1097/RLI.0000000000000530
32. Galizia MS, Töre HG, Chalian H, McCarthy R, Salem R, Yaghmai V. MDCT Necrosis Quantification in the Assessment of Hepatocellular Carcinoma Response to Yttrium 90 Radioembolization Therapy. Comparison of Two-dimensional and Volumetric Techniques. *Acad Radiol*. 2012;19(1):48–54. doi:10.1016/j.acra.2011.09.005
33. Zhou Z, Li S, Qin G, Folkert M, Jiang S, Wang J. Multi-objective based radiomic feature selection for lesion malignancy classification. *IEEE J Biomed Health Inform*. 2020;24(1):194-204. doi:10.1109/JBHI.2019.2902298
34. Ang KK, Harris J, Wheeler R, et al. Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer. *N Engl J Med*. 2010;363(1):24-35. doi:10.1056/NEJMoa0912217
35. Fakhry C, Zhang Q, Nguyen-Tan PF, et al. Human papillomavirus and overall survival after progression of oropharyngeal squamous cell carcinoma. *J Clin Oncol*. 2014;32(30):3365–3373. doi:10.1200/JCO.2014.55.1937
36. Park SH, Lim H, Bae BK, 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. doi:10.1186/s40644-021-00388-5
37. Papadimitroulas P, Brocki L, Chung NC, et al. Artificial intelligence: Deep learning

in oncological radiomics and challenges of interpretability and data harmonization. *Phys Medica*. 2021;83:108–121. doi:10.1016/j.ejmp.2021.03.009

38. Masson I, Da-ano R, Lucia F, et al. Statistical harmonization can improve the development of a multicenter CT based radiomic model predictive of non-response to induction chemotherapy in laryngeal cancers. *Med Phys*. 2021;48(7):4099:4109. doi:10.1002/mp.14948
39. Foy JJ, Al-Hallaq HA, Grekoski V, et al. Harmonization of radiomic feature variability resulting from differences in CT image acquisition and reconstruction: Assessment in a cadaveric liver. *Phys Med Biol*. 2020;65(20):205008. doi:10.1088/1361-6560/abb172
40. Da-Ano R, Visvikis D, Hatt M. Harmonization strategies for multicenter radiomics investigations. *Phys Med Biol*. 2020;65(24):24TR02. doi:10.1088/1361-6560/aba798
41. Ligero M, Jordi-Ollero O, Bernatowicz K, et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *Eur Radiol*. 2021;31(3):1460–1470. doi:10.1007/s00330-020-07174-0
42. Noortman WA, Vriens D, Mooij CDY, et al. The influence of the exclusion of central necrosis on [18F]FDG PET radiomic analysis. *Diagnostic*. 2021;11:1296. doi:10.3390/diagnostics11071296
43. Papanikolaou N, Matos C, Koh DM. How to develop a meaningful radiomic signature for clinical use in oncologic patients. *Cancer Imaging*. 2020;20:33. doi.1.1186/s40644-020-00311-4

Clinical variables and magnetic resonance imaging-based radiomics predict human papillomavirus status of oropharyngeal cancer

Head Neck, Oct 2020;43(2);485-495
DOI: 10.1002/hed.26505

Paula Bos
Michiel W.M. van den Brekel
Zeno A.R. Gouw
Abrahim Al-Mamgani
Marjaneh Taghavi
Selam Waktola
Hugo J.W.L. Aerts
Jonas A. Castelijns
Regina G.H. Beets-Tan
Bas Jasperse

5

ABSTRACT

Background: Human papillomavirus (HPV)-positive oropharyngeal squamous cell carcinoma (OPSCC) have better prognosis and treatment response compared to HPV-negative OPSCC. This study aims to noninvasively predict HPV status of OPSCC using clinical and/or radiological variables.

Methods: Seventy-seven magnetic resonance radiomic features were extracted from T1-weighted postcontrast images of the primary tumor of 153 patients. Logistic regression models were created to predict HPV status, determined with immunohistochemistry, based on clinical variables, radiomic features, and its combination. Model performance was evaluated using area under the curve (AUC).

Results: Model performance showed AUCs of 0.794, 0.764, and 0.871 for the clinical, radiomic, and combined models, respectively. Smoking, higher T-classification (T3 and T4), larger, less round, and heterogeneous tumors were associated with HPV-negative tumors.

Conclusion: Models based on clinical variables and/or radiomic tumor features can predict HPV status in OPSCC patients with good performance and can be considered when HPV testing is not availabl*e.*

INTRODUCTION

Human papillomavirus (HPV) infection is an important factor in the development and disease course of oropharyngeal squamous cell carcinoma (OPSCC)[1,2]. HPV-related OPSCC has a better progression-free survival and overall survival after (chemo)radiation treatment than HPV-negative OPSCC[3-5]. Despite these differences in prognosis and treatment response, HPV-positive and HPV-negative OPSCC are currently not treated differently. Only recently, it was shown that cetuximab cannot replace cisplatin in HPV-positive OPSCC[6]. Ongoing de-escalation trials will further elucidate whether HPV-positive tumors can be treated with less aggressive treatment regimens in the future to reduce treatment-related toxicity (trial number NCT03952585). This is especially relevant as HPV-positive OPSCC patients tend to be younger with an associated higher life expectancy than HPV-negative OPSCC patients[5,7,8]. Adding to the importance of HPV status of OPSCC is the increasing relative incidence of HPV-positive OPSCC compared to HPV-negative OPSCC over the past years despite declining overall age adjusted incidence of head and neck cancer in developed countries. These changes are probably due to a decline in alcohol and especially nicotine abuse combined with an increase in sexual promiscuity with a high risk of HPV transmission[9]. For these reasons, HPV tumor status is increasingly important and has therefore been included in the most recent eighth edition of the TMN classification[10].

HPV infection is detected using p16/p53 immunohistochemistry and/or HPV DNA polymerase chain reaction (PCR) on biopsy material[11,12]. Determination of tumor HPV status from just clinical and/or tumor features extracted from imaging would be ideal, and could possibly reduce the need for time consuming and expensive immunochemistry and PCR techniques. Recent literature showed that tumor biology can be assessed noninvasively in other tumor types using advanced imaging analysis or radiomics[13,14]. The same approach may be used to determine predictive features for the HPV status in OPSCC. Multiple studies reported that the CT-based radiomic features, such as shape and homogeneity, are associated with HPV positivity in OPSCC tumors[15-17]. To our knowledge, MRI-based radiomics to predict HPV status has not been performed previously. Clinical variables associated with HPV-positive tumors are well known and include male gender, younger age, and less exposure to tobacco and alcohol[9]. These variables have been used to predict HPV status of head and neck cancer, including OPSCC[18-21].

This study aims to assess and compare the ability of clinical variables, MR-based radiomic features, or a combination of these variables to predict HPV status of OPSCC.

MATERIALS AND METHODS

This study is approved by the local institutional review board (IRBd18047). Due to the retrospective design, informed consent was waived.

Clinical data

A total of 240 consecutive patients with histologically proven primary OPSCC, treated with CRT (70 Gy radiation with three planned cycles cisplatin-based chemotherapy [100 g/m^2]) at our Institute between January 2010 and December 2015, were considered for this study. Patients were excluded when pretreatment MRI of the primary tumor was not available (n=38), image quality was poor (n=7), tumors were undetectable on MRI (n=17), a second head and neck tumor was present (n=1), or when HPV status of the tumor was missing (n=24). This resulted in a total of 153 patients eligible for this study.

Age, gender, smoking status, tumor subsite, and TNM classification (TNM seventh edition), were collected for each patient. T-classification and N-classification was determined in multidisciplinary consensus based on clinical and radiological information, including MRI, ultrasound staging with fine needle aspiration cytology, and, when available, PET images. Smoking status was classified into the categories nonsmoker, current smoker, and former smoker (quit more than 2 years prior to diagnosis) at the initial visit to the outpatient clinic. T-classification was dichotomized in low (T1+T2) or high T-classification (T3+T4). N-classification was dichotomized in node-positive (N >0) or node-negative disease (N=0). Differences in clinical variables between HPV-positive and HPV-negative tumors were assessed by applying the Fisher exact test and independent t-test for age. P values of <.05 were considered statistically significant.

Determination of HPV tumor status

A combination of p16 and p53 immunohistochemistry on tumor biopsy material was performed to determine HPV positivity or negativity of the tumor for each patient. p53 positivity was concluded when at least 80% of the tissue sample showed strong nuclear staining or completely negative. No p53 staining of tumor tissue with positive staining of surrounding normal tissue was regarded as tumor mutation for which p53 positivity was concluded. p16 positivity was concluded when at least 70% of tumor tissue stained positive for p16. A known HPV-positive tonsil sample, surrounding tissue of the tested biopsy sample and appendix, was used as positive internal and external control. HPV positivity was concluded when tumor biopsy material tested positive for p16 and negative for p53 staining. HPV negativity was concluded when tumor biopsy material tested negative for p16, regardless of p53; see Henneman et al.[22] for further details on the HPV testing scheme.

MRI data

All patients underwent an MRI examination of the primary tumor for pretreatment staging purposes as part of the routine clinical workup. Imaging was performed at 1.5 T or 3.0 T (Achieva, Philips Medical System, Best, The Netherlands) using a standard head and neck coil (SENSE-NV-16). The imaging protocol included T1-weighted (T1W), T2-weighted (T2W), postcontrast 3D T1W, perfusion, and diffusion-weighted sequences. Imaging details are summarized in Table 1 and Supplementary Table S1.1. The axial slices of 3D T1W high-resolution isotropic volume excitation (THRIVE) after gadolinium injection (postcontrast 3DT1W) were used to manually delineate primary tumor volumes. One nonexpert observer (PB, 1 year experience in head and neck diagnosis) manually delineated the tumor volumes (i.e. nonexpert delineations), which were verified and corrected by an experienced head and neck radiologist (BJ, 7 years of experience in head and neck diagnosis) (i.e. expert-corrected delineations). The observers were allowed to review other available pretreatment MR imaging sequences and available PET scans as reference to improve delineations.

Table 1. Postcontrast 3DT1W MRI image acquisition parameters stratified by MRI magnet strength, 1.5 Tesla and 3.0 Tesla.

MRI field strength	**1.5 Tesla** **n=74**	**3.0 Tesla** **n=79**
HPV+	41	35
Slice thickness [mm]	0.8-1.0	0.8
Pixel spacing [mm]	0.4-1.0	0.2-0.8
Repetition time [ms]	9.4-10	4.3 – 5.3
Echo time [ms]	4.6	1.7-2.4
Echo train length	60	90
Flip angle [ᵅ]	10	10

Note: MRI indicates Magnetic Resonance Imaging; HPV, Human papillomavirus; 3DT1W, 3D T1-weighted

Radiomic feature extraction

Signal intensities for each individual MRI scan were normalized (with zero mean and unit SD) prior to further analysis to reduce intensity variations between MRI scans obtained from different patients. Image resampling to isotropic voxels of 1.0 mm was performed using B-spline interpolation. Image discretization was applied to allow quantification of texture images in fixed bin width of five. In total, 1184 radiomic features per patient were calculated from the postcontrast 3DT1W MRI within the primary tumor volumes using the open-source package PyRadiomics 2.2.0[23], which were categorized into the five groups: shape, intensity, texture, wavelet transform, and Laplacian of Gaussian filter. Wavelet features were calculated in seven decompositions and texture coarseness is determined by four levels modifying the Gaussian radius parameter from 0.5 to 2.0 mm, in steps of 0.5 mm. Detailed definitions of the radiomic features can be found elsewhere[28].

After quality control, features with zero variance were excluded. Stable features were selected using the interclass correlation coefficient with regard to the nonexpert and expert-corrected tumor delineations and the MannWhitney U test in features with regard to the different MRI field strengths. Features with an interclass correlation coefficient greater than 0.75 and a significance level equal to or above .05 in the Mann-Whitney U test were considered stable. From the selected stable features, collinear features (Pearson correlation coefficient > 0.9) were removed, where for each pair the feature that has the largest mean absolute correlation is deleted. The remaining 77 features (see consort diagram in Supplementary Figure S1.1) eligible for radiomic analysis were normalized with zero mean and unit variance for analysis.

Machine learning analysis

From the total of 153 patients, 60% (n=91) were randomly allocated to a training/ validation subset and 40% (n=62) to a test subset, stratifying for HPV status and MRI magnet strength (1.5 or 3.0 T).

Then, separate logistic regression models[24] were build based on solely clinical variables (i.e., age, gender, smoking status, T-classification, N-classification, and subsite of cancer) (clinical model), only radiomic features (radiomic model) and a model where both clinical and radiomic features were combined (combined model). As data from other cancer registries may be missing smoking status and/ or TN-classification, we constructed a combined model without smoking status and/or TN-classification (see Supplementary Material II).

Feature dimensionality is reduced by applying a sequential backward wrapper feature selection approach (recursive feature elimination). This method obtains the optimal feature set for the given classifier (in this case logistic regression) by iteratively removing the weakest feature assessed by its feature importance score. The optimal set of features is used to train the model[25,26].

In the training phase, Bayesian optimization was used to obtain optimal hyperparameters employing 1000 iterations of 4-fold cross-validation on a 75% (n=68) training and 25% (n=23) validation set. During this process, the regularization parameter (λ, 0.005-200), a parameter for the complexity of the model, and the number of features (k, 1-77 [radiomic model] or 1-86 [combined model]) were tuned based on the four training performances obtained during cross-validation. Area under the curve (AUC) was calculated as measure of model performance, where the loss function is minimized. The loss function was defined as 1−mean(AUC)+SD(AUC), where mean(AUC) aims to maximize model performance and SD(AUC) aims to minimize model generalization[27-29].

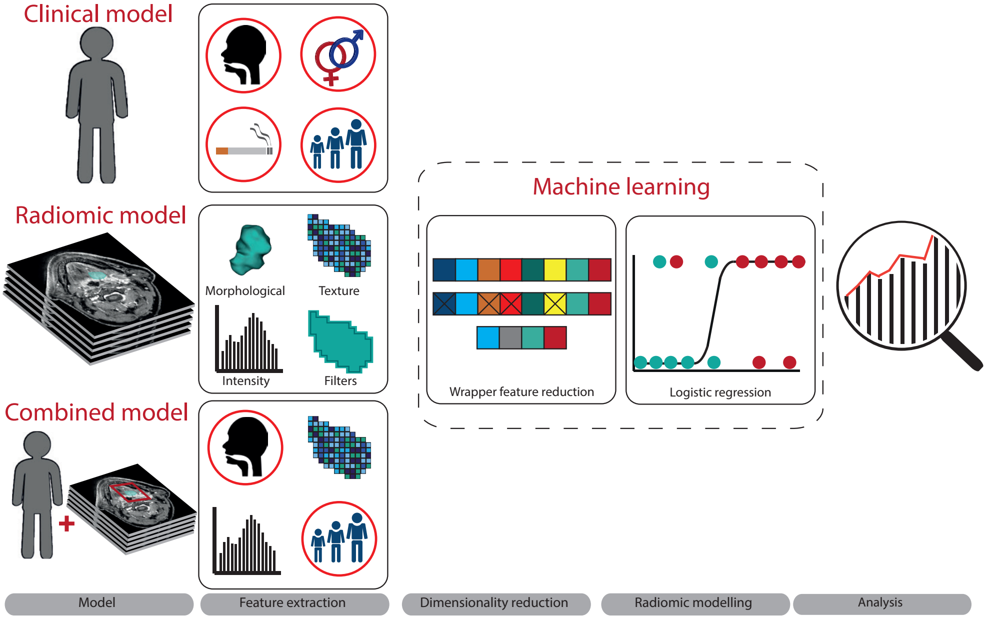


Fig. 1. Analysis pipeline. Three models were created to predict human papillomavirus (HPV) status of oropharyngeal squamous cell carcinoma (OPSCC). A clinical model (based on the clinical variables, age, gender, smoking status, T-classification, N-classification, and tumor subsite), a radiomic model based on radiomic features, and a combined model based on both clinical variables and radiomic features. Morphological, texture, intensity, and filter-based radiomic features were computed from within the tumor delineations on the postcontrast 3DT1 MRI images. Feature reduction was performed using the wrapper feature selection approach by recursive feature elimination, resulting in an optimal subset of features as input for the logistic regression models. The three separate models were created using logistic regression analysis on the training subset. Resulting models were tested using bootstrapping with 500 iterations. Model performance on the test set was evaluated using median area under the curve, sensitivity, accuracy, and its 95% confidence intervals.

The optimized hyperparameters obtained in the training phase were then used to verify the predictive model in the test phase, applying bootstrapping on the test subset. Bootstrapping calculated model performance (AUC) of 500 randomly selected samples (with replacement) of the test subset. Median AUC and the 95% confidence interval (95% CI) of these 500 iterations were then calculated to reflect the model performance that can be attained of HPV prediction. All analyses were implemented in python 3.5 and SPSS version 25.0 (SPSS Inc. Chicago). The complete machine learning pipeline is shown in Figure 1.

A clinically applicable nomogram was constructed from the clinical logistic regression model using R software package RMS (version 3.6.3)[30]. Points were assigned to each prognostic variable from the clinical model based on the distribution of the regression coefficients, maximizing sensitivity and specificity for discrimination between HPV-positive and HPV-negative tumors. The probability of HPV positivity can be deducted from the sum of these points.

RESULTS

Table 2 summarizes patient characteristics for the total patient cohort and subgroups stratified by HPV status. The clinical characteristics of the whole patient group have an equal distribution of HPV (n=77 HPV negative and n=76 HPV-positive tumors) and T status (51% patients have T1+T2 tumors, 49% T3+T4 tumors). Tumors were mostly located in the tonsils. Patients were categorized as either smoking or nonsmoking, no patients were categorized as former smokers.

OPSCC patients with HPV-positive tumors were younger (median age: 63 vs 59 year, P=.007), less likely to smoke (P<.001), and had a lower T-classification (T1-T2 vs T3-T4; P<.001) compared to patients with HPV-negative tumors. For node-positive disease (P=.051) and male gender (P=.067), these differences were borderline significant at the 5% level. Tumors of the soft palate (P=.017) were significantly more frequent in HPV-negative tumors.

Performance of logistic regression models

Performance of the three logistic regression models is summarized in Table 3. All models showed good performance in the prediction of tumor HPV status for the training set (AUC: 0.872-0.923) and test set (AUC 0.764-0.871). Figure 2 shows the receiver-operating characteristic (ROC) curves of the three models. The clinical model (Test AUC: 0.794, Sens: 0.71, Spec: 0.81, PPV: 0.79, NPV: 0.74, Acc: 0.76) performed slightly better than the radiomic model (Test AUC: 0.764, Sens: 0.76, Spec: 0.71, PPV: 0.72, NPV: 0.75, Acc: 0.73). The combined model had the most favorable performance, outperforming the other models (Test AUC: 0.871, Sens: 0.88, Spec: 0.68, PPV: 0.73, NPV: 0.85, Acc: 0.78). Model performance was similar when only smoking status (Test AUC: 0.837) or TNM classification (T0.873) was omitted from model construction, but drops when both clinical variables were omitted (Test AUC: 0.756); see Supplementary Material II for detailed results of the subanalysis.

Table 2. Patient characteristics, for all patients and subgroups stratified by human papillomavirus (HPV)-status of the tumor. The number of patients and its percentage in parentheses is given. Differences between HPV-negative and HPV-positive patient groups, calculated with independent t-test ([a]) or Fishers exact test ([b]), are shown in the last column. Significant values are summarized with an asterisk. Patients were categorized as either smoking or non-smoking, no patients were categorized as former smokers.

Patients	**Total n=153**	**HPV negative n=77**	**HPV positive n=76**	**p-value**
Age, median y [IQR]	61 [56-66]	63 [57-67]	59 [55-65]	0.007[a]*
Male, *n* (%)	96 (63)	54 (70)	42 (55)	0.067[b]
Smoking, *n* (%)	114 (75)	72 (94)	42 (55)	< 0.001[b]*
T-classification, *n* (%)				< 0.001[b]*
T1+T2	78 (51)	25 (32)	53 (70)	
T3+T4	75 (49)	52 (68)	23 (30)	
N-classification (N>0), *n*(%)	127 (83)	59 (77)	68 (89)	0.051[b]
Subsite of cancer, *n* (%)				
Tonsil	88 (58)	42 (55)	46 (60)	0.514[b]
Soft palate	13 (8)	11 (14)	2 (3)	0.017[b]*
Base of tongue	48 (31)	20 (26)	28 (37)	0.166[b]
Posterior wall	4 (3)	4 (5)	0 (0)	0.120[b]

Note: HPV indicates Human Papillomavirus

Selected features of logistic regression models

Table 4 summarizes all prognostic variables selected for the three models with their regression coefficients, SE and odds ratios (OR) (95% CI). Selected features were obtained in the training phase, during the last cross-validation fold, and then used to train the predictive model with the full training dataset. In the clinical model, smoking (OR: 0.47 [0.24-0.91]), node-negative disease (OR: 0.69 [0.33-1.42]), male gender (OR: 0.76 [0.44-1.34]), tumor located on the soft palate (OR: 0.69 [0.04-13.15]), and tumor located on the posterior wall of the oropharynx (OR: 0.80 [0.02-29.97]) were associated with HPV-negative tumors. A low T-classification (OR: 1.70 [0.96-3.03]) and tumor located in the tonsil (OR: 1.24 [0.07-20.73]) was associated with HPV-positive tumors. The clinical model is presented in a nomogram in Figure 3, where a cutoff value of 134 points has the maximum sensitivity (76%) and specificity (73%). A sum of points below 134 is indicative of HPV negativity.

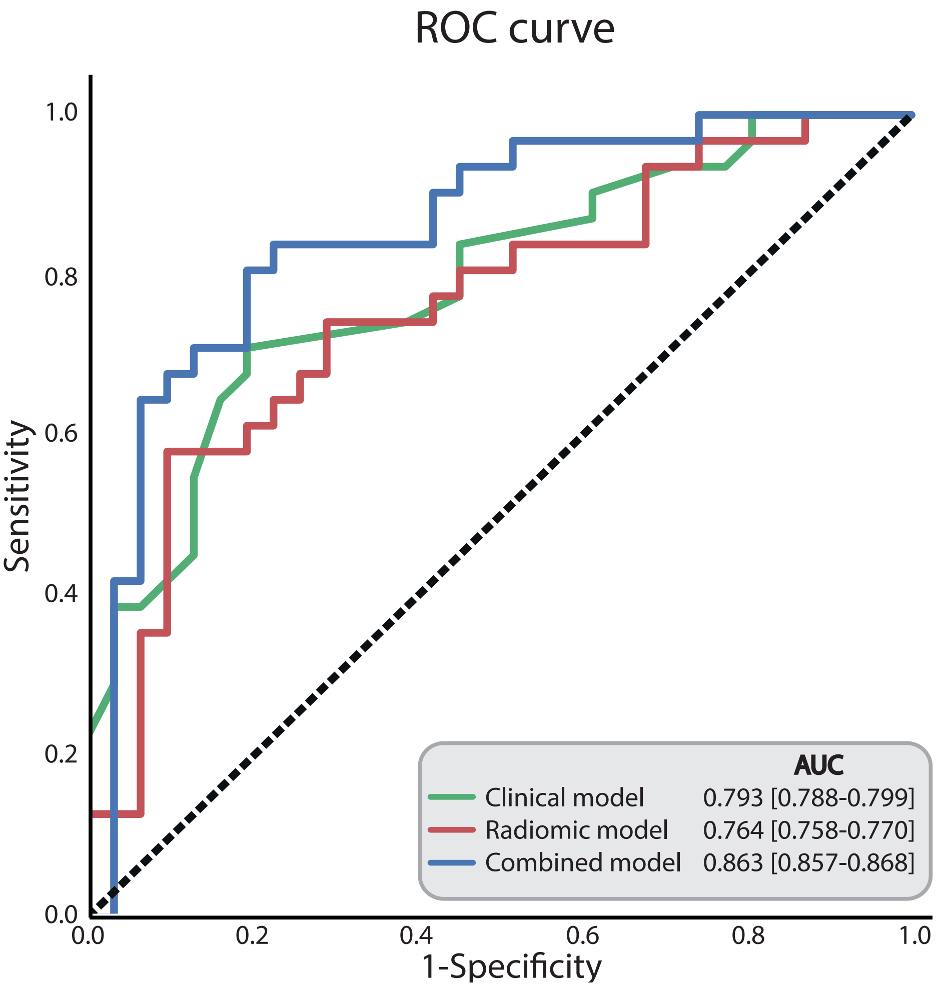


Fig. 2. Receiver-operating characteristic (ROC) curve for prediction of human papillomavirus (HPV) status of the tumor. The combined model had a higher area under the curve (AUC) than the clinical and radiomic model.

Out of the 77 initial radiomic features, three prognostic features were selected in the radiomic model after model construction. Fourteen radiomic features were

Table 3. Model performance of the logistic regression prognostic models for human papillomavirus (HPV)-status. Performance is defined as median AUC with its 95% CI in parenthesis calculated from AUC values of the cross-validation and bootstrapping for the training and test set respectively.

Model	Training AUC [CV]	Test AUC [CI bootstrap]	Sensitivity [CI bootstrap]	Specificity [CI bootstrap]	PPV [CI bootstrap]	NPV [CI bootstrap]	Accuracy [CI bootstrap]
Clinical	0.872 [0.819-0.938]	0.794 [0.788-0.800]	0.71 [0.70-0.72]	0.81 [0.80-0.82]	0.79 [0.78-0.79]	0.74 [0.73-0.75]	0.76 [0.75-0.76]
Radiomic	0.885 [0.826-0.934]	0.764 [0.758-0.770]	0.71 [0.70-0.72]	0.71 [0.70-0.72]	0.72 [0.71-0.73]	0.75 [0.74-0.76]	0.73 [0.73-0.74]
Combined	0.923 [0.868-0.983]	0.871 [0.866-0.876]	0.88 [0.87-0.89]	0.68 [0.67-0.69]	0.73 [0.72-0.74]	0.85 [0.84-0.86]	0.78 [0.77-0.78]

Note: AUC indicates Area under the curve; CI, Confidence Interval; HPV, Human Papillomavirus; PPV, Positive predicted value; NPV, Negative predicted value

selected in the combined model, along with six clinical variables that were included in the clinical model. Radiomic features indicated smaller, rounder, more homogeneous, and more regular texture in HPV-positive tumors. Figure 4 illustrates textural differences between a patient with HPV-negative and HPV-positive tumor. The interpretation of all selected radiomic features is summarized in Supplementary Table S1.2.

DISCUSSION

This retrospective study shows that logistic prediction models based on clinical and/or MR-based radiomic features are able to predict HPV status in OPSCC with good performance. The model combining radiomic features and clinical variables performed better than separate models based on clinical and radiological features.

The variables included in the clinical model were variables that can be expected to differentiate HPV-negative and HPV-positive tumors (ie, smoking status, age, gender, T-classification, N-classification, and tumor location). This underscores that the clinical model, besides the good overall performance, is biologically plausible.

The discriminatory MRI features in the radiomicbased models probably reflect differences in tumor biology between HPV-positive and HPV-negative tumors. HPV-positive tumors are characterized by less-invasive exophytic growth,

Table 4. Selected features in the radiomic and combined model with regression coefficients ranked from high to low, standard errors and odds ratio (OR) (with 95% confidence interval (CI)). Positive regression coefficients or an OR above 1 indicates a higher likelihood of Human Papillomavirus (HPV) positive tumor. Negative coefficients indicate a higher likelihood of HPV negative tumors. * features in the combined model that are also included in the clinical or radiomic model.

Selected feature	**Regression coefficient**	**Standard Error**	**Odds ratio [95% CI]**
Clinical model (n=7)			
Smoking	-0.76	0.17	0.47 [0.24-0.91]
Low T-classification	0.53	0.15	1.70 [0.96-3.03]
Node-negative disease	-0.38	0.19	0.69 [0.33-1.42]
Subsite of cancer: Soft palate	-0.37	0.75	0.69 [0.04-13.15]
Male gender	-0.27	0.14	0.76 [0.44-1.34]
Subsite of cancer: Posterior wall of oropharynx	-0.22	0.92	0.80 [0.02-29.97]
Subsite of cancer: Tonsil	0.21	0.72	1.24 [0.07-20.73]
Radiomic model (n=3)			
Shape Sphericity	0.16	0.90	1.18 [0.03-40.59]
Gray Level Co-occurrence Matrix Inverse Difference Moment (Laplacian of Gaussian (2mm))	0.13	0.11	1.13 [0.73-1.76]
Kurtosis (wavelet)	0.12	0.22	1.13 [0.48-2.67]
Combined model (n=20)			
* Smoking	-0.74	0.44	0.44 [0.09-2.64]
Neighbouring Gray Tone Difference Matrix Busyness (Wavelet) (2x)	-0.39 -0.21	0.88 0.38	0.68 [0.02-21.01] 0.81 [0.18-3.61]
* Node-negative disease	-0.33	0.60	0.72 [0.07-7.53]
Skewness (Wavelet)	-0.33	0.32	0.72 [0.21-2.51]
* Shape Sphericity	0.33	0.46	1.39 [0.23-8.35]
* Gray Level Co-occurrence Matrix Inverse Difference Moment (Laplacian of Gaussian (2mm))	0.30	0.12	1.35 [0.86-2.12]
* Subsite of cancer: Soft palate	-0.30	0.44	0.74 [0.13-4.25]
* Low T-classification	0.29	0.55	1.33 [0.15-11.61]
* Kurtosis (Wavelet) (3x)	0.29 -0.19 -0.18	0.19 0.26 0.38	1.33 [0.64-2.77] 0.83 [0.30-2.26] 0.83 [.019-3.68]

	Neighbouring Gray Tone Difference Matrix Complexity (Wavelet)	-0.26	0.00	0.77 [0.77-0.77]
	Maximum (Wavelet)	-0.23	0.01	0.79 [0.77-0.82]
	Gray Level Co-occurrence Matrix Cluster Prominence (Wavelet)	-0.23	0.00	0.80 [0.80-0.80]
*	Subsite of cancer: Tonsil	0.22	0.34	1.25 [0.33-4.74]
*	Male gender	-0.22	0.50	0.80 [0.11-5.71]
	Neighbouring Gray Tone Difference Matrix Contrast (2x) (Laplacian of Gaussian (0.5mm), Wavelet)	-0.21 -0.18	0.10 0.10	0.81 [0.55-1.20] 0.83 [0.56-1.24]
	Maximum2DDiameter	-0.19	0.04	0.82 [0.71-0.96]

Note: CI indicates Confidence Interval

nonkeratinizing histopathology, genetic stability, and well-defined surroundings[31]. These histopathological differences are likely to be reflected in the selected radiomic features indicating rounder tumors, lower maximum intensity values, and texture homogeneity. Conversely, HPV-negative tumors are genetically more unstable[32], which can lead to focal hypoxia or varying grades of dedifferentiation within a tumor, likely to be reflected in the selected MR features of heterogeneity in the radiomic models.

Although no direct comparison was made, our MR-based predictive radiomic model suggests similar performance (AUC=0.76) compared to CT[16,17]. This suggests that postcontrast 3DT1W MRI and CT reveal, at least partly, similar textural properties relevant for the discrimination of HPV-positive and HPV-negative tumors in radiomic analysis. Intuitively, features from MRI and CT should at least be able to characterize tumor size and morphology in a similar way, explaining similar performance. Whether structural MRI or CT is better for determination of HPV status of OPSCC by radiomic analysis is not entirely clear at this point. In our opinion, MRI is preferable over CT for staging and radiomic analysis for OPSCC due to the better soft tissue contrast of MRI in this anatomically challenging area. But in the end, the choice for CT or MRI will largely depend on the preference and experience of the radiologists within the center. The radiomic model presented in this article seems to have better predictive performance compared to fluorodeoxyglucose-positron emission tomography (FDG-PET) (AUC:0.64)[33]. This can be expected as FDG-PET images are less able to provide textural detail of tumor tissue.

The models in this article are less sensitive (88%) and specific (71%) compared to pathological methods (p16 immunohistochemistry: sensitivity 56-100%, specificity

79-93%; DNA PCR: sensitivity: 100% specificity 89% or the combination of latter techniques: Sensitivity and specificity 100%[34]) to determine HPV status of the tumor. However, these pathological methods are expensive and time consuming and are not always available (for instance, in retrospective studies when no biopsy is performed or biopsy/tissue samples are not available), making predictive models based on clinical and/or radiomic features a useful alternative.

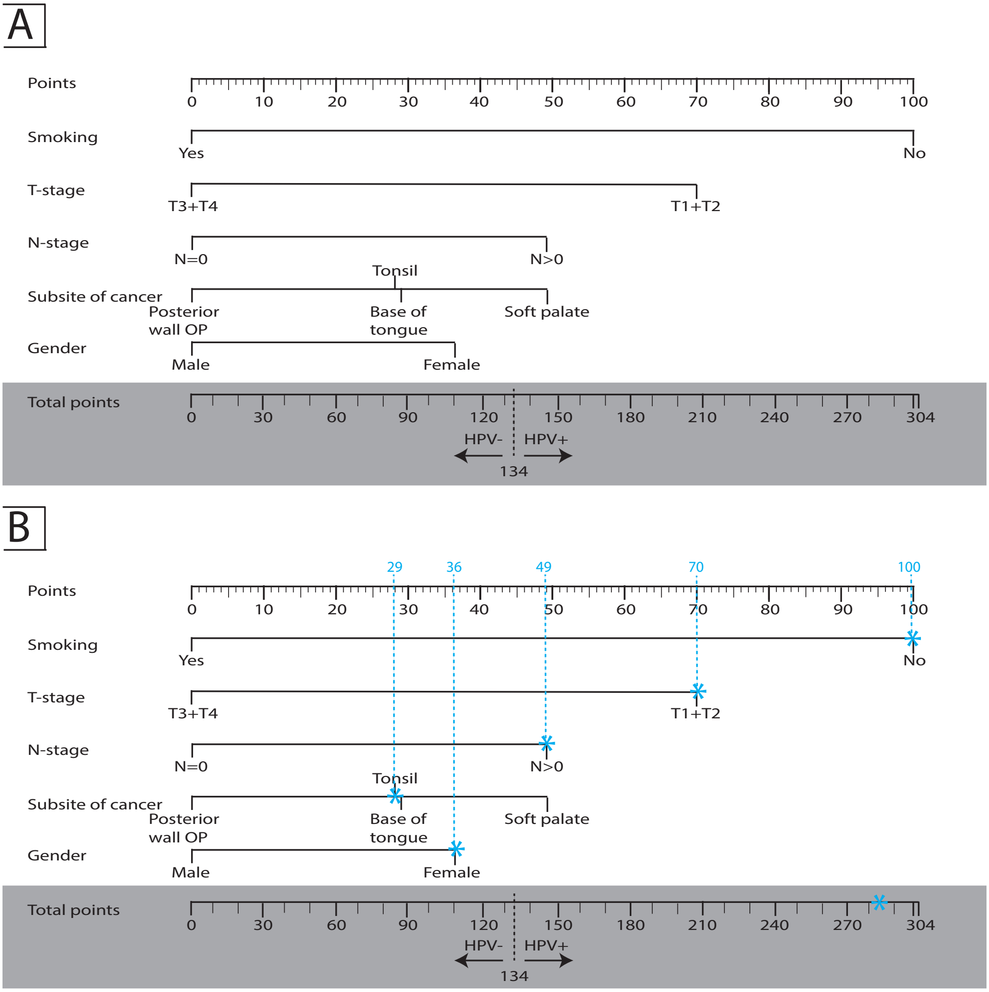


Fig. 3. Nomogram for the clinical model to predict human papillomavirus (HPV) positivity. A: Points are given to each clinical variable by drawing a line between the clinical variable with the “Points” line (top row) ranging from 0 to 100. The sum of all points for the individual clinical variables result in a total score (total points). A total score of ≥134 points is indicative of HPV positivity of the tumor. B: Worked example. A nonsmoking female with a T1 tumor of the tonsil region, including node-positive disease had a total score of 284 points, corresponding to HPV positivity of the tumor.

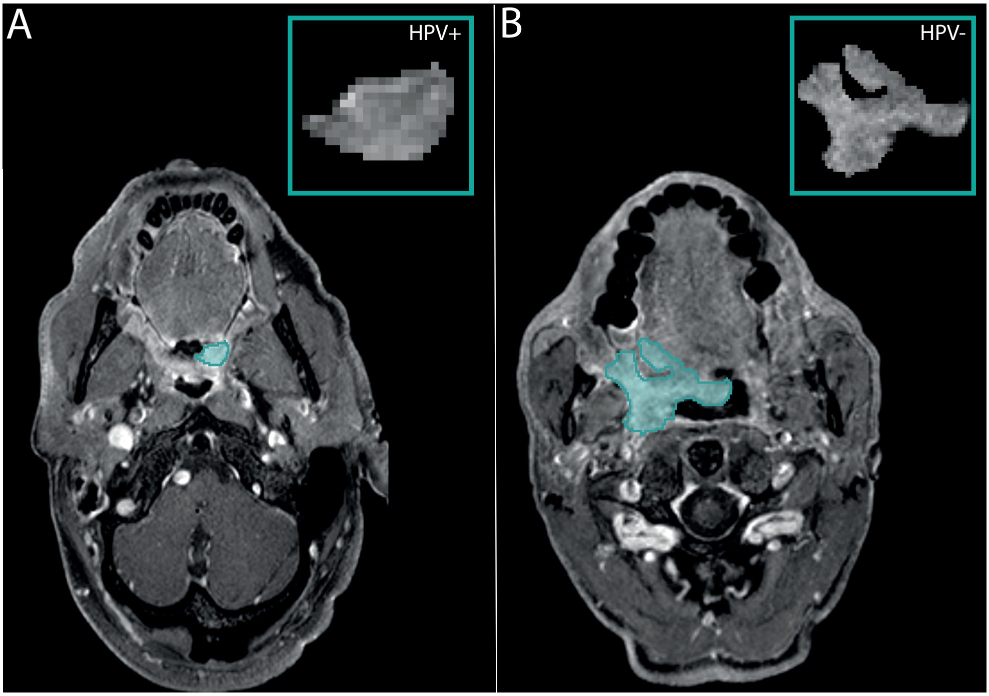


Fig. 4. Magnetic resonance image of a patient with human papillomavirus (HPV)-positive (A) and HPV-negative (B) tumor status (blue marked area) showing differences in textural appearances. The patient with a HPV-positive tumor status has a smaller and rounder tumor. Intensity values were less variated and less change of intensitites were visible.

This study is, to our knowledge, the largest radiomic study on MRI in head and neck squamous cell carcinomas[35]. However, our sample size is still quite limited compared to previous studies evaluating CT-based radiomics[15-17]. Clearly, larger populations, preferably in a multicenter setting, are needed to confirm our findings and create radiomic models that are more generalizable across scanners and populations.

The present study included patients from a single center, without an external cohort to validate our results, which is obviously a recommendation for further work. Another, minor, limitation might be the accuracy of the self-reporting variables, especially smoking status. This is partly overcome by categorizing smoking status into three robust categories (current-, former-, and nonsmoker), where former smokers stopped for at least 2 years prior to diagnosis. Only postcontrast 3DT1W MRIs were used in this study to limit the number of features with our available cases. Other MR sequences might give additional radiomic features for prediction of HPV status and is a topic for further study. In a preliminary study, we included all available MRI sequences, revealing mainly radiomic features from the postcontrast

3DT1W sequence, suggesting that other sequences would not contribute to the eventual predictive models. Finally, time-consuming manual tumor delineations were used for feature extraction, which introduces interobserver variability. Stable features with regard to delineations were selected to minimize the effect of interobserver variability in the eventual models. Ideally, this interobserver variability should be eliminated. Automated tumor delineation algorithms by, for instance, convolutional neural networks may overcome interobserver delineation variability[36]. In addition, automated tumor delineation would greatly reduce the workload of manual tumor delineation, making clinical implementation of radiomic analysis more feasible. Another approach would be to use deep-learning models or other unsupervised machine learning techniques to predict HPV status of head and neck tumors. However, adequate training of these models is challenging due to the relatively small tumors in a large and challenging anatomical area. Radiomic analysis therefore seems to be the most straight forward approach at this point in time.

CONCLUSION

This study shows that logistic regression models based on clinical variables, MR-based radiomic features, or a combination of clinical and radiomic features can accurately predict HPV status in OPSCC patients. Although a model based on clinical and radiomic features performs best, the clinical model would be the method of choice due to its ease of implementation. These models have a place in determination of HPV tumor status in settings where tumor biopsy material, tumor samples, immunohistochemistry, and/or DNA polymerase chain reaction techniques are not available. HPV testing is becoming more a routine in hospitals, but not everywhere, especially not in the past when the importance of HPV status of the tumor was not known. Medical images, on the other hand, are widely available due to the advantage of storage capability of medical images for a long time, making it a good alternative to assess HPV tumor status.

SUPPLEMENTARY INFORMATION

Password: PhD_PaulaBos

REFERENCES

1. Brakenhoff RH, Wagner S, Klussmann JP. Molecular patterns and biology of HPV-associated HNSCC. *Recent Results Cancer Res*. 2017;206:37-56. doi:10.1007/978-3-319-43580-0_3
2. Van Houten VMM, Snijders PJF, Van Den Brekel MWM, et al. Biological evidence that human papillomaviruses are etiologically involved in a subgroup of head and neck squamous cell carcinomas. *Int J Cancer*. 2001;93(2):232-235. doi:10.1002/ijc.1313
3. Ang KK, Harris J, Wheeler R, et al. Human Papillomavirus and Survival of Patients with Oropharyngeal Cancer. *N Engl J Med*. 2010;363(1):24-35. doi:10.1056/NEJMoa0912217
4. Fakhry C, Zhang Q, Nguyen-Tan PF, et al. Human papillomavirus and overall survival after progression of oropharyngeal squamous cell carcinoma. *J Clin Oncol*. 2014;32(30):3365–3373. doi:10.1200/JCO.2014.55.1937
5. Rietbergen MM, Brakenhoff RH, Bloemena E, et al. Human papillomavirus detection and comorbidity: critical issues in selection of patients with oropharyngeal cancer for treatment de-escalation trials. *Ann Oncol*. 2013;24(11):2740-2745. doi:10.1093/annonc/mdt319
6. Mehanna H, Robinson M, Hartley A, et al. Radiotherapy plus cisplatin or cetuximab in low-risk human papillomaviruspositive oropharyngeal cancer (de-escalate HPV): an open-label randomised controlled phase 3 trial. *Lancet*. 2019;393(10166):51-60. doi:10.1016/S0140-6736(18)32752-1
7. Elrefaey S, Massaro M, Chiocca S, Chiesa F, Ansarin M. HPV in oropharyngeal cancer: The basics to know in clinical practice. *Acta Otorhinolaryngol Ital*. 2014;34(5):299-309.
8. Mirghani H, Blanchard P. Treatment de-escalation for HPV-driven oropharyngeal cancer: where do we stand? *Clin Transl Radiat Oncol*. 2018;8:4-11. doi:10.1016/j.ctro.2017.10.005
9. Budu VA, Decuseara T, Balica NC, et al. The role of HPV infection in oropharyngeal cancer. *Rom J Morphol Embryol*. 2019;60(3):769-773.
10. Huang SH, O'Sullivan B. Overview of the 8th edition TNM classification for head and neck cancer. *Curr Treat Options Oncol*. 2017;18(7):40. doi:10.1007/s11864-017-0484-y
11. Smeets SJ, Hesselink AT, Speel EJM, et al. A novel algorithm for reliable detection of human papillomavirus in paraffin embedded head and neck cancer specimen. *Int J Cancer*. 2007;121(11):2465-2472. doi:10.1002/ijc.22980
12. Kim KY, Lewis JS, Chen Z. Current status of clinical testing for human papillomavirus in oropharyngeal squamous cell carcinoma. *J Pathol Clin Res*. 2018;4(4):213-226. doi:10.1002/cjp2.111
13. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by

noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006. doi:10.1038/ncomms5006

14. Fruehwald-Pallamar J, Hesselink JR, Mafee MF, Holzer-Fruehwald L, Czerny C, Mayerhoefer ME. Texture-Based analysis of 100 MR examinations of head and neck tumors - Is it possible to discriminate between benign and malignant masses in a multicenter trial? *Rofo*. 2016;188(2):195-202. doi:10.1055/s-0041-106066
15. Yu K, Zhang Y, Yu Y, et al. Radiomic analysis in prediction of Human Papilloma Virus status. *Clin Transl Radiat Oncol*. 2017;7:49-54. doi:10.1016/j.ctro.2017.10.001
16. Bogowicz M, Riesterer O, Ikenberg K, 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):921-928. doi:10.1016/j.ijrobp.2017.06.002
17. Leijenaar RTH, Bogowicz M, Jochems A, 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. doi:10.1259/bjr.20170498
18. Nauta IH, Rietbergen MM, van Bokhoven AAJD, et al. Evaluation of the eighth TNM classification on p16-positive oropharyngeal squamous cell carcinomas in The Netherlands and the importance of additional HPV DNA testing. *Ann Oncol*. 2018;29(5):1273-1279. doi:10.1093/annonc/mdy060
19. Beesley LJ, Bartlett JW, Wolf GT, Taylor JMG. Multiple imputation of missing covariates for the Cox proportional hazards cure model. *Stat Med*. 2016;35(26):4701-4717. doi:10.1002/sim.7048
20. Habbous S, Chu KP, Lau H, et al. Human papillomavirus in oropharyngeal cancer in Canada: analysis of 5 comprehensive cancer centres using multiple imputation. *CMAJ*. 2017;189(32):E1030-E1040. doi:10.1503/cmaj.161379
21. Ren J, Xu W, Su J, et al. Multiple imputation and clinicoserological models to predict human papillomavirus status in oropharyngeal carcinoma: an alternative when tissue is unavailable. *Int J Cancer*. 2020;146(8):2166-2174. doi:10.1002/ijc.32548
22. Henneman R, van Monsjou HS, Verhagen CVM, et al. Incidence changes of human papillomavirus in oropharyngeal squamous cell carcinoma and effects on survival in the Netherlands Cancer Institute, 1980-2009. *Anticancer Res.* 2015;35(7):4015-4022.
23. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104–e107. doi:10.1158/0008-5472.CAN-17-0339
24. Sperandei S. Understanding logistic regression analysis. *Biochem Med*. 2014;24(1):12-18. doi:10.11613/BM.2014.003
25. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46:389-422.
26. Kohavi R, John GH. Wrappers for feature subset selection. *Artif Intell*. 1997;97:273324.

27. Bergstra J, Yamins D, Cox D. Hyperopt: a python library for optimizing the hyperparameters of machine learning algorithms. *Proc SciPy.* 2013;13–20. doi:10.1088/1749-4699/8/1/014008

28. Ferri C, Hernandez-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. *Pattern Recognit Lett*. 2009;30:27-38. doi:10.1016/j.patrec.2008.08.010

29. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. *Data Min Knowl Disc*. 2019;9:e1301. doi:10.1002/widm.1301

30. Zhang Z, Kattan MW. Drawing nomograms with R: applications to categorical outcome and survival data. *Ann Transl Med*. 2017;5(10):211. doi:10.21037/atm.2017.04.01

31. Cantrell SC, Peck BW, Li G, Wei Q, Sturgis EM, Ginsberg LE. Differences in imaging characteristics of HPV-positive and HPV-negative oropharyngeal cancers: a blinded matched-pair analysis. *Am J Neuroradiol*. 2013;34(10):2005-2009. doi:10.3174/ajnr.A3524

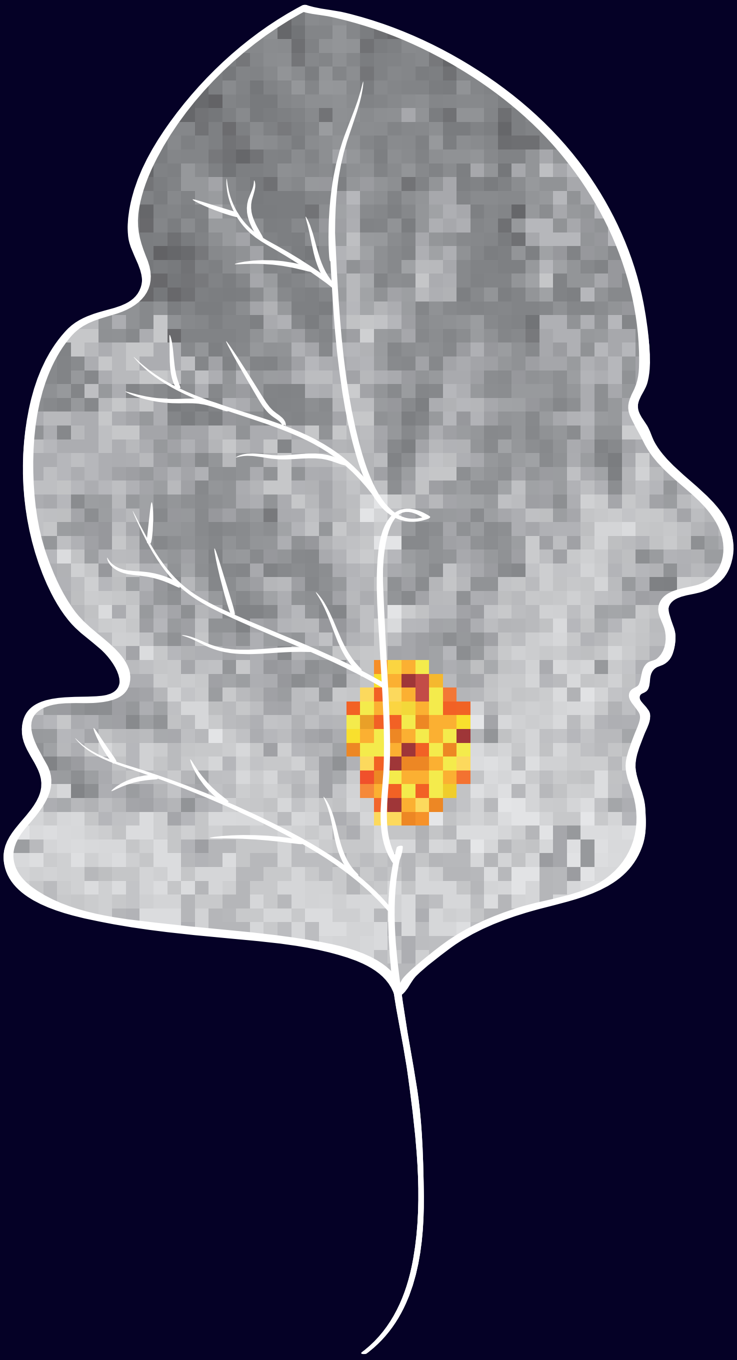
32. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*. 2015;517(7536):576-582. doi:10.1038/nature14129

33. Vallieres M, Kumar A, Sultanem K, El Naqa I. FDG-PET image-derived features can determine HPV status in head-and-neck cancer. *Radiat Oncol Biol*. 2013;87(2):S467.dDoi:10.1016/j.ijrobp.2013.06.1236

34. Robinson M, Schache A, Sloan P, Thavaraj S. HPV specific testing: a requirement for oropharyngeal squamous cell carcinoma patients. *Head Neck Pathol*. 2012;6(1):S83-S90. doi:10.1007/s12105-012-0370-7

35. Jethanandani A, Lin TA, Volpe S, et al. Exploring applications of radiomics in Magnetic Resonance Imaging of head and neck Cancer: A systematic review. *Front Oncol*. 2018;8:131. doi:10.3389/fonc.2018.00131

36. Pavic M, Bogowicz M, Wurms X. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol*. 2018;57(8):1070-1074. doi:10.1080/0284186X.2018.1445283



Part III

Simplification or automation of delineation techniques to improve clinical adoption of MR-based radiomics for OPSCC patients

Largest diameter delineations can substitute 3D tumor volume delineations for radiomics prediction of human papillomavirus status on MRIs of oropharyngeal cancer

Accepted for publication
Physica Medica, July 2022

Paula Bos
Michiel W.M. van den Brekel
Marjaneh Taghavi
Zeno A.R. Gouw
Abrahim Al-Mamgani
Selam Waktola
Hugo J.W.L. Aerts
Regina G.H. Beets-Tan
Jonas A. Castelijns
Bas Jasperse

6

ABSTRACT

Purpose: Laborious and time-consuming tumor segmentations are one of the factors that impede adoption of radiomics in the clinical routine. This study investigates model performance using alternative tumor delineation strategies in models predictive of human papillomavirus (HPV) in oropharyngeal squamous cell carcinoma (OPSCC).

Methods: Of 153 OPSCC patients, HPV status was determined using p16/p53 immunohistochemistry. MR-based radiomic features were extracted within 3D delineations by an inexperienced observer, experienced radiologist or radiation oncologist, and within a 2D delineation of the largest axial tumor diameter and 3D spheres within the tumor. First, logistic regression prediction models were constructed and tested separately for each of these six delineation strategies. Secondly, the model trained on experienced delineations was tested using these delineation strategies. The latter methodology was repeated with the omission of shape features. Model performance was evaluated using area under the curve (AUC), sensitivity and specificity.

Results: Models constructed and tested using single-slice delineations (AUC/Sensitivity/Specificity: 0.84/0.75/0.84) perform better compared to 3D experienced observer delineations (AUC/Sensitivity/Specificity: 0.76/0.76/0.71), where models based on 4mm sphere delineations (AUC/Sensitivity/Specificity: 0.77/0.59/0.71) show similar performance. Similar performance was found when experienced and largest diameter delineations (AUC/Sens/Spec: 0.76/0.75/0.65 vs 0.76/0.69/0.69) was used to test the model constructed using experienced delineations without shape features.

Conclusion: Alternative delineations can substitute labor and time intensive full tumor delineations in a model that predicts HPV status in OPSCC. These faster delineations may improve adoption of radiomics in the clinical setting. Future research should evaluate whether these alternative delineations are valid in other radiomics models.

INTRODUCTION

Radiomics is a promising tool for the non-invasive detection of clinically relevant tumor characteristics. These characteristics can be used to predict treatment response[1,2], classify tumor types[3,4] or discriminate tumor properties[5,6]. Radiomics analysis requires various steps that include image acquisition, image pre-processing, tumor delineation, feature extraction, feature selection and model construction. These steps can be controlled easily within research settings, but poses challenges with regard to reproducibility and repeatability in daily clinical practice[7-9]. Even if these challenges and other requirements for clinical implementation[10,11] are overcome, time consuming expert tumor delineations, taking valuable hours to complete, hampers further adoption of radiomics in daily clinical practice[12].

Time reduction with regard to tumor delineation can be achieved by either automated delineation strategies or manual delineation strategies which are easier to implement. Previous studies have shown that variability of tumor delineations can impact model performance. However, these studies[9,13] mainly focused on the consequences of (semi-)automatic alteration of available manual full tumor delineations on model performance. The methods used in these studies cannot be translated to adequate delineation strategies that would reduce time and labor consumption of manual tumor delineations needed for the implementation of radiomics in a clinical setting. A study comparing models based on rough and precise tumor delineations found that radiomic features extracted from precise delineations were more informative for prediction of overall survival in non-small cell lung cancer patients[14]. These interesting findings show that the choice of delineation strategy can lead to substantial variations in radiomic results[14]. Consensus of the most suitable delineation strategy is therefore highly recommended to standardize the radiomic workflow and increase clinical implementation.

In this study we investigate whether the performance of a previously published[5] radiomics model predictive of human papillomavirus (HPV) status of oropharyngeal squamous cell carcinomas (OPSCC) is similar when fast (“simple”, “rough”) or readily available tumor delineations are used compared to the time consuming standard expert tumor delineations. The following fast or readily available tumor delineations will be considered: tumor volumes delineated by a non-experienced observer, the readily available gross tumor volumes (GTV) delineated by radiation oncologists, tumor delineations extracted on the axial slice with the largest diameter and a simple strategy where a sphere was drawn within the tumor volume. Radiomic features (i.e. radiomics signature) are selected during model construction and may depend on the delineation strategy used. To ensure that the same radiomic features are detected when the model is applied to a new case, one can assume that the same delineation strategy should be used when implementing the model. Under this assumption, separate models will be constructed for each delineation strategy. On the other hand, alternative delineations may be able to adequately quantify relevant features that were selected in a model trained using the optimal expert 3D tumor delineations. Under this assumption, the performance of the model constructed using optimal delineations will be applied using the alternative delineations. The latter approach will be repeated while omitting shape and size features, as some of the alternative delineations are not able to quantify these features.

MATERIALS AND METHODS

The study was approved by the local institutional review board (IRBd18047). Due to the retrospective nature of the study, informed consent was waived.

Study population

A cohort of 240 patients with histologically proven primary OPSCC, treated with chemoradiation (CRT) between January 2010 and December 2015 at our Institute was considered. All patients had no history of previous head and neck malignancies. The main exclusion criteria were (a) no determined HPV status of the tumor, (b) no available pretreatment MRI, (c) poor image quality, (d) undetectable tumors, and, (e) a second head and neck primary tumor. In total, 153 patients were eligible for this study. HPV status of the tumor was determined on biopsy material using p16 and p53 immunohistochemistry using the methodology described in Henneman et al.[15].

Image acquisition

Pretreatment MR and CT images were acquired as part of the clinical routine. T1-weighted postcontrast (postcontrast T1W) MRI was used for analysis, with a slice thickness ranging between 0.8 and 1.0 mm, TR/TE: 4300-10000/1.7-4.6 ms, echo train length of 60-90 and 10° flip angle.

CT images for GTV delineation were acquired during treatment planning from two scanners. All CT images had a slice thickness of 3mm, a tube current of 120 kV, and an exposure ranging from 19 to 509 mAs.

Tumor delineations

Primary tumors were delineated using six delineation strategies (see below and Figure 1), including three delineations covering the whole tumor volume and three delineations including only a part of the tumor ("simple delineations", e.g. spherical volumes). Whole tumor volumes represent the full 3D tumor volume, where "simple delineations" evaluates tumor delineation strategies which might easily implementable in the clinic. Tumors were delineated on postcontrast T1W MRI, except for the GTV delineation. Observers were allowed to review other available imaging modalities to improve tumor delineation and were blinded to HPV status. Delineations were performed using the 3D slicer software (version 4.8.0, www.slicer.org). The annotation time for each delineation time was recorded.

1. *3D Non-experienced observer*: One observer in training (PB, 1 year of experience in head and neck diagnosis) delineated the 3D tumor volume.
2. *3D Experienced observer*: An experienced radiologist (BJ, >7 years of expertise in head and neck diagnosis) reviewed and corrected the *Non-experienced* tumor delineation.
3. *3D GTV*: GTV was delineated on contrast-enhanced planning CT-scan for radiotherapy treatment purposes by a radiotherapist, with the allowance to review planning MRI when available. Planning CT and its GTV contouring were registered to post contrast T1W using B-spline registration (SimpleElastix[16], see Appendix A).
4. *2D Largest Diameter*: The slice with the largest axial tumor diameter was automatically selected from the 3D *Experienced* manual tumor delineation using Python scripting (version 3.4, www.python.org).
5. *3D Spherical ⌀4mm*: A sphere of 4mm was placed in the most solid part of the tumor by the *Non-experienced* observer. A size of 4 mm was selected since this was the minimum maximal tumor diameter included in the cohort.
6. *3D Spherical ⌀BestFit*: A sphere with the largest possible diameter (best fit) was placed in the most solid tumor area by the *Non-experienced* observer.

The spherical tumor delineations were delineated one year after initial delineation of the *Non-experienced* observer, blinded to the initial delineation to prevent memory bias.

Image pre-processing

Prior to analysis, MR images were normalized, resampled and discretized. Image normalization was applied with zero mean and unit standard deviation to avoid inhomogeneity between MRI scans. Comparable quantification of radiomic features in all directions was obtained by resampling MR images to isotropic voxels of 1.0 mm using B-spline interpolation. Finally, MR intensity values were discretized into

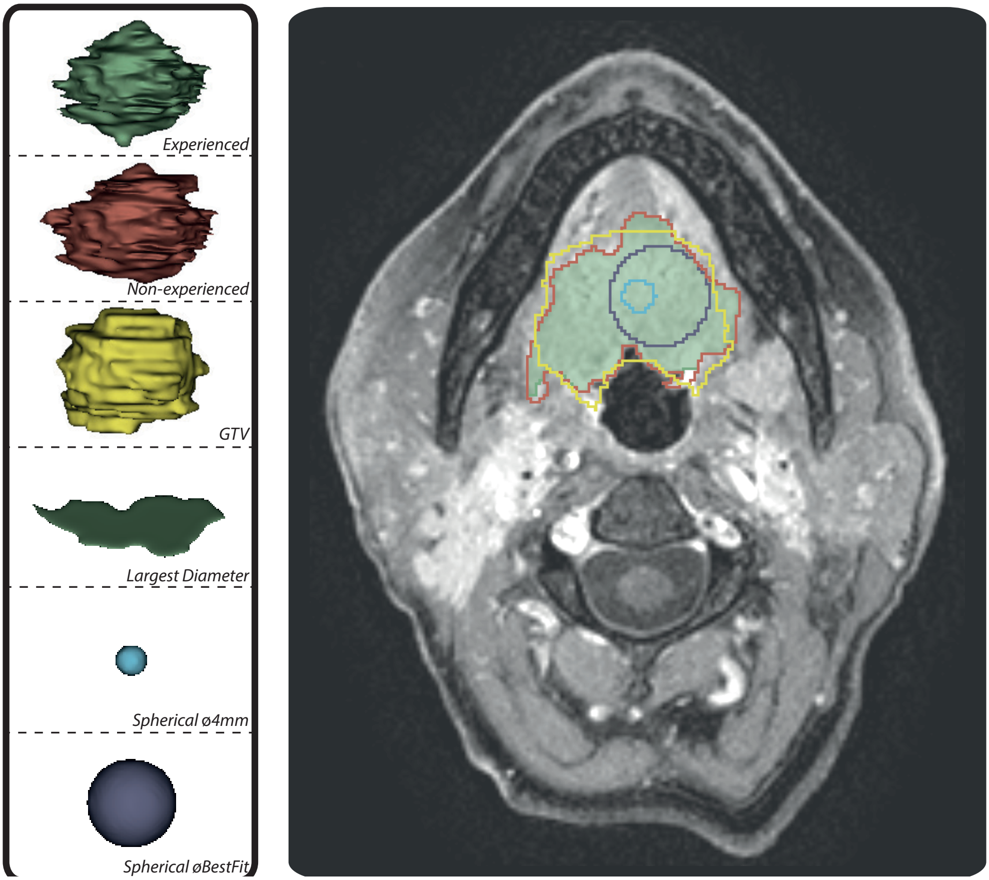


Fig. 1. An illustration of the six manual delineations. The six individual delineations are visualized in the left box. The right box illustrates these delineations on postcontrast T1w MRI on the slide with the largest axial diameter.

a fixed bin width of five intensity values to allow quantification of texture. All image pre-processing steps were performed using the open-source package PyRadiomics[17].

Radiomic features

Radiomic features were extracted using PyRadiomics[17] for each separate delineation strategy. Features were divided into the categories shape, intensity and texture. These features were extracted from the original image, the image with a wavelet filter and the image with a Laplacian of Gaussian (LoG) filter. A wavelet filter was used to examine different spatial frequencies of the image in 8 decompositions, where a LoG filter determines different texture coarseness (4 levels, sigma of 0.5, 1.0, 1.5 and 2.0mm). A total of 1184 radiomic features were extracted for each delineation.

Stable features were assessed by intraclass correlation coefficient (ICC) and Mann-Whitney U test for each separate delineation strategy used for model construction. First, radiomic features were considered to be stable when ICC between the radiomic features extracted from the experienced radiologist and the appropriate tumor delineation (*Non-experienced*, *GTV*, *Largest Diameter*, *Spherical ⌀4mm* and *Spherical ⌀BestFit*) was higher than 0.75. For the *Experienced* model, ICC was calculated between features extracted from the *Experienced* reader and *Non-experienced* reader. ICC calculated stable features were assessed by Mann-Whitney U test to exclude differences of magnetic field strength. Features without significant differences (p-value ≥ 0.05) were considered stable. Finally, collinearity between the remaining stable features was assessed by Pearson correlation (>0.9), removing the features with the largest collinearity. The stable features for each separate delineation strategy were used as input for the prediction model.

Features were standardized per delineation strategy, using zero mean and unit variance, to obtain scalar homogeneity in each approach. Then, recursive feature elimination[18] was used to select a feature subset by iteratively removing the feature with the weakest importance score. The remaining feature subset was used for analysis by the logistic regression classifier to predict HPV tumor status and subsequent model testing.

For the prediction model, the cohort was divided into a training (60%, n = 91) and test (40%, n=62) subset, stratified by magnetic field strength and HPV status of the tumor. Hyperparameters for classification were optimized using 1000 iterations of Bayesian hyperparameter optimization on the training subset. During this step, four-fold cross-validation was applied to calculate the minimal loss function. Then, the optimal hyperparameters were applied on the unseen test set to evaluate prediction performance. A detailed description of the workflow can be found in our previous publication[5]. The radiomic pipeline is summarized in Figure 2.

The impact of tumor delineation variability on the prediction performance of HPV was investigated using three methods.

Method 1: Separate model construction and testing for each delineation strategy
Prediction models were built (trained and validated) and tested on each tumor delineation separately (*Experienced, Non-experienced, GTV, Largest diameter, Spherical ⌀4mm and Spherical ⌀BestFit)*, resulting in six separate models. To prevent artificial inflation of model performance, all models were forced to select the same number of features as selected in the experienced model.

Method 2: Testing the Experienced model using the alternative delineations
Performance of the prediction model that was trained and validated using *Experienced* delineations was tested on the test subset using each of the six tumor delineation strategies.

Method 3: Testing the Experienced model without shape and size features using the alternative delineations
As spherical or 2D delineations do not reliably represent shape and size features, the *Experienced* model was trained and validated without shape and size features and tested using the six alternative delineations.

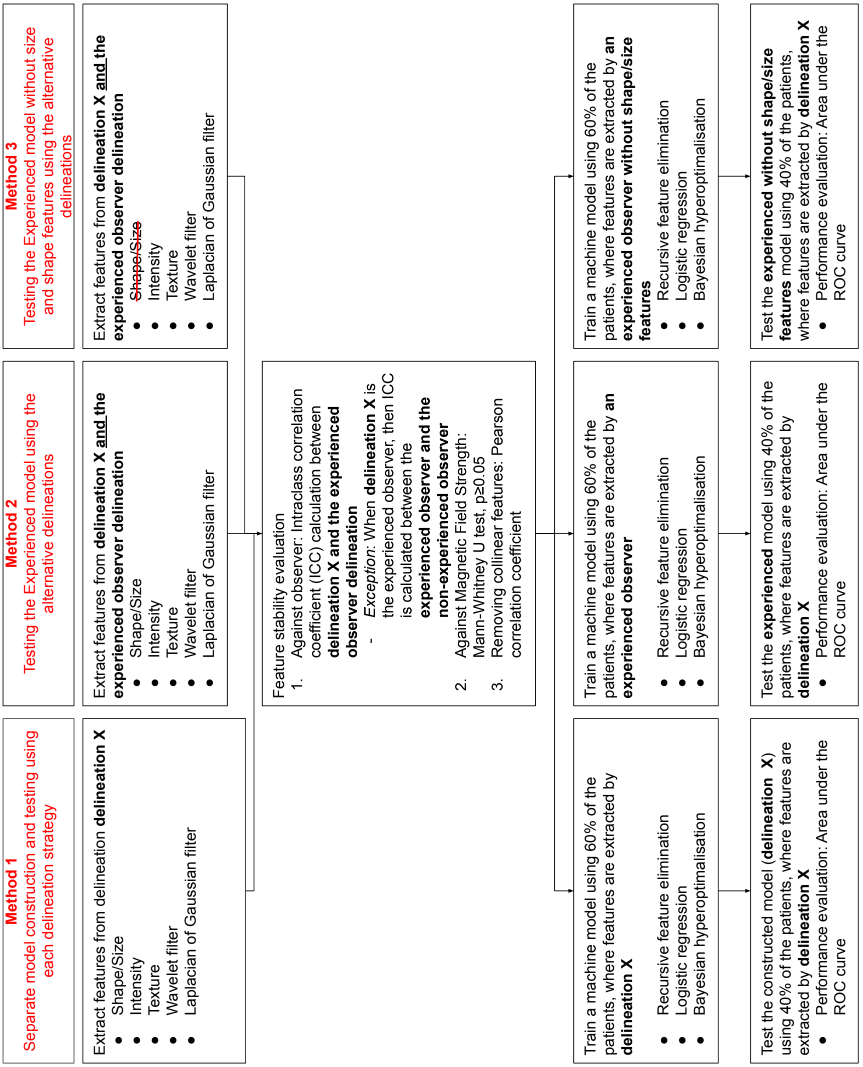

Method 1
Separate model construction and testing using each delineation strategy
Extract features from delineation **delineation X**
• Shape/Size
• Intensity
• Texture
• Wavelet filter
• Laplacian of Gaussian filter
Method 2
Testing the Experienced model using the alternative delineations
Extract features from **delineation X and the experienced observer delineation**
• Shape/Size
• Intensity
• Texture
• Wavelet filter
• Laplacian of Gaussian filter
Method 3
Testing the Experienced model without size and shape features using the alternative delineations
Extract features from **delineation X and the experienced observer delineation**
• ~~Shape/Size~~
• Intensity
• Texture
• Wavelet filter
• Laplacian of Gaussian filter
Feature stability evaluation
1. Against observer: Intraclass correlation coefficient (ICC) calculation between **delineation X and the experienced observer delineation**
- *Exception*: When **delineation X** is the experienced observer, then ICC is calculated between the **experienced observer and the non-experienced observer**
2. Against Magnetic Field Strength: Mann-Whitney U test, p≥0.05
3. Removing collinear features: Pearson correlation coefficient
Train a machine model using 60% of the patients, where features are extracted by **delineation X**
• Recursive feature elimination
• Logistic regression
• Bayesian hyperoptimalisation
Train a machine model using 60% of the patients, where features are extracted by **an experienced observer**
• Recursive feature elimination
• Logistic regression
• Bayesian hyperoptimalisation
Train a machine model using 60% of the patients, where features are extracted by **an experienced observer without shape/size features**
• Recursive feature elimination
• Logistic regression
• Bayesian hyperoptimalisation
Test the constructed model (**delineation X**) using 40% of the patients, where features are extracted by **delineation X**
• Performance evaluation: Area under the ROC curve
Test the **experienced** model using 40% of the patients, where features are extracted by **delineation X**
• Performance evaluation: Area under the ROC curve
Test the **experienced without shape/size features** model using 40% of the patients, where features are extracted by **delineation X**
• Performance evaluation: Area under the ROC curve

Fig. 2. A flowchart describing the radiomic workflow of the three methods. Delineation X can be one of the six delineation strategies, including the experienced observer, non-experienced observer, gross tumor volume (GTV), largest diameter on the single slice, a sphere with a diameter of 4mm or a sphere with a diameter best fitted in the tumor volume.

Statistical analysis

An independent t-test was applied to calculate differences in age for both HPV status groups. Fisher´s exact test was applied to the other clinical variables. A p-value below 0.05 was considered as significant. Spatial agreement between the six delineation strategies was calculated by using the Dice Similarity Coefficient (DSC)[19] and Hausdorff Distance (HD)[20].

Performance of the prediction models was evaluated by area under the curve (AUC), sensitivity and specificity. Median values, with its 95% confidence interval (95% CI) were calculated using 500 iterations of bootstrap (with replacement) using the test set.

RESULTS

Patient demographics

Patient demographics are summarized in Table 1. The patients show an equal distribution for HPV tumor classification (n=77 HPV negative tumors, n=76 HPV positive tumors). Younger (p=0.007), non-smoking patients (p<0.001) with a high T-classification (p=<0.0001) or tumor not located in the soft palate (p=0.017) were more likely to have HPV positive tumors. Other cancer subsites and gender were not significantly different between HPV negative and positive tumors. N-classification was slightly higher in HPV positive compared to HPV negative tumors with near significance (p=0.051).

Time recordings

The *Non-experienced* observer delineated a tumor with a median of 34 minutes [range: 25-65], and was checked and corrected in a median of 9 minutes [range: 6-14] by the *Experienced* observer. The time required to place a ROI with a diameter of 4 mm or user-determined diameter was 1.5 and 3 minutes, respectively. *Largest Diameter* delineations were automatically extracted, and therefore, obtained within seconds. Time recordings of *GTV* delineations were not available, since those were previously delineated for radiotherapy purposes.

Tumor delineation agreement

Agreement between tumor volumes was calculated with DSC and HD, see Table A.1 and Table A.2. The *Experienced* and *Non-experienced* observer show reasonable similarity with a mean DSC of 0.84 and mean HD of 18.7mm. *GTV* tumor delineation shows a lower similarity with *Experienced* observer (DSC: 0.43, HD: 183.3mm).

Table 1. Patient characteristics for the total cohort and subgroups stratified by HPV status. Summaries are given as number of patients and % of the total group between parentheses. Median and interquartile range (IQR) are used to summarize continuous variables. [a]Independent t-test, [b]Fisher's exact test and [c]Chi-square test. Values were statistic significant (marked with an asterisk) if p-value was below 0.05 (p<0.007 after Bonferroni correction).

	Total cohort	**HPV negative**	**HPV positive**	**P-value**
Patients, *n*	153	77	76	-
Age, *median y* (IQR)	61 (56-66)	63 [57-67]	59 [55-65]	0.007[a]*
Sex, *n male* (%)	96 (63)	54 (70)	42 (55)	0.067[b]
Smoking, *n* (%)	114 (75)	72 (94)	42 (55)	<0.001[b]*
T-stage, *n* (%)				<0.001[b]*
T1 + T2	78 (51)	25 (32)	53 (70)	-
T3 + T4	75 (49)	52 (68)	23 (30)	-
N-stage (N>0), *n* (%)	127 (83)	59 (77)	68 (89)	0.051[b]
Subsite of cancer				0.406[c]
Tonsillar tissue	88 (58)	42 (55)	46 (60)	0.514[b]
Soft palate	13 (8)	11 (14)	2 (3)	0.017[b]*
Base of tongue	48 (31)	20 (26)	28 (37)	0.166[b]
Posterior wall	4 (3)	4 (5)	0 (0)	0.120[b]

*Note: HPV indicates Human Papillomavirus

Logistic regression prediction model

Prediction performances of all models for the three methods are summarized in Table 2, ROC curves are visualized in Figure 3.

Method 1: Separate model construction and testing for each delineation strategy

0.3 to 6.5% of the total features were defined as stable (see Table 3), resulting in 77, 10, 20, 4 and 13 radiomic features as input for the *Experienced/Non-experienced*, *GTV*, *Largest Diameter*, *Spherical ⌀4mm* and *Spherical ⌀BestFit* model, respectively.

The model built and tested based on *Largest Diameter* delineation shows higher performance, higher specificity and similar sensitivity (AUC/Sens/Spec: 0.84/0.75/0.84) compared to the standard *Experienced* model (AUC/Sens/Spec: 0.76/0.76/0.71). Prediction performance of the *Spherical ⌀4mm* delineations model was comparable to standard *Experienced* delineation model with slightly lower sensitivity and similar specificity (AUC/Sens/Spec: 0.77/0.59/0.71). Performance of models based on *Non-experienced* (AUC/Sens/Spec: 0.68/0.69/0.55), *GTV* (AUC/

Sens/Spec: 0.71/0.69/0.58) and *Spherical⌀BestFit* (AUC/Sens/Spec: 0.64/0.59/0.62) delineations were considerably lower than the standard *Experienced* model.

Table A.3 summarizes the selected features for each model. The models based on *Experienced* and *Largest Diameter* delineation include shape/size features (sphericity and maximum 2D diameter respectively), as well as textural features. Models based on the other delineations included only textural features.

Method 2: Testing the Experienced model using alternative delineations

The standard *Experienced* model shows the highest performance when tested on expert radiologist tumor delineations (AUC/Sens/Spec: 0.76/0.76/0.71). Overall performance and specificity were considerably lower when the *Experienced* model was tested using the *Non-experienced* (AUC/Sens/Spec: 0.63/0.76/0.50) delineations. Test performance approached randomness when tested with the remaining delineations. Sensitivity and specificity for testing with the 2D or spherical delineations were 0 and 1 or vice versa.

Method 3: Testing the Experienced model without shape and size features using the alternative delineations

Of the extracted 1184 radiomic features, 14 features belong to the shape and size group. Those 14 features were excluded when shape and size features were omitted. Of the remaining 1170 features, 71 (6.1%) features were considered as stable (see Table 3).

Performance of the *Experienced* model without shape and size features was comparable to the standard Experienced model *with* shape and size features (AUC/Sens/Spec: 0.76/0.75/0.65 vs 0.76/0.76/0.71). This performance is similar to the *Largest Diameter* model (AUC/Sens/Spec: 0.76/0.69/0.69). Prediction performances increased when the *Experienced model without shape and size features* was tested using *Non-experienced* delineations (AUC/Sens/Spec: 0.82/0.76/0.80). Performance of this model using G*TV*, *Spherical ⌀BestFit* or *Spherical ⌀4mm* delineations was considerably lower, as summarized in Table 2.

DISCUSSION

This study shows that less labor-intensive, easily applicable, delineations might substitute labor-intensive experienced delineations in the application of radiomics models to predict HPV status. Moreover, some of these alternative delineation strategies seem to increase model performance compared to standard expert delineations.

Table 2. Performances (expressed in AUC, sensitivity and specificity) of the models of all three methods in predicting human papillomavirus (HPV) status of the tumor. Confidence intervals were calculated from 500 times bootstrapping. Stable features were calculated between features extracted from the experienced delineation ([a]) and the listed delineations ([b]).

Method		**1**			**2**			**3**		
Model construction specifics	Stable feature selection based on	Experienced and listed delineations [a,b]			Experienced and Non-experienced delineations			Experienced and Non-experienced delineations		
	Features removed	None			None			Shape and size features		
	Model construction based on	Listed delineations*			Experienced			Experienced		
Model testing results	**Delineation**	**Test AUC [CI]**	**Sensitivity [CI]**	**Specificity [CI]**	**Test AUC [CI]**	**Sensitivity [CI]**	**Specificity [CI]**	**Test AUC [CI]**	**Sensitivity [CI]**	**Specificity [CI]**
	Experienced[a]	0.76 [0.76-0.77]	0.76 [0.75-0.77]	0.71 [0.70-0.72]	0.76 [0.76-0.77]	0.76 [0.75-0.77]	0.71 [0.70-0.72]	0.76 [0.76-0.77]	0.75 [0.74-0.76]	0.65 [0.64-0.66]
	Non-experienced[b]	0.68 [0.68-0.69]	0.69 [0.68-0.70]	0.55 [0.54-0.56]	0.63 [0.63-0.64]	0.76 [0.76-0.77]	0.50 [0.49-0.51]	0.82 [0.81-0.82]	0.76 [0.75-0.77]	0.80 [0.79-0.81]
	GTV[b]	0.71 [0.70-0.72]	0.69 [0.68-0.70]	0.58 [0.56-0.59]	0.53 [0.52-0.54]	0.33 [0.32-0.34]	0.61 [0.60-0.62]	0.70 [0.69-0.70]	0.75 [0.74-0.76]	0.65 [0.64-0.66]
	Largest Diameter[b]	0.84 [0.83-0.85]	0.75 [0.74-0.76]	0.84 [0.83-0.85]	0.53 [0.52-0.54]	0	1	0.76 [0.75-0.76]	0.69 [0.68-0.70]	0.69 [0.68-0.70]
	Spherical ⌀4mm[b]	0.77 [0.76-0.77]	0.59 [0.58-0.60]	0.71 [0.70-0.72]	0.39 [0.38-0.40]	1	0	0.58 [0.57-0.58]	0.56 [0.55-0.57]	0.50 [0.49-0.51]
	Spherical ⌀BestFit[b]	0.64 [0.64-0.65]	0.59 [0.58-0.60]	0.62 [0.60-0.63]	0.52 [0.51-0.53]	1	0	0.67 [0.66-0.68]	0.59 [0.58-0.60]	0.68 [0.67-0.69]

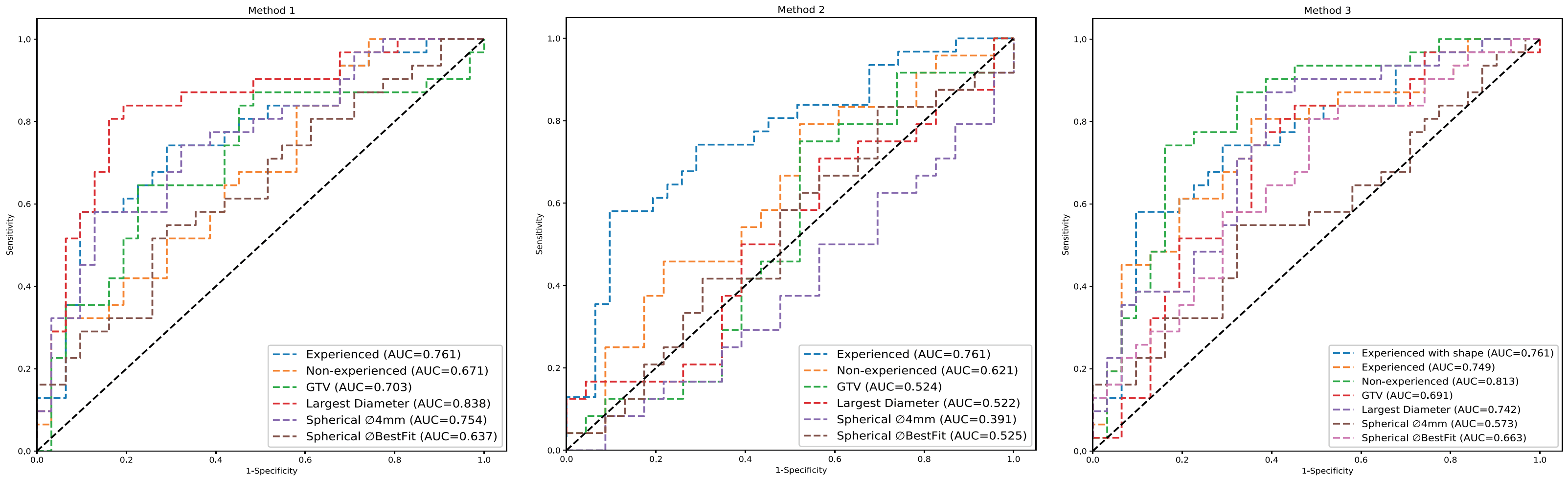


Fig. 3. Receiver operating characteristic (ROC) curves of the three methods. Performances of the test set for each individual delineation are assessed by the area under the curve (AUC).

Table 3. The number of features after each stability check for each observer versus the experienced delineation model. The number of stable features are given, with the percentage of the total number of features between parenthes

Experienced vs observer delineation \ Stability check	**Non-experienced[a] (%)**	**GTV (%)**	**Largest Diameter (%)**	**Spherical ∅4mm (%)**	**Spherical ∅BestFit (%)**	**Experienced without shape and size features(%)**
None	1184 (100)	1184 (100)	1184 (100)	1184 (100)	1184 (100)	1170 (100)
for delineation (ICC > 0.75)	926 (78.2)	241 (20.4)	483 (40.8)	90 (7.6)	310 (26.2)	913 (78.0)
for magnetic field strength (mwu ≥0.05)	240 (20.3)	34 (2.9)	68 (5.7)	11 (0.9)	64 (5.4)	231 (19.7)
for collinear features (Pearson > 0.9)	77 (6.5)	10 (0.8)	20 (1.7)	4 (0.3)	13 (1.1)	71 (6.1)

Note: [a]Stable features for the *Experienced* model *with* shape and size features were also selected by this comparison. ICC represents Interclass correlation Coefficient; Mwu, Mann-Whitney U Test

In contrast to our expectations, all delineations (except *Spherical ⌀BestFit)* show good prediction performance, regardless of delineation precision. This suggest that each separate delineation capture information with regard to tumor biology in a different matter.

The model based and tested on largest tumor diameter delineations appeared to outperform the standard experienced delineation based model. This may be explained by the effect of interpolation on the radiomic features. Interpolation is recommended as necessary preprocessing step to correct for pixel size and slice thickness variance for 3D volumes. This interpolation to isotropic voxels induces smoothing effects that might remove relevant feature information from 3D delineations that will be present in (unsmoothed) 2D tumor delineations[21]. Additional experiments (see appendix B) supports this hypothesis, as performance of a model based on 3D tumor volumes delineated by an experienced observer (AUC: 0.74) increases when interpolation was omitted (AUC: 0.81).

Poor model performance was observed when the standard experienced model was applied to the test subset using the alternative delineations. This poor test performance might be explained by the reduced ability of the “faster” delineations to adequately quantify the sphericity feature (see appendix Table A.3) that is part of the experienced model. This does not rule out that applying the experienced model using alternative delineations may be useful in other predictive models that only rely on textural features.

Removal of shape and size features (method 3) did not change the performance when the model was constructed and tested using the expert radiologist delineations. As expected, prediction performances were considerably better when this experienced model (constructed *without* shape and size features) was tested with the alternative delineations compared to the standard experienced model (constructed *with* shape and size features (method 2)). Taken together, this implies that the loss of shape and size features might be adequately compensated with textural features without losing predictive properties.

To make radiomics clinically applicable, substitution of the labor-intensive time-consuming delineations is desirable. This study shows that easy delineation strategies needed a shorter time to perform the delineation (*Non-experienced delineation* vs *Spherical* delineation: 34 min vs 3 min). While no direct comparison can be made for the 2D delineation, it can be safely assumed that delineating only a single slice requires less time compared to the full 3D tumor delineation. Taking prediction performance and ease of implementation into account, the largest diameter seems to be the most preferable alternative delineation strategy.

Evidently, the findings of this study are only applicable to models predicting HPV in OPSCC. Other delineation strategies may be more applicable for radiomics models trained to predict other outcome variables or applied to other tumor types. Besides tumor delineation and the studied outcome parameter, each step of the radiomic pipeline shows large variations, limiting reproducible and repeatable results[7-9,22]. Preselected choices in image acquisition, tumor delineation, feature selection and/or machine learning model construction parameters directly affect the radiomic pipeline and therefore the set of predictive features. Though all these variances, direct and reliable comparison between studies is limited.

A good example of this are the contrary results between findings of this study and Lang *et al*.[23] regarding the superiority of 2D delineations over 3D tumor volumes in the prediction of HPV status. Significant differences within the methodology (e.g. MR images vs CT images, machine learning model vs deep learning model, feeding one vs multiple 2D slices in the model) impede critical evaluation.

As our study aimed to find suitable delineation alternatives to full tumor delineations by an experienced observer, observer variability of model performance was not assessed. Observer variability of delineations should be addressed in future studies, or studies aiming to adopt this alternative delineation approach. It is obvious that observer variability is less of an issue in the proposed faster delineations compared to full tumor volume delineations as tumor margins are not delineated. Another important limitation of this study is the bias introduced by interdependency of delineations. The single slice delineations are calculated from the expert 3D delineations, which may inflate the performance of single slice delineations compared to the 3D delineations. Furthermore, the results presented for the single slice delineations do not represent the real-world scenario of an observer manually selecting and delineating the largest tumor diameter from the image. Additionally, expert and non-expert delineations are not totally independent, as the expert delineations are basically the corrected non-expert delineations. Future research should take these limitations into account by evaluating independently acquired manual delineations.

Besides the easy implementation of radiomics in the clinical workflow, the alternative delineations would also benefit standardization of radiomics analysis. Reliable automatic segmentation of tumors would be the best solution to time and labor-intensive delineations while eliminating interobserver bias[10,11]. Multiple studies investigated the potential of deep learning in auto segmentation in head and neck cancer patients, where substantial overlap (DSC>0.74) between the manual and automatic delineations was shown[24,25]. Other studies proposed multi-task deep learning to combine automatic segmentations with models predictive of treatment outcome[26] or HPV status[23]. However, to our knowledge, no reliable automatic tools for the delineations of complex oropharyngeal tumors based on MR images are available at this point in time, and therefore automatic delineations are not included in this study.

As mentioned earlier, various factors can influence robustness and stability of individual features and should be used to select the most suitable feature for every radiomics model. Feature stability across delineations was used as a selection criterion in this study, where features were defined as stable when agreement between the experienced radiologist and the appropriate delineation was high. By selecting features with only high agreement, features prognostic for HPV status might be eliminated since they were different across full tumor and single slice delineation. Additionally, feature robustness can be influenced by the MRI scanner used and circumstances under which the MRI scan was performed[22]. Evidently, this could not be addressed in this single center study, and should be addressed in future projects.

Recently, advances have been made to increase performance of radiomics models by improving image quality using AI techniques. For instance, Chen *et al*. have improved the predictive performance of a radiomics model by denoising CT images using Generative Adversarial Networks[27]. These techniques could also be employed to improve the quality of MRI images and/or the similarity of MRI image acquired from different scanners. By improving predictive performance of radiomics models, these technique might also increase performance of the alternative delineation strategies proposed in this study.

CONCLUSIONS

In conclusion, this study shows that alternative delineations with low labor/time consumption can substitute labor and time intensive full tumor delineations in the application of a model that predicts HPV status in OPSCC. These faster delineations may improve adoption of radiomics in the clinical setting. Evidently, the findings in this paper are only relevant to the radiomics model predicting HPV status used in this paper, future research should evaluate whether these alternative delineations are valid in other radiomics models.

SUPPLEMENTARY INFORMATION

Password: PhD_PaulaBos

REFERENCES

1. Chu CS, Lee NP, Adeoye J, Thomson P, Choi SW. Machine learning and treatment outcome prediction for oral cancer. *J Oral Pathol Med*. 2020;49(10):977-985. doi:10.1111/jop.13089
2. Yuan Y, Ren J, Shi Y, Tao X. MRI-based radiomic signature as predictive marker for patients with head and neck squamous cell carcinoma. *Eur J Radiol*. 2019;117:193-198. doi:10.1016/j.ejrad.2019.06.019
3. Fruehwald-Pallamar J, Hesselink JR, Mafee MF, Holzer-Fruehwald L, Czerny C, Mayerhoefer ME. Texture-Based analysis of 100 MR examinations of head and neck tumors - Is it possible to discriminate between benign and malignant masses in a multicenter trial? *Rofo*. 2016;188(2):195-202. doi:10.1055/s-0041-106066
4. Zheng YM, Xu WJ, Hao DP, et al. A CT-based radiomics nomogram for differentiation of lympho-associated benign and malignant lesions of the parotid gland. *Eur Radiol*. 2021;31(5):2886-2895. doi:10.1007/s00330-020-07421-4
5. Bos P, van den Brekel MWM, Gouw ZAR, et al. Clinical variables and magnetic resonance imaging-based radiomics predict human papillomavirus status of oropharyngeal cancer. *Head Neck*. 2021;43(2):485–495. doi:10.1002/hed.26505
6. Romeo V, Cuocolo R, Ricciardi C, et al. Prediction of tumor grade and nodal status in oropharyngeal and oral cavity squamous-cell carcinoma using a radiomic approach. *Anticancer Res*. 2020;40(1):271-280. doi:10.21873/anticanres.13949
7. Berenguer R, Del Rosario Pastor-Juan M, Canales-Vázquez J, et al. Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters. *Radiology*. 2018;288(2):407-415. doi:10.1148/radiol.2018172361
8. Pfaehler E, Zhovannik I, Wei L, et al. A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features. *Phys Imaging Radiat Oncol*. 2021;20:69-75. doi:10.1016/j.phro.2021.10.007
9. Liu R, Elhalawani H, Radwan Mohamed AS, et al. Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer. *Clin Transl Radiat Oncol*. 2020;21:11-18. doi:10.1016/j.ctro.2019.11.005
10. Castiglioni I, Rundo L, Codari M, et al. AI applications to medical images: From machine learning to deep learning. *Physica Medica.* 2021;83:9-24. doi:10.1016/j.ejmp.2021.02.006
11. Balagurunathan Y, Mitchell R, El Naga I. Requirements and reliability of AI in the medical context. *Physica Medica.* 2021;83:72-78. doi:10.1016/j.ejmp.2021.02.024
12. Harari PM, Song S, Tomé WA. Emphasizing Conformal Avoidance Versus Target Definition for IMRT Planning in Head-and-Neck Cancer. *Int J Radiat Oncol Biol Phys.* 2010;77(3):950–958. doi:10.1016/j.ijrobp.2009.09.062
13. Zhang X, Zhong L, Zhang B, et al. The effects of volume of interest delineation on MRI-based radiomics analysis: Evaluation with two disease groups. *Cancer*

Imaging. 2019;19(1):89. doi:10.1186/s40644-019-0276-7

14. Sepehri S, Tankyevych O, Iantsen A, Visvikis D, Hatt M, Cheze Le Rest C. Accurate tumor delineation vs rough volume of interest analysis for 18F-FDG PET/CT Radiomics-based prognostic modeling in Non-Small Cell Lung Cancer. *Front Oncol.* 2021;11:726865. doi:10.3389/fonc.2021.726865
15. Henneman R, van Monsjou HS, Verhagen CVM, et al. Incidence changes of human papillomavirus in oropharyngeal squamous cell carcinoma and effects on survival in the Netherlands Cancer Institute, 1980-2009. *Anticancer Res.* 2015;35(7):4015-4022.
16. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. Elastix: a toolbox for intensity based medical image registration. *IEEE Trans Med Imaging.* 2010;29(1):196-205. doi:10.1109/TMI.2009.2035616
17. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104–e107. doi:10.1158/0008-5472.CAN-17-0339
18. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46:389-422.
19. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945;26(3):297–302. doi:10.2307/1932409
20. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. *IEEE Trans Pattern analysis and machine intelligence*. 1993;15(9):850-863. doi:10.1109/CVPR.1992.223209
21. Park SH, Lim H, Bae BK, 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. doi:10.1186/s40644-021-00388-5
22. Sun M, Baiyasi A, Liu X, et al. Robustness and reproducibility of radiomics in T2 weighted images from magnetic resonance image guided linear accelerator in a phantom study. *Physica Medica.* 2022;96:130-139. doi:10.1016/j.ejmp.2022.03.002
23. Lang DM, Peeken JC, Combs SE, Wilkens JJ, Bartzsch S. Deep learning based HPV status prediction for oropharyngeal cancer patients. *Cancers*. 2021;13(4):786. doi:10.3390/cancers13040786
24. Fontaine P, Andrearczyk V, Oreiller V, et al. Fully automatic head and neck cancer prognosis prediction in PET/CT. M*ultimodal learning for clinical decision Support.* Springer. 2021;59-68. doi:10.1007/978-3-030-89847-2_6
25. Andrearczyk V, Oreiller V, Jreige M, et al. Overview of the HECKTOR challenge at MICCAI 2020: Automatic head and neck tumor segmentation in PET/CT. *3D Head and Neck Tumor segmentation in PET/CT Challenge.* Springer, Cham, 2020.
26. Andrearczyk V, Fontaine P, Oreiller V, et al. Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer. *Predictive Intelligence In Medicine.* Springer, Cham, 2021. doi:10.1007/978-3-030-87602-9_14

27. Chen J, Bermejo I, Dekker A, et al. Generative model simprove radiomics performance in different tasks and different datasets: An experimental study. *Physica Medica.* 2022;98:11-17. doi:10.1016/j.ejmp.2022.04.008

Simple delineations cannot substitute full 3D tumor delineations for MR-based radiomics prediction of locoregional control in oropharyngeal cancer

European Journal of Radiology; Mar 2022;148:110167
DOI: 10.1016/j.ejrad.2022.110167

Paula Bos
Michiel W.M. van den Brekel
Marjaneh Taghavi
Zeno A.R. Gouw
Abrahim Al-Mamgani
Selam Waktola
Hugo J.W.L. Aerts
Regina G.H. Beets-Tan
Jonas A. Castelijns
Bas Jasperse

7

ABSTRACT

Background: Manual delineation of head and neck tumor contours for radiomics analyses is tedious and time consuming. This study investigates if fast or readily available tumor contours can substitute full tumor contours by an experienced observer for an MR-based radiomics model to predict locoregional control (LRC) in oropharyngeal squamous cell carcinoma (OPSCC) tumors.

Materials and methods: Radiomic features were extracted from postcontrast T1-weighted MRIs of 177 OPSCC primary tumors using six different manual delineation strategies. LRC prediction models based on recursive feature elimination combined with logistic regression were built. Models were trained and tested on data from each separate delineation. Additionally, the model derived from segmentations from the experienced reader was tested by each of the alternative delineations. Complementary, this was repeated with removal of size and shape features. Model performance was evaluated using area under the curve (AUC).

Results: Prediction performance of the experienced radiologist tumor delineation (AUC: 0.74) was superior compared to all other delineations when trained and tested (AUCs: 0.41–0.56) or trained on experienced delineations and tested (AUCs: 0.56–0.67) on alternative segmentations. Removal of size and shape features considerably decreases prediction performance (AUC: 0.54). Applying the model based on expert delineations to spherical or single slice delineations makes prediction worthless since these models predict one class.

Conclusion: Fast or readily available contours cannot substitute full expert tumor delineations in radiomics models predictive of LRC in OPSCC.

INTRODUCTION

In the last decade radiomics has been showing promising value to characterize biological tumor properties[1-3] or predict treatment response[4-6]. Due to its complex methodology, reproducibility and repeatability of radiomics are a major concern. Several factors might play a role, among which is tumor delineation. Studies show that inter- observer variability[7] or alterations of delineations[8] can impact model performance. However, these studies only consider (semi)automatic alterations of manual tumor delineations.

An previous study[9] has shown that delineations of the largest tumor diameter on a single slice can substitute the standard time consuming manual delineations of the full 3D tumor volume for an MR-based radiomics model predicting human papillomavirus (HPV) in oropharyngeal squamous cell carcinoma (OPSCC). This finding would greatly reduce the time needed to create tumor delineations, thereby facilitating the adoption of radiomics in clinical practice.

However, alternative delineations that are able to substitute full tumor delineations in a model predictive of HPV, might not be able to substitute full tumor delineations for a model predictive of another variable. In the construction of radiomics models, tumor features are selected based on their relationship with the variable that needs to be predicted. Thereby, the number and types of features in the eventual model will vary based on the variable to be predicted and the imaging characteristics of the tumor of interest. The extent to which the alternative simple delineations can substitute full tumor delineations depends on their ability to adequately quantify the features considered for construction or testing of the radiomics model.

The aim of this study was to investigate whether the substitution of full tumor delineations by fast or readily available tumor delineations used for the radiomics model to predict HPV is feasible in a model predictive of locoregional tumor control (LRC) for OPSCC using the same methodology.

More specifically, simple spherical tumor volumes, tumor delineations on the slice with the largest diameter, delineations by a non-experienced observer, and the already available gross tumor volume (GTV) delineations for radiation therapy were considered as alternative delineation strategies. The performance of models constructed and/or tested using these alternative delineations are compared to the standard model constructed and tested using tumor delineations from an experienced radiologist.

MATERIALS AND METHODS

This retrospective study was approved by the local institutional review board (IRBd18047). Informed consent was waived due to the retrospective design of the study.

Patient population

Patients treated with chemoradiation (CRT) for histologically proven primary OPSCC between January 2010 and December 2015 were retrospectively collected resulting in 240 consecutive patients. Exclusion criteria were (1) no available pretreatment MRI examination, (2) poor image quality, (3) small undetectable tumors, (4) synchronous tumors and (5) history of previous head and neck cancer. Clinical variables (in particular age, gender, smoking status, tumor subsite, HPV status, TNM-classification (7th edition), follow-up data on date of tumor recurrence, site of recurrence and lymph node metastasis were collected for all patients. LRC was defined as the absence of local recurrence and/or lymph node metastases within 2 years after treatment initiation, determined by clinical, radiological, and, if needed, histological assessment.

Image acquisition

MR images were acquired as part of standard staging of OPSCC in our institute at 1.5T (n=82 patients) or 3T (n=95 patients) (Achieva, Philips Medical System, Best, The Netherlands)[10]. The imaging protocol included T1-weighted (T1W), T2-weighted (T2W), T1-weighted postcontrast (postcontrast T1W), and dynamic MRI scans (diffusion and perfusion). 3D isotropic postcontrast T1W was used for analysis (postcontrast 3D-T1W), acquired with a slice thickness ranging from 0.8 to 1.0 mm (TR/TE: 4300–10000/1.7–4.6 ms, echo train length: 60–90, flip angle: 10°).

Computer Tomography (CT) scans for radiotherapy planning were made on two CT scanners (Siemens Sensation Open, Siemens Healthcare, Erlangen, Germany; Philips Gemini TF Big Bore, Philips, Eindhoven, The Netherlands). All CT scans were reconstructed with a slice thickness of 3 mm and acquired with a tube current of 120 kV and an exposure ranging from 19 to 509 mAs.

Radiomics methods

Tumor delineations, image pre-processing, radiomic feature extraction, construction of radiomics models and statistical analysis was essentially the same in the previous study[10] on alternative delineations for the prediction of LRC in OPSCC.

In summary: Six different primary tumor contours were manually delineated on postcontrast 3D T1W MRI using 3D slicer software (version 4.8.0, www.slicer.org) (except *GTV*). *GTV* was delineated on the pretreatment radiotherapy planning CT, using the corresponding MRI as reference. Figure A.1 demonstrates the six delineation strategies:

1. *3D Non-experienced observer*: The 3D volume of the tumor was contoured by an observer in training (PB) with 1 year of experience in head and neck diagnosis.
2. *3D Experienced observer*: The tumor volume of the *3D Non-experienced observer* was controlled and corrected by an experienced observer (BJ), with >7 years of experience in head and neck diagnosis.
3. *3D GTV delineation*: The already available tumor contouring used for radiotherapy treatment was collected. The planning CT-based GTV delineation, interpreted by a radiotherapist, was registered to postcontrast 3D T1W MRI using B-spline registration by the open-source software SimpleElastix[11] (See Appendix B).
4. 2D largest diameter delineation: The slice with the largest axial tumor diameter was automatically extracted from the 3D Experienced tumor delineation using Python (version 3.4, www.python.org).
5. 3D Spherical ⌀4mm delineation: A predefined sphere with a diameter of 4 mm was placed in the most solid part of the tumor by the non-experienced observer. The size of 4 mm was chosen, as result of the maximum fitting diameter in the smallest tumor of the patient cohort.
6. 3D Spherical ⌀BestFit delineation: A sphere with an adjustable diameter was placed in the most solid part of the tumor by the non-experienced observer. The selected diameter was the largest possible diameter (best fit) fitting in each tumor volume.

MRI images were normalized using zero mean, resampled to 1.0 mm isotropic voxels and discretized with a fixed bin width of five. 1184 radiomics features, including shape, intensity, texture, wavelet transform (8 decompositions) and Laplacian of Gaussian (LoG) filter (sigma 0.5, 1.0, 1.5 and 2.0 mm), were extracted using PyRadiomics (version 2.2.0)[12] for all six tumor delineations separately.

Stable features were selected as input for the machine learning pipeline. Features were considered stable when the intraclass correlation coefficient (ICC) was above 0.75, calculated between features extracted from the listed delineations and the experienced delineation. Stable features for the experienced model were assessed by evaluating the agreement between features extracted from the experienced and non-experienced delineation. Additionally, the remaining features were examined against magnetic field strength (calculated using the Mann- Whitney *U* test) and collinearity (calculated using Pearson correlation).

The cohort was divided into a training (70%, n=124) and test (30%, n=53) subset, stratified by magnetic field strength, HPV status and LRC. For model creation, four-fold cross validation was used to determine optimal model hyperparameters[13] using recursive feature elimination[14] with logistic regression on the training subset. The resulting model was applied to the test subset. Area under the curve (AUC), sensitivity and specificity were used as evaluation parameters, with a 95% confidence interval (95% CI), calculated using 500 iterations of bootstrap (with replacement).

Model construction and testing was performed using three different methods: Separate model construction and testing for each delineation method (method 1), testing the model constructed on experienced delineations (*Experienced* model) using the alternative delineations (method 2) and testing the *Experienced* model without shape and size features using the alternative delineations (method 3). The radiomic workflow is visualized in Appendix Figure A.2.

Statistical analysis

Fishers' exact test, independent *t*-test, and, Chi-square test were applied to calculate differences between clinical variables and LRC. P-values below 0.05 were considered statistically significant (p<0.007 after Bonferroni correction). Agreement between the respective six tumor volumes was calculated using Dice Similarity Coefficient (DSC)[15] and Hausdorff distance (HD)[16]. DSC values ranged between 0 (no overlap) and 1 (complete overlap). A higher value represents more spatial overlap between the two volumes. A DSC above 0.6 is considered to be appropriate. A smaller HD indicates that the surfaces of both volumes are closer to each other, with thereby a better agreement between both tumor volumes.

RESULTS

Patient demographics

In total, 177 patients were included in this study, for which patient demographics are summarized in Table 1. In total 145 (82%) patients had LRC after 2 years. Patients with LRC were more likely to have HPV positive tumor status (47% vs 25%, p=0.012) and low T-stage (58% vs 32%, p=0.013). Age, gender, smoking status, N-stage and tumor subsite were comparable for both groups. No significant differences were seen in patient characteristics with regard to LRC.

Table 1. Patient demographics. Baseline characteristics and outcome after CRT for all patients and subsets stratified by LRC. Summaries are given as number of patients and % of the total group between parentheses. Median and interquartile range (IQR) are used to summarize continuous variables. ^{a}Independent *t*-test, ^{b}Fisher exact test and ^{c}Chi-square test. Values were statistic significant (marked with an asterisk) if p-value was below 0.05 (p<0.007 after Bonferroni correction).

	Total cohort	**Patients with LRC**	**Patients with LRF**	**P-value**
Patients, *n*	177	145	32	–
Age, *median y* (IQR)	61 (56–66)	62 (56–66)	60 (57–66)	0.548^{a}
Sex, *n male* (%)	111 (63)	89 (61)	22 (69)	0.427^{b}
Smoking, *n* (%)	134 (76)	108 (74)	26 (81)	0.500^{b}
HPV				0.012^{c}
Negative, *n* (%)	77 (44)	56 (39)	21 (66)	–
Positive, *n* (%)	76 (43)	68 (47)	8 (25)	–
Unknown, *n* (%)	24 (13)	21 (14)	3 (9)	–
T-stage, *n* (%)				0.013^{b}
T1 + T2	94 (53)	84 (58)	10 (32)	–
T3 + T4	83 (47)	61 (42)	22 (68)	–
N-stage (N > 0), *n* (%)	141 (80)	115 (79)	26 (81)	0.815^{b}
Subsite of cancer				0.406^{c}
Tonsillar tissue	99 (56)	83 (57)	16 (50)	–
Soft palate	18 (10)	14 (10)	4 (13)	–
Base of tongue	56 (32)	46 (32)	10 (31)	–
Posterior wall	4 (2)	2 (1)	2 (6)	–
Clinical endpoint				–
LRC < 2 year, *n* (%)	145 (82)	0 (0)	32 (100)	–
Time to LRF in months, *median* (IQR)	6 (4–13)	–	6 (4–13)	–

*Note: HPV indicates Human Papillomavirus; LRC Locoregional control; LRF Locoregional failure

Agreement between tumor delineations

The interobserver agreement of delineations, calculated with DSC and HD, between radiomic features of each individual tumor delineation method is summarized in Appendix Table A.1, Figure A.3 and Figure A.4. *Experienced* tumor delineation shows reasonable overlap with *Non-experienced* tumor delineation (DSC: 0.83, HD: 18.2 mm), and decreasing overlap with *GTV* and *Spherical ⌀BestFit* (DSC: 0.44/0.39, HD:

184.9/30.2 mm respectively). Overlap with the other delineations was extremely low.

Model performance for each method

In total, 1184 radiomic features were extracted from each delineation method, including 14 shape features. As expected, the number of stable features decreased to a varying degree with the alternative delineation methods as compared to stable feature determination using the delineations of the entire tumor (4.0% to 0.3%) This was the case for both the experienced and non-experienced observer (see Table 2). All prediction performances are summarized in Table 3. Selected features of the individual models are summarized in Table A.2.

Method 1: Separate model construction and testing using each delineation method

AUC of models trained on each separate tumor segmentation ranged 0.40 to 0.74 (See Figure 1). The model based on and tested using the *Experienced* tumor delineations outperformed all other models (AUC/Sensitivity/Specificity: 0.74/0.75/0.60). Performance of the other models was near random: *GTV* delineation (AUC/Sensitivity/Specificity: 0.56/0.67/0.50), *Spherical ⌀4mm(*AUC/Sensitivity/Specificity:0.54/0.43/0.71), *Non-experienced* delineations (AUC/Sensitivity/Specificity: 0.52/0.66/0.50), *Largest Diameter* (AUC/Sensitivity/Specificity: 0.46/0.67/0.50), and *Spherical ⌀BestFit* (AUC/Sensitivity/Specificity: 0.40/0.62/0.29)).

Method 2: Testing the Experienced model using the alternative delineations

As illustrated in Figure 2, predictive performance was highest for the *Experienced* model (AUC/Sensitivity/Specificity: 0.74/0.75/0.60). Prediction performance decreased when applying the model using *Spherical ⌀BestFit* (AUC/Sensitivity/Specificity: 0.67/1.00/0.00), *Non-experienced* (AUC/Sensitivity/Specificity: 0.66/0.69/0.50) and *Spherical ⌀4mm* (AUC/Sensitivity/Specificity: 0.65/1.00/0.00) delineations. The performance was considerably lower for *Largest diameter* (AUC/Sensitivity/Specificity: 0.61/0.00/1.00) and *GTV* (AUC/Sensitivity/Specificity: 0.56/0.34/0.71) delineations. Models based on spherical or single slice delineations had a sensitivity or specificity of 0.00 or 1.00.

Method 3: Testing the Experienced model without size and shape features using the alternative delineations

Removal of size and shape features dramatically decreased performance of the prediction model using *Experienced* tumor delineations (AUC/Sensitivity/

Table 2. The number of features after each stability checks for each observer versus the ground truth (*Experienced* delineation). The number of stable features are given, with the percentage of the total number of features between parentheses.

Experienced vs observer delineation \ Stability check	**Non-experienced* (%)**	**GTV (%)**	**Largest diameter (%)**	**Spherical Ø4mm (%)**	**Spherical ØBestFit (%)**	**Experienced without shape and size features (%)**
None	1184 (100)	1184 (100)	1184 (100)	1184 (100)	1184 (100)	1170 (100)
for delineation (ICC>0.75)	902 (76.2)	263 (22.2)	493 (41.6)	102 *8.6)	326 (27.5)	913 (78)
for magnetic field strength (mwu ≥0.05)	113 (9.5)	49 (4.1)	65 (5.5)	14 (1.2)	72 (6.1)	210 (17.9)
for collinear features (Pearson > 0.9)	47 (4.0)	10 (0.8)	24 (2.0)	4 (0.3)	12 (1.0)	75 (6.4)

Note: *Stable features for the *Experienced* model were also selected by this comparison. ICC represents Intraclass correlation Coefficient; Mwu, Mann-Whitney U Test

Table 3. Performances (expressed in AUC, sensitivity and specificity) of the models predicting locoregional control (LRC). Confidence intervals were calculated from 500 times bootstrapping. Stable features were calculated between features extracted from the listed delineations (marked with an asterisk(*)), and the experienced delineation (marked with two asterisks(**)).

Method		**1**			**2**			**3**		
Model construction specifics	Stable feature selection based on	Experienced** and listed delineations*			Experienced and Non-experienced delineations			Experienced and Non-experienced delineations		
	Features removed	None			one			Size and shape features		
	Model construction based on	Listed delineations			Experienced			Experienced		
Model testing results	**Delineation**	**Test AUC [CI]**	**Sensitivity [CI]**	**Specificity [CI]**	**Test AUC [CI]**	**Sensitivity [CI]**	**Specificity [CI]**	**Test AUC [CI]**	**Sensitivity [CI]**	**Specificity [CI]**
	Experienced**	0.74 [0.73-0.75]	0.75 [0.74-0.76]	0.60 [0.58-0.62]	0.74 [0.73-0.75]	0.75 [0.74-0.76]	0.60 [0.58-0.62]	0.54 [0.53-0.55]	0.82 [0.82-0.83]	0.30 [0.28-0.32]
	Non-experienced*	0.52 [0.51-0.53]	0.66 [0.65-0.66]	0.50 [0.48-0.52]	0.66 [0.65-0.67]	0.69 [0.68-0.70]	0.50 [0.48-0.52]	0.46 [0.45-0.47]	0.77 [0.77-0.78]	0.20 [0.19-0.21]
	GTV*	0.56 [0.55-0.57]	0.67 [0.66-0.67]	0.50 [0.48-0.52]	0.56 [0.55-0.57]	0.34 [0.34-0.35]	0.71 [0.70-0.73]	0.58 [0.57-0.59]	0.55 [0.54-0.56]	0.50 [0.48-0.52]
	Largest Diameter*	0.46 [0.45-0.47]	0.67 [0.66-0.67]	0.50 [0.48-0.52]	0.61 [0.60-0.63]	0	1	0.63 [0.62-0.64]	1	0
	Spherical ø4mm*	0.54 [0.53-0.55]	0.43 [0.42-0.44]	0.71 [0.70-0.73]	0.65 [0.65-0.66]	1	0	0.72 [0.71-0.74]	1	0
	Spherical øBestFit*	0.40 [0.38-0.41]	0.62 [0.61-0.62]	0.29 [0.27-0.30]	0.67 [0.66-0.68]	1	0	0.61 [0.60-0.62]	0.83 [0.82-0.83]	0.20 [0.19-0.21]

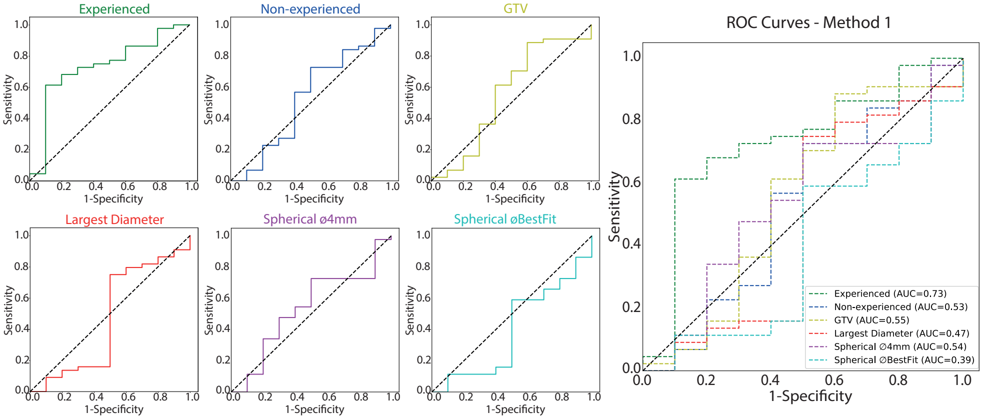


Fig. 1. ROC curves of performances of the test set, assessed by AUC, for method 1.

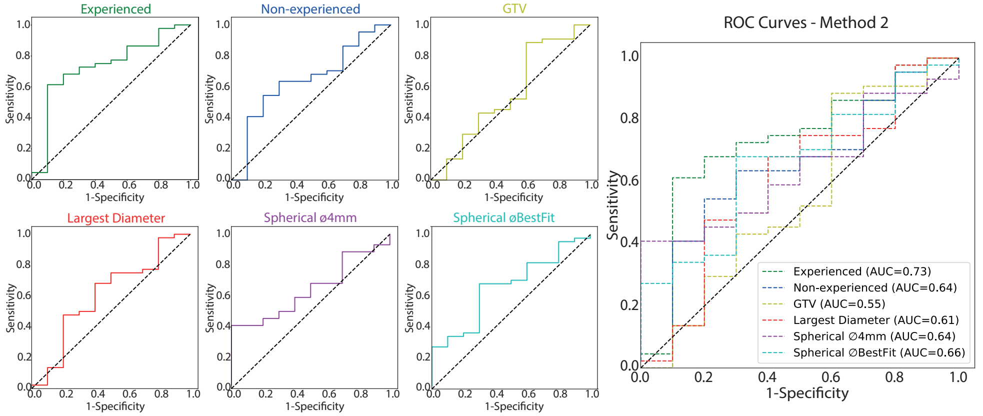


Fig. 2. ROC curves of performances of the test set, assessed by AUC, for method 2.

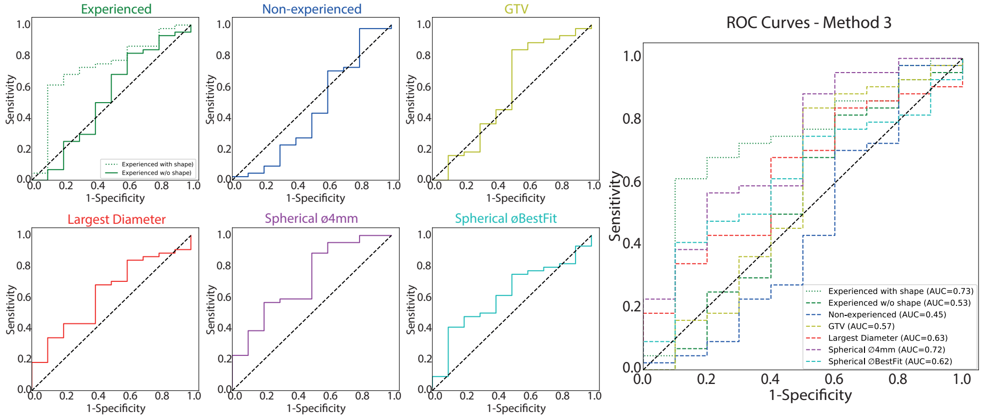


Fig. 3. ROC curves of performances of the test set, assessed by AUC, for method 3.

Specificity: 0.54/0.82/0.30), with a slightly better performance when GTV delineations (AUC/Sensitivity/Specificity: 0.58/0.55/0.50) were applied. Interestingly, *Spherical ⌀BestFit* (AUC/Sensitivity/Specificity: 0.60/0.83/0.20) tumor delineations outperformed *Experienced* delineation when used to apply the model to the test set. *Spherical ⌀4mm* (AUC/Sensitivity/Specificity: 0.72/1.00/0.00) and *Largest Diameter* (AUC/Sensitivity/Specificity: 0.63/1.00/0.00) tumor delineations outperformed the *Experienced* model, but showed a sensitivity of 1.00 and specificity of 0.00. Prediction performance was near random when *Non-experienced* delineations (AUC/Sensitivity/ Specificity: 0.46/0.77/0.20) were used. The ROC curves are visualized in Figure 3.

DISCUSSION

The main finding of this study is that faster or readily available tumor delineations do not provide a reliable alternative to tumor delineations from an experienced radiologist for the creation or application of an MR-based radiomics model predictive of LRC in patients with OPSCC.

This finding is contrary to the findings of Bos *et al.*[9], where performance using the largest tumor diameter on a single slice was higher compared to the time consuming manual delineations of the full 3D tumor volume in models predictive of HPV. The features that are included in the final constructed models are for the greater part different for the prediction of LRC and HPV[9] (see Table A.3). Apparently, at least part of the features included in the HPV model can be reasonably quantified by faster delineations, allowing substitution of expert full tumor delineations by some of these alternative delineations. This was not the case for features included in the models constructed to predict LRC. Therefore, faster alternative delineations may be used in some but not all radiomics models. Whether alternative delineations can be used, and which delineations and model construction approach is appropriate needs to be determined for each radiomics model separately, as confirmed by other studies[7,17].

Another explanation for the contrary findings between the two studies, might be the effect of interpolation of voxel values on the values and dispersion of features. Changes in voxel values due to interpolation can be expected to affect features that are directly derived from these voxel values, like histogram-based features such as mean, kurtosis, etc. In contrast, shape based features, like sphericity or maximum diameter, are not directly based on voxel values and should therefore not be greatly affected by interpolation effects. This might explain the observed differences between the two studies, as relatively more shape features were selected in the HPV model and relatively more histogram-based features in the LRC model. A sub-analysis seems to confirm this hypothesis, (see Appendix Table A.4), showing that the difference in model AUC between the largest diameter and full 3D tumor volume delineations decreased for the LRC, and to a lesser extent, the HPV models. Still, the LRC model based 3D full tumor delineations outperformed the model based on largest diameter delineations when correcting for interpolation.

Besides the shape and histogram-based features, peripheral surface information can also be the consequence of contrary performances using single slice or whole tumor delineation in the prediction of HPV status[9] and LRC. Invasive tumors generally have worse treatment outcomes, where HPV status does not depend on surface characteristics. This implies that surface information is of more relevance in the distinguishing between patients with a good or poor treatment response when compared to HPV status determination.

To our knowledge, only two studies compared the performance of 2D and 3D delineations. Shen et al.[18] developed prediction models for survival in non-small cell lung cancer (NSCLC) patients based on a single slice (2D) and whole tumor (3D) delineations. Findings showed that prediction performance was slightly better using 2D compared to 3D features (C-index: 0.68 vs 0.63). This is in contrast to the findings of Yang et al.[19] who reported that 3D features were favourable in a nomogram for predicting survival (C-index: 0.62 vs 0.70). Additionally, a nomogram combining 2D and 3D features was superior to models based on only a single slice or whole tumor delineations. This assumes that features extracted from a single slice and whole tumor delineations are complementary to each other and may both have particular predictive power. Therefore different types of cancer may require different approaches to delineation[18].

Another interesting finding concerns the performance of the models based on expert tumor delineation with (method 1) and without (method 3) shape/size features. Removal of shape and size features did not change performance in a model predictive of HPV (AUC: 0.76 vs 0.76)[9], when performance decreased considerably for the prediction of LRC (AUC: 0.74 vs 0.54). This implies that HPV prediction is mainly driven by texture features, where LRC prediction is more associated with tumor contour characteristics.

It is important to note that a reasonable AUC was found for testing the experienced model using some fast alternative delineations in this study (i.e. *Largest diameter*, *Spherical ⌀4mm*, and *Spherical ⌀BestFit model*). However, the results found for these delineations had a sensitivity of 1 and specificity of 0 or vice versa. Evidently, the combination of the experienced model with these fast delineations classified either all cases as positive or negative for LRC, rendering them useless for clinical application.

The need to explore the impact of every single factor of the radiomic pipeline on feature variability is already described in previous literature[20,21]. Of these factors several are already explored[22,23], such as acquisition and reconstruction parameters. The evaluation of the influence of tumor delineation variability is limited[7,17], especially in MR-based images, and, lacking in the use of manual delineation approaches. The evaluation of six different manual delineation approaches, representative for clinical purposes, using MR images, on radiomic feature variability makes this study unique.

A limitation of our study is the selection of stable features. Stable features are selected by calculating ICC of features extracted by tumor delineations from the expert radiologist and extracted from the appropriate delineation. This methodology requires time-consuming expert contours and might eliminate features yielding a better prognostic value than the remaining features, resulting in better prediction performance. Moreover, the automatic selection of the single slice with the largest axial diameter requires also the expert tumor delineations. Theoretically, this methodology selects the slice reflecting the broad tumor heterogeneity, while a slice representative for the predictive outcome variable (HPV status[9] and LRC) is desirable. Therefore, analysis with multiple single slices (i.e. two slices above/below the largest axial tumor diameter) is recommended for future research.

CONCLUSION

In conclusion, this study shows that MR-based radiomic models constructed and applied using alternative delineations cannot substitute delineations from an experienced radiologist for the prediction of LRC in OPSCC. This is in contrast to previous findings on alternative delineations for radiomics models predictive of HPV in OPSCC. The applicability of alternative delineations needs to be determined separately for each radiomics model.

SUPPLEMENTARY INFORMATION

Password: PhD_PaulaBos

REFERENCES

1. Yu K, Zhang Y, Yu Y, et al. Radiomic analysis in prediction of Human Papilloma Virus status. *Clin Transl Radiat Oncol*. 2017;7:49-54. doi:10.1016/j.ctro.2017.10.001
2. Bos P, van den Brekel MWM, Gouw ZAR, et al. Clinical variables and magnetic resonance imaging-based radiomics predict human papillomavirus status of oropharyngeal cancer. *Head Neck*. 2021;43(2):485–495. doi:10.1002/hed.26505
3. Romeo V, Cuocolo R, Ricciardi C, et al. Prediction of tumor grade and nodal status in oropharyngeal and oral cavity squamous-cell carcinoma using a radiomic approach. *Anticancer Res*. 2020;40(1):271-280. doi:10.21873/anticanres.13949
4. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006. doi:10.1038/ncomms5006
5. Yuan Y, Ren J, Shi Y, Tao X. MRI-based radiomic signature as predictive marker for patients with head and neck squamous cell carcinoma. *Eur J Radiol*. 2019;117:193-198. doi:10.1016/j.ejrad.2019.06.019
6. Chu CS, Lee NP, Adeoye J, Thomson P, Choi SW. Machine learning and treatment outcome prediction for oral cancer. *J Oral Pathol Med*. 2020;49(10):977-985. doi:10.1111/jop.13089
7. Zhang X, Zhong L, Zhang B, et al. The effects of volume of interest delineation on MRI-based radiomics analysis: Evaluation with two disease groups. *Cancer Imaging.* 2019;19(1):89. doi:10.1186/s40644-019-0276-7
8. Pavic M, Bogowicz M, Wurms X. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol*. 2018;57(8):1070-1074. doi:10.1080/0284186X.2018.1445283
9. Bos P, van den Brekel MWM, Taghavi M, et al. Largest diameter delineations can substitute 3D tumor volume delineations for radiomics prediction of human papillomavirus status on MRIs of oropharyngeal cancer. *Accepted for publication in Physica Medica*. July 2022.
10. Bos P, van den Brekel MWM, Gouw ZAR, et al. Improved outcome prediction of oropharyngeal cancer by combining clinical and MRI features in machine learning models. *Eur J Radiol*. 2021;139:109701. doi:10.1016/j.ejrad.2021.109701
11. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. Elastix: a toolbox for intensity based medical image registration. *IEEE Trans Med Imaging.* 2010;29(1):196-205. doi:10.1109/TMI.2009.2035616
12. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104–e107. doi:10.1158/0008-5472.CAN-17-0339
13. Bergstra J, Yamins D, Cox D. Hyperopt: a python library for optimizing the hyperparameters of machine learning algorithms. *Proc SciPy.* 2013;13–20.

doi:10.1088/1749-4699/8/1/014008

14. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46:389-422.
15. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945;26(3):297–302. doi:10.2307/1932409
16. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. *IEEE Trans Pattern analysis and machine intelligence*. 1993;15(9):850-863. doi:10.1109/CVPR.1992.223209
17. Liu R, Elhalawani H, Radwan Mohamed AS, et al. Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer. *Clin Transl Radiat Oncol*. 2020;21:11-18. doi:10.1016/j.ctro.2019.11.005
18. Shen C, Liu Z, Guan M, et al. 2D and 3D CT radiomics features prognostic performance comparison in non-small cell lung cancer. *Transl Oncol*. 2017;10(6):886–894. doi:10.1016/j.tranon.2017.08.007
19. Yang L, Yang J, Zhou X, et al. Development of a radiomics nomogram based on the 2D and 3D CT features to predict the survival of non-small cell lung cancer patients. *Eur Radiol.* 2019:29(5):2196–2206. doi:10.1007/s00330-018-5770-y
20. Berenguer R, Del Rosario Pastor-Juan M, Canales-Vázquez J, et al. Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters. *Radiology*. 2018;288(2):407-415. doi:10.1148/radiol.2018172361
21. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys*. 2018;102(4):1143–1158. doi:10.1016/j.ijrobp.2018.05.053
22. Foy JJ, Gertsenshteyn IH, Al-Hallaq H, Armato III SG, Sensakovic WF. Dependence of radiomics features on CT image acquisition and reconstruction parameters using a cadaveric liver. *SPIE Med Imag*. 2020;11314. doi:10.1117/12.2551155
23. van Velden FHP, Kramer GM, Frings V, et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. *Mol Imaging Biol*. 2016;18(5):788–795. doi:10.1007/s11307-016-0940-2

Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning

Physics and Imaging in Radiation Oncology;
Jul 2021;19:39-44
DOI: 10.1016/j.phro.2021.06.005

Roque Rodríguez Outeiral
Paula Bos
Abrahim Al-Mamgani
Bas Jasperse
Rita Simões
Uulke A. van der Heide

8

ABSTRACT

Background: Segmentation of oropharyngeal squamous cell carcinoma (OPSCC) is needed for radiotherapy planning. We aimed to segment the primary tumor for OPSCC on MRI using convolutional neural networks (CNNs). We investigated the effect of multiple MRI sequences as input and we proposed a semi-automatic approach for tumor segmentation that is expected to save time in the clinic.

Materials and methods: We included 171 OPSCC patients retrospectively from 2010 until 2015. For all patients the following MRI sequences were available: T1-weighted, T2-weighted and 3D T1-weighted after gadolinium injection. We trained a 3D UNet using the entire images and images with reduced context, considering only information within clipboxes around the tumor. We compared the performance using different combinations of MRI sequences as input. Finally, a semi-automatic approach by two human observers defining clipboxes around the tumor was tested. Segmentation performance was measured with Sørensen–Dice coefficient (Dice), 95th Hausdorff distance (HD) and Mean Surface Distance (MSD).

Results: The 3D UNet trained with full context and all sequences as input yielded a median Dice of 0.55, HD of 8.7 mm and MSD of 2.7 mm. Combining all MRI sequences was better than using single sequences. The semi-automatic approach with all sequences as input yielded significantly better performance ($p<0.001$): a median Dice of 0.74, HD of 4.6 mm and MSD of 1.2 mm.

Conclusion: Reducing the amount of context around the tumor and combining multiple MRI sequences improved the segmentation performance. A semi-automatic approach was accurate and clinically feasible.

INTRODUCTION

Worldwide, there are more than 679,000 new cases of head and neck cancer (HNC) per year and 380,000 of those cases result in death[1]. Radiotherapy (RT) is indicated for 74% of head and neck cancer patients, and up to 100% in some subsites[2]. Tumor delineation is needed for RT planning. In clinical practice, tumor contouring is done manually, which is time consuming and suffers from interobserver variability. Thus, accurate automatic segmentation is desirable.

Convolutional neural networks (CNNs) are considered the current state of the art for computer vision techniques, such as automatic segmentation. Specifically for tumor segmentation, promising results have been obtained for various tumor sites such as brain[3], lung[4], liver[5] and rectum[6].

For HNC, previous literature[7,8] focused on the segmentation of other RT-related target volumes rather than the primary tumor and without special focus on any particular HNC subsite, such as nasopharyngeal or oropharyngeal cancer. However, anatomy and imaging characteristics of tumors and their surrounding tissue vary greatly across subsites. Nasopharyngeal tumors are bounded by the surrounding anatomy and thus they present with lower spatial variability. Men et al.[9] proposed an automatic segmentation method for nasopharyngeal primary tumors. To the best of our knowledge, no studies have been published on automatic segmentation of primary tumors in oropharyngeal squamous cell cancer (OPSCC). Tumors in this category are quite variable in shape, size and location compared to other subsites in head and neck cancer and their delineation suffers from high interobserver variability[10].

The modalities of choice in other works for HNC automatic segmentation are PET and/or CT[7,8]. PET presents low spatial resolution and only shows the metabolically active part of the tumor while CT has low soft tissue contrast. MRI is now becoming a modality of interest in RT and provides improved soft tissue contrast compared to other modalities, being better suitable for oropharyngeal tumor segmentation. In line with this, previous works have suggested that the use of MRI for head and neck cancer delineation provides unique information compared to PET/CT or CT[11].

We investigated the effect on segmentation performance of different MRI sequences and its combination as inputs to the model. We hypothesized that by decreasing the amount of context around the tumor, thereby simplifying the task, the performance of the segmentation model would improve. Hence, we proposed a semi-automatic approach in which a clipbox around the tumor is used to crop

8

the input image. We demonstrated its clinical applicability by having two observers (including one radiation oncologist) manually selecting the clipbox. The aim of this study was to develop a CNN model for segmenting OPSCC on MRI images.

MATERIALS AND METHODS

Data

A cohort of 171 patients treated at our institute between January 2010 and December 2015 was used for this project. Mean patient age was 60 (Standard deviation ± 7 years) and 62% of the patients were male. Further details on tumor stage and HPV status can be found in the Supplemental Material Table S.1. All patients had histologically proven primary OPSCC and pre-treatment MRI, acquired for primary staging. The institutional review board approved the study (IRBd18047). Informed consent was waived considering the retrospective design. Any identifiable information was removed.

All MRI scans were acquired on 1.5T (n=79) or 3.0T (n=92) MRI scanners (Achieva, Philips Medical System, Best, The Netherlands). The imaging protocol included: 2D T1-weighted fast spin-echo (T1w), 2D T2-weighted fast spin-echo with fat suppression (T2w) and 3D T1-weighted high-resolution isotropic volume excitation after gadolinium injection with fat suppression (T1gd). Further details on the MRI protocols are given in the Supplemental Material Table S.2. The primary tumors were manually contoured in 3D Slicer (version 4.8.0, www.slicer.org) by one observer with 1 year of experience (PB). Afterwards, they were reviewed and adjusted, if needed, by a radiologist with 7 years of experience (BJ). All tumor volumes were delineated on the T1gd but observers were allowed to consult the other sequences.

For the experimental set-up, we split the data set in three subsets: training set (n=131), validation set (n=20) and test set (n=20). The test set was not used for training or hyper-parameter tuning. We stratified the three subsets for tumor volume, subsite, and aspect ratio since these features are likely relevant for segmentation. Subsites were defined as tonsillar tissue, soft palate, base of tongue and posterior wall. Aspect ratio was defined as the ratio between the shortest and the longest axis of the tumor. All images were resampled to a voxel size of 0.8 mm × 0.8 mm × 0.8 mm.

Model architecture

The UNet architecture was chosen as the basis for our experiments because of the promising results on segmentation of medical structures[5,12-15]. Given the 3D nature of the images, we chose a 3D UNet as the architecture in this work[12,16]. We used Dice as loss function[17], the Adam optimizer[18] and early stopping. Dropout and data augmentation were used for regularization. Further details on the training procedure can be found in the Supplemental Material Tables S.3 and S.4.

Fully automatic approach

We trained the 3D UNet using the full 3D scans. We studied the effect of incorporating multiple MRI sequences into the training by introducing the available MRI sequences as input channels. Five networks were trained for the following MRI sequences and combinations thereof: T1w, where the tumor is hypo-intense but homogeneous; T2w, where the tumor is hyper-intense; T1gd, since the tumor presents with clearer boundaries; combining T1gd and T2w, and combining all sequences together (T1gd, T2w and T1w), to explore all the available information.

Semi-automatic approach

We proposed a semi-automatic approach in which we trained the networks with only the information within a clipbox around the tumor instead of with the full image as input.

During training, the clipbox was computed from the tumor delineations. First, the bounding box was calculated (i.e. the minimal box around the tumor). Then, random shifts of up to 25 mm were applied to all of the six directions to make clipboxes of different sizes and allow off-centered positioning of the tumors. We considered that shifts of more than 25 mm would represent unrealistic errors during clipbox selection. Examples of inputs possibly seen by the network are shown in Figure 1.

To study the clinical feasibility of this semi-automatic approach, two human observers were asked to manually select a clipbox around the tumor for each test set patient. The clipboxes were selected using 3D Slicer on the T1gd with access to the other sequences. The first observer (PB) had delineated the tumors two years earlier. The second observer was a radiation oncologist with 16 years of experience (AA) and had no information about the tumor delineations. To mitigate the risk of the observers defining too small clipboxes, cropping the tumor, the clipboxes were dilated 5 mm so as to ensure that they encompass the tumors. We consider it unlikely that a human observer would crop the tumor by more than 5 mm.

Experiments

For the fully automatic approach, the performance of the networks trained with different sequences (T1w, T2w, T1gd, T1gd/T2w, and all sequences combined) was compared for the patients on the separate test set.

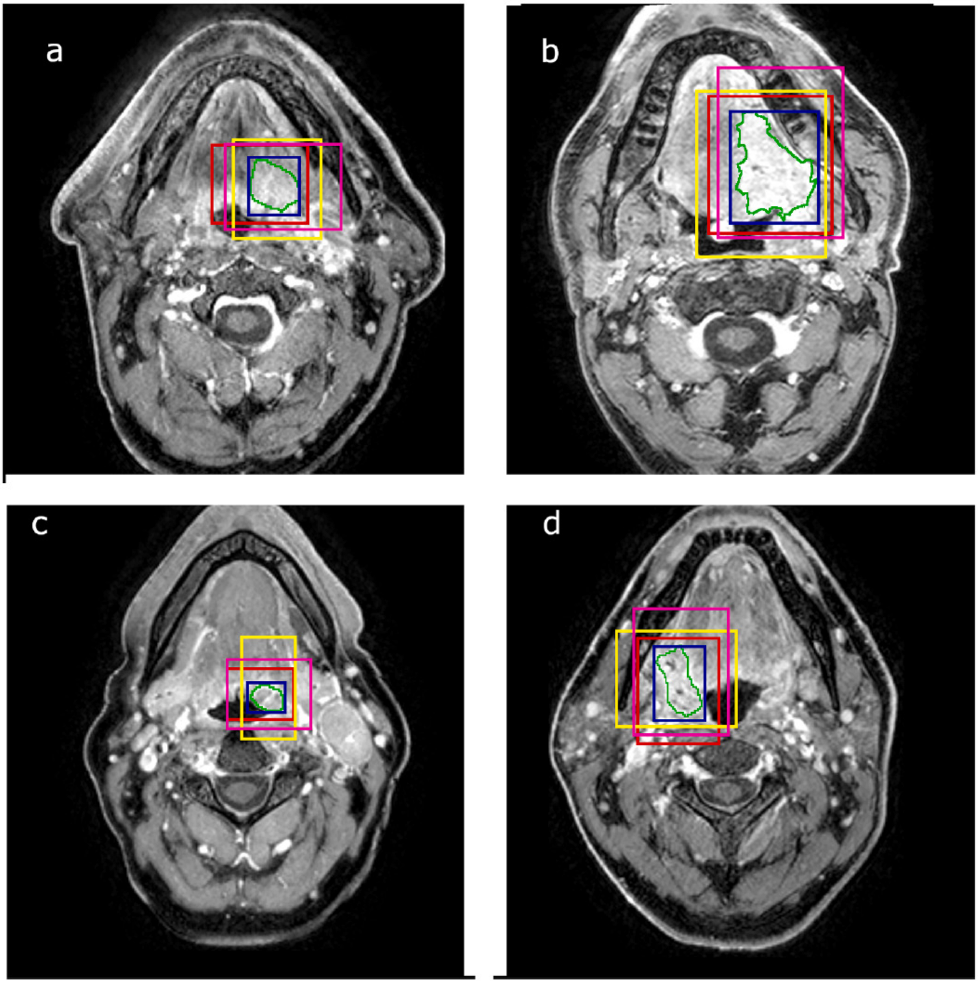


Fig. 1. Original MRI image with the manual segmentation (green) of the oropharyngeal tumor. The blue boxes are the bounding boxes of the tumor. The rest of the boxes are used as inputs to the network during training.

Because of memory constraints, scans were resized to a lower resolution by a factor of ~2.5 to 1.9 mm × 1.9 mm × 1.9 mm. Thus, even the smallest tumors were seen by the network. As a control experiment, to assess the impact of the resulting loss of resolution, we additionally trained a 2D UNet with full resolution axial slices. We checked for significant differences in performance of both approaches.

For the semi-automatic approach, one network was trained with all the sequences as input. The results with the clipboxes of the two observers were compared to the fully automatic approach experiment when combining all sequences as input (baseline).

To evaluate the robustness of the semi-automatic approach to off-centered tumors inside the clipboxes, we presented the trained model with increasingly shifted versions of the clipboxes, starting from the bounding box. The artificially induced shifts were applied in the 6 possible directions of the clipbox and expressed as two metrics: the centroid displacement and the relative difference in clipbox diagonal length before and after the shifts.

Statistics

To confirm that the three subsets were balanced in subsite, volume and aspect ratio, we used a Kruskal-Wallis test for continuous variables (volume and aspect ratio) and a chi-square test for independence for the categorical data (subsite).

Automatic contours were compared against the delineations from the human experts using common segmentation metrics: Sørensen–Dice coefficient (Dice), 95th Hausdorff Distance (HD) and Mean Surface Distance (MSD), implemented using the Python package from DeepMind (https://github.com/deepmind/surface-distance). Differences among experiments were assessed by the Wilcoxon signed-ranked test. P-values below 0.05 were considered statistically significant. Statistical analyses were performed with the SciPy package (version 1.1.0) and Python 3.6. Other relevant libraries can be found in the Supplemental Material Table S.5. The code is publicly available and can be found in: https://github.com/RoqueRouteiral/oroph_segmentation.git.

RESULTS

Summary of tumor characteristics

Tumor characteristics (location, volume and aspect ratio) of our cohort are described in Table S.6. No significant differences were found in the distributions of subsite, volume and aspect ratio between the training, validation and test sets.

Fully automatic approach

As shown in Figure 2, combining all MR sequences resulted in the best performance, with a median Dice of 0.55 (range 0–0.78), median 95th HD of 8.7 mm (range 2.8–84.8 mm) and median MSD of 2.7 mm (range 1.0–26.8 mm), and the least variability among patients. The control experiment showed that by training a 2D UNet with full resolution scans the results were not significantly better than when using its 3D counterpart (Table S.7).

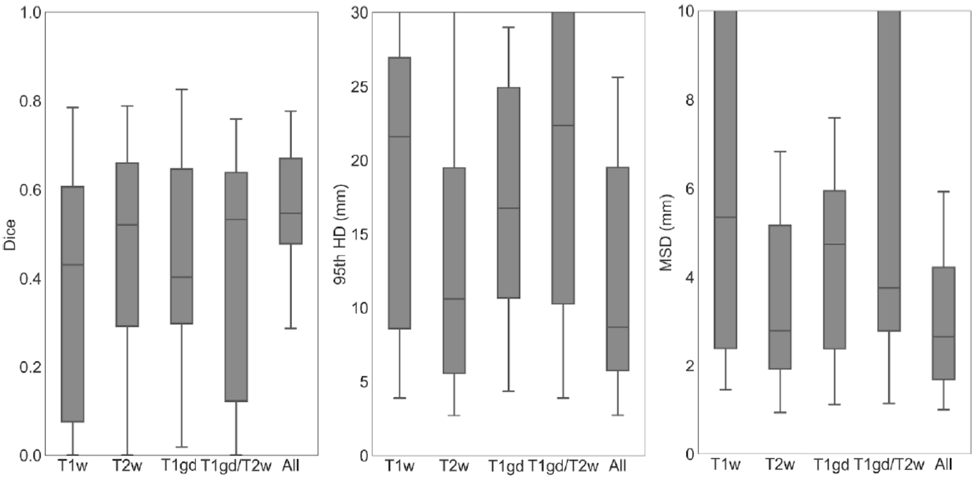


Fig. 2. Segmentation performance in terms of Dice, 95th HD and MSD for the 3D. The different boxes show different MRI sequences as input: T1w (T1 weighted), T2w (T2 weighted), T1gd (T1 3D after gadolinium injection), T1gd and T2w combined (T1gd/T2w) and all sequences combined (All). The box includes points within the interquartile range (IQR) while the whiskers show points within 1.5 times the IQR.

Semi-automatic approach

In Figure 3, it is observed that the semi-automatic approach using the boxes of the first observer achieved a median Dice score of 0.74 (range 0.32–0.80), HD of 4.6 mm (range 2.2 mm–10.5 mm) and MSD of 1.2 mm (range 0.6 mm- 2.9 mm). For the second observer, the network achieved a median Dice score of 0.67 (range 0.28–0.87), HD of 7.2 mm (range of 3.0 mm–19.9 mm) and MSD of 1.7 mm (range of 0.9 mm–4.9 mm).

The semi-automatic approach significantly outperformed the fully automatic approach in all of the metrics for the first observer (p <0.001) and in Dice and MSD for the second observer (p < 0.01). These results were expressed for 19 out of the 20 patients in the test set (also for the fully automatic approach - equivalent to “All’ in Figure 2), as one of the observers did not detect one of the tumors when asked to draw the clipbox.

The average time to draw the boxes was of 7.5 min per patient for the first observer and 2.8 min for the second observer.

Robustness to shifts

Figure 4 shows the segmentation performance of the network trained for the semi-automatic approach as a function of the artificially induced shifts applied to the

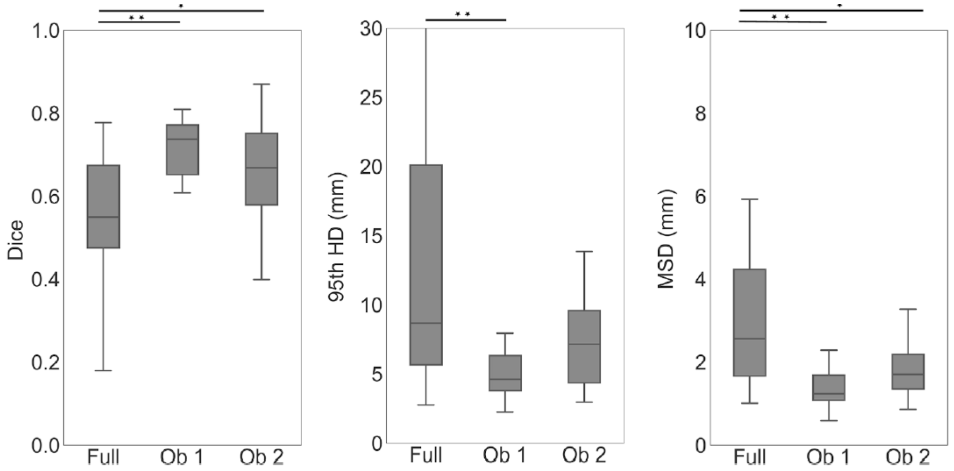


Fig. 3. Segmentation performance of the semi-automatic approach with boxes drawn by two human observers. We compare the semi-automatic results (Ob1 and Ob2) to the fully automatic approach (Full). The box includes points within the interquartile range (IQR) while the whiskers show points within 1.5 times the IQR. Significance is represented as one asterisk (*) for p<0.01 and two asterisks (**) for p<0.001.

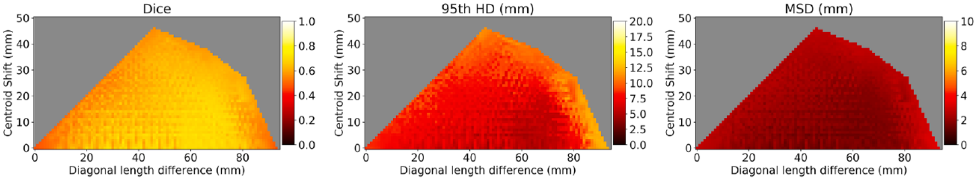


Fig. 4. Robustness analysis. Segmentation performance in terms of median Dice, 95th HD and MSD for the semi-automatic approach as a function of the tumor centroid displacement and the clipbox diagonal length difference. The grey areas correspond to undetermined values due to the geometric constraints (i.e. no combination of shifts can achieve those values of centroid displacement and diagonal length difference).

tumor within the clipbox. For centroid displacements below 20 mm and diagonal length differences of between 25 mm and 60 mm the Dice was consistently greater than 0.70, the HD was lower than 6.5 mm and the MSD was lower than 1.7 mm.

Qualitative results

Figure 5a and 5b show examples in which the shape of the semi-automatic approach output and ground truth segmentation agreed while the fully automatic approach oversegmented (a) or undersegmented (b) the tumor. Figure 5c shows a case where the segmentation by the network trained with the fully automatic approach showed a similar shape to the ground truth segmentation but there were

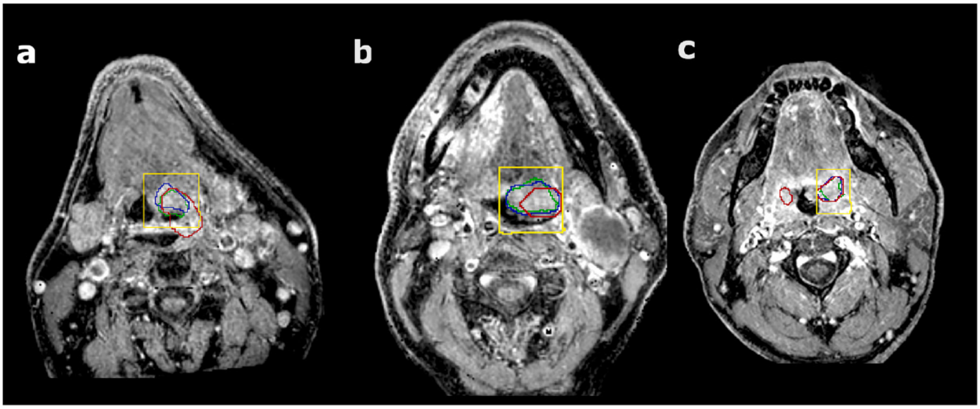


Fig. 5. Comparison of the oropharyngeal segmentations in three different patients (a, b, c) trained with the fully automatic approach (red contour), with the semi-automatic approach (blue contour) and the manual delineation (green contour). The yellow boxes are the boxes drawn by the observer.

additional false positive volumes on the image.

DISCUSSION

It was shown that using multiple MRI sequences yielded better results compared to using a single sequence as input. Also, decreasing the amount of context given to the CNN improved the segmentation performance. Finally, a functional semi-automatic approach that outperformed the fully automatic baseline was proposed and it was shown to be robust to clipbox selection errors, suggesting its potential clinical applicability.

Our network resulted in worse performance in terms of Dice compared to other tumor sites as reported by Sahiner et al.[19], where the authors provide a comparison of CNN segmentations for different tumor/lesions (Dices: 0.51–0.92). However, lower performance for oropharyngeal tumor segmentation is consistent with what is known about the inter-observer variability for this subsite: Blinde et al.[10] have shown differences in volume of up to 10 times among observers when segmenting OPSCC on MR, indicating the complexity of this task even for human observers. In this study, the mean Dice between our observers was 0.8. However, this number is an overestimation of the interobserver variability, considering that one of the observers corrected the other's delineation.

No significant differences were found between training the network with full context in 3D compared to its 2D counterpart. This shows that reducing the resolution due to memory constrains in the 3D case is not critical for the segmentation performance when the full image is used as input.

When restricting the context, the network outperformed significantly the full context approach for all metrics. This means that local textural differences between tumor and immediate surrounding tissues are sufficient for delineation.

Using clipboxes drawn by human observers demonstrates the feasibility of a semi-automatic approach for OPSCC primary tumor segmentation. Additionally, these boxes were drawn by two independent observers with different backgrounds and levels of expertise, suggesting that the method is not highly sensitive to the observer. This is supported by the results of our robustness analysis, which showed that when training with shifted versions of the clipbox, the networks were fairly robust to these shifts. More concretely, the network was robust centroid displacements below 20 mm and diagonal length differences of between 25 mm and 60 mm, which we consider a fair estimate of the maximum error an observer can make when selecting the clipbox.

A fully manual segmentation can take from 30 min to almost 2 h (depending on the shape and size of the tumor), the average time between our two observers for the semi-automatic approach can take an average of 5 min (average of our two observers). Although after the proposed semi-automatic approach, some manual adaptations may be needed by a radiation oncologists to make the contours clinically acceptable, the overall process is expected to be less labor-intensive. Additionally, in the clinic it would be possible to use software designed to draw the clipboxes faster. Consequently, a functional semi-automatic system is not only feasible in terms of segmentation performance but also relevant for speeding up the radiotherapy workflow.

There are limitations in this study. First, given the high interobserver variability of OPSCC delineation, we are likely training the network with imperfect ground truths. However, we palliated the possible errors on the delineations by having the second observer correcting the first observer’s delineation. Secondly, we used a standard 3D UNet in our studies. Despite the extensive literature on deep learning architecture modifications, investigating the best architecture for this task is outside of our scope. Thirdly, our results would need validation with an independent cohort in a multi-center study. Furthermore, the scan protocols were not standardized in our dataset. Arguably, that makes the network robust to such differences (e.g. TR/TE), given that the network has learned from a diverse dataset. Finally, our work can still be improved by adding other MRI sequences into the training (such as DWI) or by fully automatizing our semi-automatic approach, but we leave that as future work.

There is an increasing interest in the literature about differences on the tumors depending on their HPV status. According to Bos et al.[20], HPV positive tumors present on MRI post contrast with rounder shapes, lower maximum intensity values, and texture homogeneity. One strength of our work is that we include both HPV positive and HPV negative tumors in the training set, making the networks able to segment both subtypes of OPSCC. To check that the network is not biased to the HPV status, we compared the performance of the network stratified per HPV status and found non-significant results. We also did not find any relationship between performance and size.

In conclusion, this is the first study of primary tumor segmentation in the OPSCC site on MRI images with CNNs to the best of our knowledge. We trained a standard 3D UNet architecture using full MRI images as input. We showed that combining MRI sequences is beneficial for OPSCC segmentation with CNNs. Additionally, the CNN trained with reduced context around the tumor outperformed the fully automatic baseline and approaches that of other tumor sites reported in the literature. Hence, our proposed semi-automatic approach can save time in the clinic while achieving competitive performance and being robust to the choice of observer and manual clipbox selection errors.

SUPPLEMENTARY INFORMATION

Password: PhD_PaulaBos

REFERENCES

1. Fitzmaurice C, Allen C, Barber RM, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability adjusted life-years for 32 cancer groups, 1990 to 2015: A systematic analysis for the global burden of disease study. *JAMA Oncol.* 2017;3(4):524-548. doi:10.1001/jamaoncol.2016.5688
2. Delaney G, Jacob S, Barton M. Estimation of an optimal external beam radiotherapy utilization rate for head and neck carcinoma. *Cancer.* 2005;103(11):2216–2227. doi:10.1002/cncr.21084
3. Chen L, Wu Y, Dsouza AM, Abidin AZ, Wismüller A, Xu C. MRI tumor segmentation with densely connected 3D CNN. *Proc of SPIE*. 2018;10574. doi:10.1117/12.2293394
4. Li J, Chen H, Li Y, Peng Y. A novel network based on densely connected fully convolutional networks for segmentation of lung tumors on multi-modal MR images. *ACM international conference proceeding series*. 2019;1–5. doi:10.1145/3358331.3358400
5. Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA. H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes. *IEEE Trans Med Imaging.* 2018;37(12):2663–2674. doi:10.1109/tmi.2018.2845918
6. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, et al. Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. *Sci Rep.* 2017;7(1):5301. doi:10.1038/s41598-017-05728-9
7. Guo Z, Guo N, Gong K, Zhong S, Li Q. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. *Phys Med Biol.* 2019;64(20):205015. doi:10.1088/1361-6560/ab440d
8. Cardenas CE, McCarroll RE, Court LE, et al. Deep learning algorithm for auto-delineation of high-risk oropharyngeal clinical target volumes with built-in dice similarity coefficient parameter optimization function. *Int J Radiat Oncol Biol Phys.* 2018;101(2):468–478. doi:10.1016/j.ijrobp.2018.01.114
9. Men K, Chen X, Zhang Y, et al. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. *Front Oncol.* 2017;7:315. doi:10.3389/fonc.2017.00315
10. Blinde S, Mohamed ASR, Al-Mamgani A, et al. Large interobserver variation in the international MR-LINAC oropharyngeal carcinoma delineation study. *Int J Radiat Oncol.* 2017;99(2):E639–E640. doi:10.1016/j.ijrobp.2017.06.2145
11. Anderson CM, Sun W, Buatti JM, et al. Interobserver and intermodality variability in GTV delineation on simulation CT, FDG-PET, and MR Images of Head and Neck Cancer. *Jacobs J Radiat Oncol.* 2014;1(1):006.
12. Çiçek O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *MICCAI.* 2016;424–432.

doi:10.1007/978-3-319-46723-8_49

13. Zeng G, Yang X, Li J, Yu L, Heng PA, Zheng G. 3D U-net with multi-level deep supervision: Fully automatic segmentation of proximal femur in 3D MR images. In: *Machine learning in Medical Imaging.* Springer. 2017;274-282. doi:10.1007/978-3-319-67389-9_32
14. Gordienko Y, Gang P, Hui J, et al. Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer. *Adv Intell Syst Comput.* 2019;754:638–47. doi:10.1007/978-3-319-91008-6_63
15. Norman B, Pedoia V, Majumdar S. Use of 2D U-net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry. *Radiology.* 2018;288(1):177–185. doi:10.1148/radiol.2018172322
16. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *MICCAI*. Springer. 2015;234–241. doi:10.1007/978-3-319-24574-4_28
17. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Deep Learn Med Image Anal Multimodal Learn Clin Decis Support*. 2017;240-248. doi: 10.1007/978- 3-319-67558-9_28
18. Kingma DP, Ba JL. Adam: A method for stochastic optimization. *ICLR*. 2015. doi:10.48550/arXiv.1412.6980
19. Sahiner B, Pezeshk A, Hadjiiski LM, et al. Deep learning in medical imaging and radiation therapy. *Med Phys.* 2019;46(1):e1–36. doi:10.1002/mp.13264
20. Bos P, van den Brekel MWM, Gouw ZAR, et al. Clinical variables and magnetic resonance imaging-based radiomics predict human papillomavirus status of oropharyngeal cancer. *Head Neck*. 2021;43(2):485–495. doi:10.1002/hed.26505

General discussion

9

Prognostic factors are necessary to categorise oropharyngeal cancer patients into well-defined groups, to optimally inform patients, select the most appropriate treatment and to be able to compare treatment results. Radiomics is a prognostic tool which can analyse and quantify image texture and relate it to tumor phenotype. This thesis describes the potential, pitfalls and opportunities of MR-based radiomics in primary OPSCC treated with chemoradiation therapy.

Potential: A representation of tumor biology

We found that radiomics markers can predict which OPSCC patients will respond to (chemo)radiotherapy treatment (chapter 3) and which OPSCC patient has tumoral human papillomavirus (HPV) (chapter 5). The potential of radiomics is not limited to these particular goals, but also extends to the classification of tumor biology[1–4] and profiling of tumor genetics[5–8], as described in various published studies. Radiological characteristics in current clinical practice are limited to semantic features, such as tumor shape, infiltration into surrounding tissues, the presence of cystic, necrotic or hypoxic areas, and lymph node status. Aforementioned studies show that radiomic quantitative features can provide complementary information to these semantic features, which might assist the clinician in treatment decisions in the future.

In line with this, we have also shown that performance of models based on the combination of clinical variables and radiomic features was superior to models based on only clinical variables or radiomic features (chapter 3, chapter 5)[9]. It is likely that clinical variables and radiomic features hold independent and complementary information. Most of these clinical variables are related to patient and treatment factors as well as tumor staging, whereas tumor structure and shape are more represented in radiomic features. This does not mean that all radiomics models should include radiomic features as well as clinical variables. As shown in chapter 5, a model based on clinical variables alone can be quite reliable (AUC: 0.79) in comparison to a model based on radiomics and clinical variables (AUC: 0.87), with the advantage of easy implementation without the requirement of time consuming tumor delineations on imaging. A critical and practical eye is necessary to determine if a model needs to be based on radiomic features, clinical variables or both.

Pitfall: Heterogeneity in radiomic workflow settings

Clearly, radiomics in prognosis and treatment stratification seems promising. The next big challenge is its implementation in the daily clinical routine. Before implementation in daily clinical practice, radiomics has to overcome some crucial pitfalls.

Firstly, representative patient cohort with adequate sample size is required to train a radiomic model that is appropriate in general clinical practice. This was demonstrated in models developed on a subgroup of HPV positive or HPV negative patients, both of which performed worse compared to if these groups were combined (chapter 3). The limited number of events probably resulted in the inability of the model to predict the endpoint robustly[10]. Additionally, model performance increases when the model was validated on patient groups matched for comparable demographics as the training cohort (chapter 4), inducing that inclusion and exclusion criteria of patients affect model reproducibility[11].

Another point of concern is that variability in image acquisition decreases radiomic feature reproducibility[12,13]. Standardization of acquisition protocols improves applicability of radiomic protocols, however, it limits development of (new) acquisition protocols. Harmonization in the image domain can reduce the influence of image acquisition settings between multiple vendors. This harmonization uses the transition of voxel intensities from a reference MR image towards the new acquired MR image. It is important that this has to be calculated on a sufficient sample size[14,15] before applying the trained model on a prospective patient (with different acquisition parameters).

Manual tumor delineations are still required to obtain radiomic features from the region of interest. Current radiology practice is demanding and radiologists do not have time to perform detailed tumor delineations. Speeding up this process would be an important step for wide adoption of a radiomic pipeline in clinical practice. Automatic tumor delineations have the advantage that these are less time consuming, repeatable, reproducible, and potentially improving the accuracy of the radiomic workflow by decreasing inter-observer variability[16]. Although we have proven that automatic segmentation is possible in this challenging anatomic area, the automatic delineations still requires a rough and simple manual indication of the tumor region by the radiologist to make the delineation more precise (chapter 8). Another method to speed-up the radiomic workflow is simplification of manual tumor delineations by drawing a sphere inside the tumor area or by delineating the tumor boundaries on only the slice with the largest axial tumor diameter. However, the ability to use these simplified methods as an alternative highly depends on the radiomic signature and their interaction on image interpolation (chapter 6, chapter 7)[17]. Besides speeding up the manual tumor delineations, a note has to be made that not all tumors are reliable in radiomic analysis. Small tumors (e.g. sub-centimeter nodules)[18] may not provide sufficient voxel information and should be either excluded in radiomic analysis or analysed voxel by voxel[18]. Moreover, necrotic and hypoxic areas may not representative for tumor tissue and should be excluded during tumor delineations for reliable radiomics analysis (chapter 2)[19].

Variation in radiomic performance is not only driven by image acquition and tumor delineation, but also by choices made in the machine learning pipeline[12,13,20,21]. In more detail, the selected classifier and feature selection methodology are responsible for 29% and 14% of the total variance, respectively[22]. Due to the high dimensionality of radiomic features, feature reduction is performed before feature selection (e.g. "stable features" in this thesis). The drawback of this step is the dependency on significance levels derived from the trained patient data, lacking reproducibility when applied on external datasets (chapter 4). Performing radiomic research using multicentre data might therefore be a possible solution.

A final remark regarding decreasing heterogeneity and improving clinical adoption of radiomics has to be made on quality control. Fair evaluation and comparison of published models is only possible when choices of the radiomic pipeline are well reported[23] and quality of the research can be assessed with an objective (quality) score (e.g. radiomics quality score[24]).

Opportunity: Creating the future for radiomics in OPSCC patients

Radiomics in its current format is not ready to support the radiologists to make precision diagnoses, or to provide oncologists a reliable clinical decision support tool[11]. Adaptations on each step of the radiomic workflow combined with innovations are crucial to improve reproducibility, repeatability and applicability of radiomics in clinical practice and create a future for radiomics for OPSCC patients.

The first step imperative for radiomics as clinically wide adoption are multicenter studies and large-scale validation studies. The most important reason for the lack of large multicentre populations (only 19.53% of all radiomic studies[25]) are legal and ethical privacy concerns associated with medical data sharing. Distributed learning facilitates data sharing without personal data leaving the institute[26,27]. In this approach, each institute trains a model based on their local data and sends the resulting model parameters to the central server. This central server compares the model parameters from each institute and returns the updated parameters to the individual institutes for further optimization. This iteratively privacy-preserving process only shares mathematical parameters (e.g. metadata), which cannot be traced to individual patients. There are already a number of studies that has shown the feasibility of distributed learning in real-world multi-institutional setting[26,28].

Omission of the time-consuming, observer-dependent tumor delineations is the second step towards clinical practice of radiomics. Simpler, faster or automatic tumor delineations might be more reproducible and repeatable and therefore, improving homogeneity of the radiomics workflow. Another alternative is deep learning, which can train radiomic models directly from MR images, without the need of tumor delineations. The output of their layers, the “deep features”, are used to predict an endpoint[29]. Besides this, deep learning algorithms could also be helpful in other stages of the radiomic process, such as image quality assessment or harmonizing between images[29,30].

Another step to implement radiomics in a clinical setting is proven adequate performance in a prospective trial. In such trials, the outcome of the radiomic tool has to be compared next to the outcome decided by clinicians. When the radiomics tools seems to be accurate enough for personalized medicine, it can be implemented in a clinical workflow. Additionally, further optimizing of the radiomic tool can be investigated.

The current format of radiomics can be elaborated with information of medical images acquired at multiple time moments to evaluate progression of the disease during the treatment period. This “delta-radiomics” enables the possibility to adapt the ongoing treatment strategy[31,32] or anticipate on undesirable side effects (e.g. xerostomia[33,34]). This requires MR images that are acquired with identical protocol settings, to ensure that radiomic features can be assumed to solely represent changes in tissue characteristics.

A final innovation is to include other available multi-disciplinary variables (“multi-omics”) in the prediction model to obtain a complete presentation of tumor behaviour. Recent studies conducted already the interrelations of radiomic and genomic features[35,36] and histopathology[37,38]. In patients with lung cancer, genetic information and radiomic features were combined to predict treatment outcome[39]. Combining information hidden in each multi-disciplinary evaluation might be complementary to each other, thereby improving prediction performance through a comprehensive representation of the underlying tumor biology[40,41].

In conclusion, this thesis shows the potential of radiomics to reliably predict if a patient will respond to chemoradiation treatment and determine HPV tumor characteristics in OPSCC patients. Despite this potential, the current state of radiomics needs improvement and standardization before clinical implementation, which include improvements in reproducibility, repeatability and generalizability. To do so, the essential first step is to perform multicenter large-scale validation studies. In line with this, the current thesis has already shown that a monocentric prediction model was generalizable in an external validation cohort. The second step is focused on the omission of the time-consuming, observer-dependent manual tumor delineation to improve radiomic workflow. This thesis proved that alternative delineations (faster, simpler tumor delineations) can substitute these manual delineations for at least some but not all radiomic models. Automatic tumor delineations can also be considered in this regard, although these techniques still need improvement in this challenging anatomical area. For the future, innovations such as multi-omics prediction models, delta-radiomics and deep learning approaches may greatly extend the possibilities of radiomics.

REFERENCES

1. Fruehwald-Pallamar J, Hesselink JR, Mafee MF, Holzer-Fruehwald L, Czerny C, Mayerhoefer ME. Texture-Based analysis of 100 MR examinations of head and neck tumors-Is it possible to discriminate between benign and malignant masses in a multicenter trial? *Rofo*. 2016;188(2):195-202. doi:10.1055/s-0041-106066
2. Zheng YM, Xu WJ, Hao DP, et al. A CT-based radiomics nomogram for differentiation of lympho-associated benign and malignant lesions of the parotid gland. *Eur Radiol*. 2021;31(5):2886-2895. doi:10.1007/s00330-020-07421-4
3. Romeo V, Cuocolo R, Ricciardi C, et al. Prediction of tumor grade and nodal status in oropharyngeal and oral cavity squamous-cell carcinoma using a radiomic approach. *Anticancer Res*. 2020;40(1):271-280. doi:10.21873/anticanres.13949
4. Yu K, Zhang Y, Yu Y, et al. Radiomic analysis in prediction of Human Papilloma Virus status. *Clin Transl Radiat Oncol*. 2017;7:49-54. doi:10.1016/j.ctro.2017.10.001
5. Zhu Y, Mohamed ASR, Lai SY, et al. Imaging-Genomic Study of Head and Neck Squamous Cell Carcinoma: Associations Between Radiomic Phenotypes and Genomic Mechanisms via Integration of The Cancer Genome Atlas and The Cancer Imaging Archive. *JCO Clin Cancer Inform*. 2019;3:1-9. doi:10.1200/cci.18.00073
6. Choi JW, Lee D, Hyun SH, Han M, Kim JH, Lee SJ. Intratumoural heterogeneity measured using FDG PET and MRI is associated with tumour–stroma ratio and clinical outcome in head and neck squamous cell carcinoma. *Clin Radiol*. 2017;72(6):482-489. doi:10.1016/j.crad.2017.01.019
7. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006. doi:10.1038/ncomms5006
8. Chu CS, Lee NP, Adeoye J, Thomson P, Choi SW. Machine learning and treatment outcome prediction for oral cancer. *J Oral Pathol Med*. 2020;49(10):977-985. doi:10.1111/jop.13089
9. Zhai TT, Langendijk JA, van Dijk LV, et al. Pre-treatment radiomic features predict individual lymph node failure for head and neck cancer patients. *Radiother Oncol*. 2020;146:58-65. doi:10.1016/j.radonc.2020.02.005
10. Chalkidou A, O’Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: A systematic review. *PLoS One*. 2015;10(5):e0124165. doi:10.1371/journal.pone.0124165
11. Liu Z, Wang S, Dong D, et al. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. *Theranostics*. 2019;9(5):1303-1322. doi:10.7150/thno.30309
12. Pfaehler E, Zhovannik I, Wei L, et al. A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features. *Phys Imaging Radiat Oncol*. 2021;20:69-75. doi:10.1016/j.phro.2021.10.007

13. Berenguer R, Del Rosario Pastor-Juan M, Canales-Vázquez J, et al. Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters. *Radiology*. 2018;288(2):407-415. doi:10.1148/radiol.2018172361

14. Lambin P, Roelofs E, Reymen B, et al. "Rapid Learning health care in oncology" - An approach towards decision support systems enabling customised radiotherapy. *Radiother Oncol*. 2013;109(1):159-164. doi:10.1016/j.radonc.2013.07.007

15. Qiu X, Wu H, Hu R. The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. *BMC Bioinformatics*. 2013;14:124. doi:10.1186/1471-2105-14-124

16. van der Veen J, Willems S, Deschuymer S, et al. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiother Oncol*. 2019;138:68-74. doi:10.1016/j.radonc.2019.05.010

17. Park SH, Lim H, Bae BK, 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. doi:10.1186/s40644-021-00388-5

18. Meyer HJ, Hamerla G, Höhn AK, Surov A. CT texture analysis-correlations with histopathology parameters in head and neck squamous cell carcinomas. *Front Oncol*. 2019;9:444. doi:10.3389/fonc.2019.00444

19. Limkin EJ, Sun R, Dercle L, et al. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann Oncol*. 2017;28(6):1191-1206. doi:10.1093/annonc/mdx034

20. Liu R, Elhalawani H, Radwan Mohamed AS, et al. Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer. *Clin Transl Radiat Oncol*. 2020;21:11-18. doi:10.1016/j.ctro.2019.11.005

21. Parmar C, Leijenaar RTH, Grossmann P, et al. Radiomic feature clusters and Prognostic Signatures specific for Lung and Head & Neck cancer. *Sci Rep*. 2015;5:11044. doi:10.1038/srep11044

22. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJWL. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Front Oncol*. 2015;5:272. doi:10.3389/fonc.2015.00272

23. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *BMJ*. 2015;350:g7594. doi:10.1136/bmj.g7594

24. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762. doi:10.1038/nrclinonc.2017.141

25. Song J, Yin Y, Wang H, Chang Z, Liu Z, Cui L. A review of original articles published in the emerging field of radiomics. *Eur J Radiol*. 2020;127:108991. doi:10.1016/j.ejrad.2020.108991

26. Deist TM, Jochems A, van Soest J, et al. Infrastructure and distributed learning

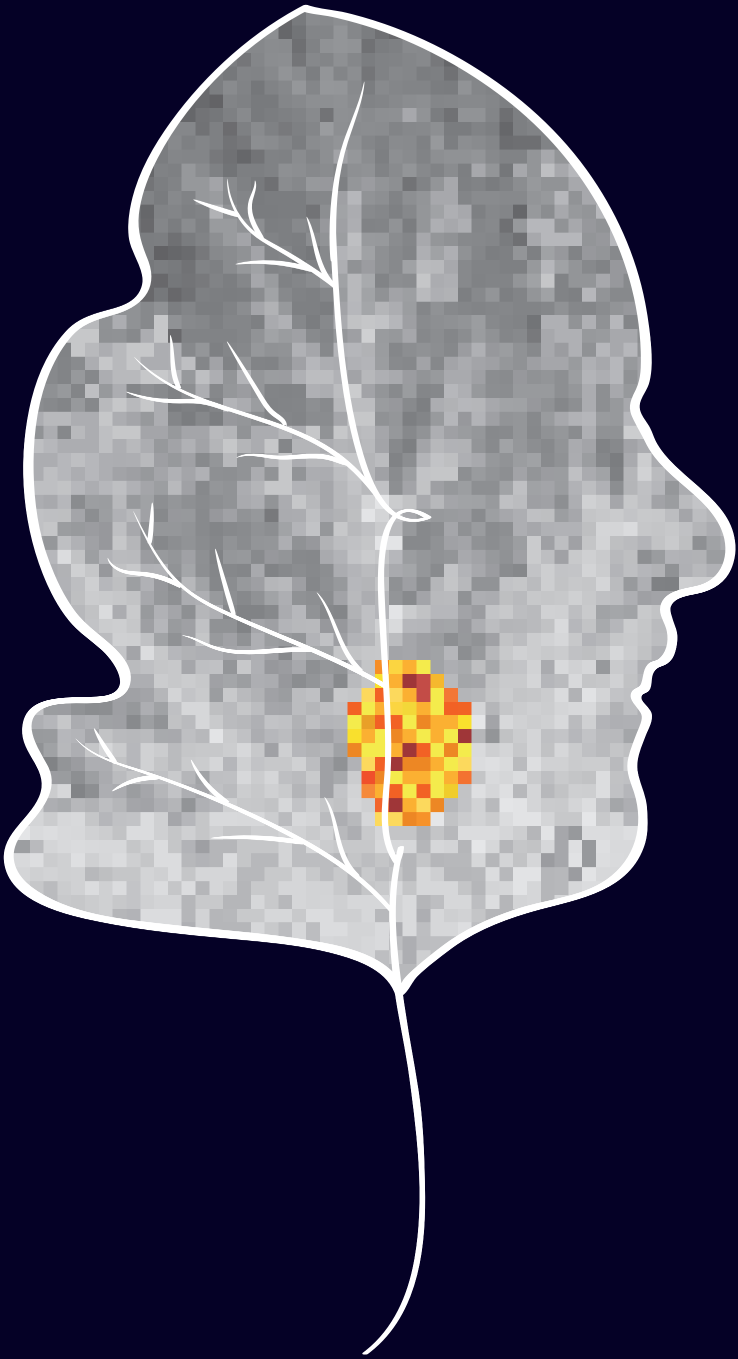
methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. *Clin Transl Radiat Oncol*. 2017;4:24-31. doi:10.1016/j.ctro.2016.12.004

27. Zerka F, Barakat S, Walsh S, et al. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. *JCO Clin Cancer Inform*. 2020;4:184-200. doi:10.1200/CCI.19.00047
28. Price G, van Herk M, Faivre-Finn C. Data Mining in Oncology: The ukCAT Project and the Practicalities of Working with Routine Patient Data. *Clin Oncol*. 2017;29(12):814-817. doi:10.1016/j.clon.2017.07.011
29. Diamant A, Chatterjee A, Vallières M, Shenouda G, Seuntjens J. Deep learning in head & neck cancer outcome prediction. *Sci Rep.* 2019;9(1):2764. doi:10.1038/s41598-019-39206-1
30. Avanzo M, Wei L, Stancanello J, et al. Machine and deep learning methods for radiomics. *Med Phys*. 2020;47(5):e185-e202. doi:10.1002/mp.13678
31. Nardone V, Reginelli A, Grassi R, et al. Delta radiomics: a systematic review. *Radiol Medica*. 2021;126(12):1571-1583. doi:10.1007/s11547-021-01436-7
32. Fatima K, Dasgupta A, DiCenzo D, et al. Ultrasound delta-radiomics during radiotherapy to predict recurrence in patients with head and neck squamous cell carcinoma. *Clin Transl Radiat Oncol*. 2021;28:62-70. doi:10.1016/j.ctro.2021.03.002
33. van Dijk LV, Langendijk JA, Zhai TT, et al. Delta-radiomics features during radiotherapy improve the prediction of late xerostomia. *Sci Rep*. 2019;9(1):12483. doi:10.1038/s41598-019-48184-3
34. Liu Y, Shi H, Huang S, et al. Early prediction of acute xerostomia during radiation therapy for nasopharyngeal cancer based on delta radiomics from CT images. *Quant Imaging Med Surg*. 2019;9(7):1288-1302. doi:10.21037/qims.2019.07.08
35. Wu L, Lin P, Zhao Y, Li X, Yang H, He Y. Prediction of Genetic Alterations in Oncogenic Signaling Pathways in Squamous Cell Carcinoma of the Head and Neck: Radiogenomic Analysis Based on Computed Tomography Images. *J Comput Assist Tomogr*. 2021;45(6):932-940. doi:10.1097/RCT.0000000000001213
36. Zwirner K, Hilke FJ, Demidov G, et al. Radiogenomics in head and neck cancer: correlation of radiomic heterogeneity and somatic mutations in TP53, FAT1 and KMT2D. *Strahlenther Onkol*. 2019;195(9):771-779. doi:10.1007/s00066-019-01478-x
37. Bogowicz M, Riesterer O, Ikenberg K, 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):921-928. doi:10.1016/j.ijrobp.2017.06.002
38. Leijenaar RTH, Bogowicz M, Jochems A, 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. doi:10.1259/bjr.20170498
39. Kirienko M, Sollini M, Corbetta M, et al. Radiomics and gene expression profile to

characterise the disease and predict outcome in patients with lung cancer. *Eur J Nucl Med Mol Imaging*. 2021;48(11):3643-3655. doi:10.1007/s00259-021-05371-7

40. Marcu LG, Marcu DC. Current omics trends in personalised head and neck cancer chemoradiotherapy. *J Pers Med*. 2021;11(11). doi:10.3390/jpm11111094

41. Zeng H, Chen L, Huang Y, Luo Y, Ma X. Integrative Models of Histopathological Image Features and Omics Data Predict Survival in Head and Neck Squamous Cell Carcinoma. *Front Cell Dev Biol*. 2020;8:553099. doi:10.3389/fcell.2020.553099



Appendices

SUMMARY

Oropharyngeal Squamous Cell Carcinoma (OPSCC) patients are currently mostly treated with chemoradiation therapy, which is successful in 55-75% of cases. Treatment success is partially explained by the extent of the local tumor and seeding of tumor to lymph nodes and other parts of the body. These disease characteristics cannot fully explain treatment success, calling for other reliable biomarkers to predict treatment response.

Recent advances in machine learning technique have made it possible to quantify tumor characteristics on medical images that cannot be appreciated by visual inspection and is commonly referred to as radiomics. These quantitative image characteristics may reveal information on tumor biology relevant for pathological classification and treatment outcome. Statistical methods applied to large groups of patients, of which the treatment outcome or pathological classification is known, can be used to find out which of these image characteristics are indeed relevant. The resulting model can then be used to predict treatment outcome or pathology for new OPSCC patients. This thesis aims to explore the potential of radiomics to predict treatment outcome and pathological classification (human papillomavirus (HPV)) for OPSCC patients using pre-treatment diagnostic magnetic resonance imaging (MRI).

Part I: Current knowledge of MR-based functional parameters in head and neck squamous cell carcinoma

Functional MRI parameters, such as perfusion and diffusion parameters, can be extracted and related to tumor characteristics. Chapter 2 provides a literature review describing prognostic pre-treatment perfusion and diffusion parameters extracted from the primary tumor of HNSCC patients . A total of 31 studies were included for quantitative analysis. Among them, 11 and 28 studies assessed perfusion and diffusion parameters, respectively. While diffusion prognosticators were studied more frequently compared to perfusion parameters, study results show high discrepancy, asking for standardization within image acquisition, tumor segmentation methodology and statistical analysis.

Part II: MR-based radiomic prediction models for tumor characterization and prognosis in OPSCC patients

To assess the feasibility of MR-based radiomics in OPSCC patients, predictive models were build and validated in chapter 3 to 5. First, predictive models for treatment outcome after chemoradiation therapy were developed and tested using a single-centre cohort of 177 OPSCC patients (chapter 3). Models were build

based on solely clinical variables, solely radiomic features and its combination. Radiomics models were able to predict treatment outcome; however, a model combining clinical variables and radiomic features outperforms models based on solely clinical variables/radiomic features. This implies that clinical variables and radiomic features hold independent information for outcome prediction. Sub-analysis with patients having HPV-negative tumors did not reach consistent meaningful predictive tumor properties due to the low number of patients.

Application of developed prediction models in patients originating from other hospitals is of high relevance for clinical utility. Chapter 4 shows that prediction performance drops slightly when our model predictive of treatment outcome was validated on data originating from an independent external center. Several factors affecting the performance were studied, where tumor delineation strategies and poorly reproducible features negatively influence generalizability. Generalizability increased when the model was validated on a patient subset matching the patient demographics or acquisition parameters of the trained dataset. Transforming the data from the external center towards the data used for training the model (data harmonization) also improves prediction performance.

In chapter 5, models based on clinical variables and/or radiomic features were built to non-invasively predict tumoral HPV status. Radiomic features show that patients with HPV positivity had rounder tumors with a higher texture homogeneity, reflecting the tumor biology with its less-invasive exophytic growth and non-keratinizing histopathology. Although a model based on clinical variables and radiomic features performs best, a model based on solely clinical variables would be the method of choice due to its ease of implementation. Findings needs to be externally validated due to the single-centre approach of this study.

Part III: Simplification or automatization of delineation techniques to improve clinical adoption of MR-based radiomics for OPSCC patients

A crucial factor hampering clinical adoption of radiomics are the required manually performed whole tumor delineations, which are laborious, time-consuming and user-dependent. Therefore, already available or simpler alternative tumor delineations are necessary. Chapter 6 investigates six manual delineation strategies in radiomic models predictive of HPV, including delineations performed by a non-experienced observer, an experienced radiologist and a radiation therapist. Besides, “simple” tumor delineations were evaluated where the tumor was delineated on the slice with the largest axial diameter or delineated by simple spherical tumor volumes (with a diameter of 4mm and largest possible diameter of the tumor). Findings show that less labour-intensive, easily applicable delineations can substitute the experienced tumor volume delineations, particularly the 2D single-slice delineations. Since the radiomic signature is unique for each outcome parameter, prediction performance might differ when alternative delineations are used.

Simpler delineations were not able to substitute whole-tumor tumor delineations performed by an experience radiologist when the same alternative delineation strategies were applied in a radiomic model predictive of treatment outcome (chapter 7). A first explanation for this contradictory can be the effect of image interpolation on radiomic features. Image interpolation on shape-based features will result towards coarser tumor shapes, which are still representative for tumor contour. Non-shape-based features, in contrast, undergo changes in voxel intensity values as result of image interpolation, resulting in alternated radiomic feature. A model predictive of treatment outcome consist mostly of shape-based radiomic features, where HPV status is mainly predicted by non-shape radiomic features, explaining the different results when alternative delineations are applied. Additionally, peripheral surface information, like tumor invasion, is more relevant in the distinguishing of treatment outcome compared to determination of HPV status implying that not enough information can be obtained in a single slice delineation to predict treatment outcome. The different conclusions for models predictive of HPV and treatment outcome mirror the need to explore the impact of every single factor of the radiomic pipeline on feature variability before radiomics can be implemented in a clinical setting.

Another solution for the user-dependent, time-consuming manual tumor delineations can be automatic tumor delineation. Chapter 8 describes a deep learning (3D UNet) architecture to obtain (semi-)automatically primary OPSCC tumor delineations. An architecture designed on multiple MRI sequences shows the highest spatial overlap compared to a tumor delineation obtained from a human observer (fully automatic approach), indicating that each MR sequence holds exclusive characteristics with regard to voxel intensity to identify tumor tissue. Manual placement of a clipbox covering the tumor (semi-automatic approach) improves this agreement and significantly reduces the workload of the clinician (30 min-2 hours vs 5 minutes), enlarging the feasibility of meaningful clinical quantitative analysis, such as radiomics.

SAMENVATTING

Stel je voor dat je werkt als KNO-arts. Vorig jaar heb je twee patiënten behandeld die op de polikliniek kwamen met vergelijkbare patiëntgeschiedenis en symptomen, namelijk keelpijn, heesheid, moeite met slikken en een zwelling in de hals. Na klinische en diagnostische evaluatie werden beide patiënten gediagnosticeerd met lokaal gevorderd keelkanker. Een zeven-week durend traject van radiotherapie in combinatie met chemotherapie werd ingezet als behandelplan voor beide patiënten. Nu, een jaar later, is de ene patiënt volledig genezen, terwijl de andere patiënt vijf maanden na het voltooien van de behandeling stierf aan een lokaal recidief. Twee vergelijkbare patiënten, twee identieke casussen, maar toch twee verschillende behandelingsuitkomsten. Als de kennis van het behandelingsresultaat van tevoren bekend geweest was, hadden we van tevoren kunnen identificieren voor welke patiënt de behandeling voldoende was, welke patiënten baat hadden bij intensivering van de behandeling en voor welke patiënten de behandeling meer kwaad dan goed had gedaan (door bijwerkingen). Met andere woorden: we zouden een behandeling op maat voor de patiënt kunnen geven, ofwel gepersonaliseerde behandeling.

De meeste patienten met (plaveiselcel) keelkanker worden behandeld met radiotherapie in combinatie met chemotherapie (chemoradiatie therapie). Helaas is deze behandeling maar in 55-75% van de gevallen succesvol. De geselecteerde behandeling (en het succes ervan) wordt bepaald aan de hand van patiëntfactoren en klinische en diagnostische evaluatie. De klinische evaluatie omvat anamnese en lichamelijk onderzoek. Tevens kan er kijkonderzoek plaatsvinden, waar mogelijk een stukje weefsel weggenomen kan worden (biopsie) om te analyseren. Bij de diagnostische evaluatie wordt er gekeken met behulp van medische beeldvorming (bijv. MRI of CT). Hierbij wordt de grootte van de lokale tumor (T), de aanwezigheid van tumorcellen in regionale lymfeklieren (N) of in andere lichaamsdelen (M) (TNM-stadiëring) geanalyseerd. Daarnaast kunnen ook functionele tumoreigenschappen (tumordiffusie/perfusie) uit deze medische beeldvorming verkregen worden. Echter kan het succes van de behandeling niet volledig worden verklaard door al deze ziektekenmerken (biomarkers), waardoor andere betrouwbare biomarkers vereist zijn die voorspelling van het resultaat van de behandeling kunnen optimaliseren.

Zowel anatomische als functionele informatie van de tumor kunnen worden verkregen door beoordeling van medische beeldvorming. Echter is er veel meer informatie verborgen in deze medische beelden, die niet met het blote oog zichtbaar zijn. Radiomics is een kwantitatieve methode die een groot aantal zichtbare en “verborgen” beeldeigenschappen kan verkrijgen uit medische beeldvorming. Deze kwantitatieve beeldeigenschappen (“radiomics beeldeigenschappen”) kunnen informatie over de biologie van de tumor onthullen die relevant zijn voor pathologische classificatie en/of behandelresultaten. Innovatie in technieken van machinaal leren hebben het mogelijk gemaakt om te achterhalen welke radiomics beeldeigenschappen relevant zijn voor deze voorspelling. Dit wordt gedaan door statistische methoden toe te passen op grote groepen patiënten, waarvan de behandeluitkomst of pathologische classificatie bekend is. Het resulterende model kan vervolgens worden gebruikt om het behandelresultaat of pathologie voor nieuwe keelkanker patiënten te voorspellen.

Het doel van deze thesis was om het potentieel van radiomics te onderzoeken om het behandelresultaat en pathologische classificatie (humaan papillomavirus (HPV)) voor keelkanker patiënten te voorspellen met behulp van diagnostische magnetische resonantie beeldvorming (MRI) die voor de behandeling van de patiënt is gemaakt.

Deel I: Huidige kennis van op het gebied van MR-gebaseerde functionele parameters bij plaveiselcelcarcinoom van hoofd-hals patiënten

Functionele MRI-parameters, zoals perfusie- en diffusieparameters, kunnen worden geëxtraheerd en gerelateerd aan tumorkenmerken. Hoofdstuk 2 geeft een literatuuroverzicht van perfusie- en diffusieparameters die prognostisch zijn voor de behandeluitkomst. Alleen parameters verkregen uit de primaire tumor van hoofd-hals kanker patiënten beoordeeld op MRI-diagnostiek afgenomen voor de behandeling van de patiënt zijn meegenomen. In totaal werden 31 studies gebruikt voor de kwantitatieve analyse, waaronder respectievelijk 11 en 28 studies perfusie- en diffusieparameters onderzochten. Hoewel prognostische diffusieparameters vaker werden onderzocht in vergelijking met perfusieparameters, vertonen beide parameters grote discrepantie binnen de onderzoeksopzet van de meegenomen studies. Hierdoor wordt standaardisatie binnen de beeldacquisitie, tumorsegmentatie methoden en statistische analyse in de toekomst streng aanbevolen.

Deel II: MR-gebaseerde radiomics predictiemodellen voor tumorkarakterisering en prognose bij keelkanker patiënten

Om de haalbaarheid van MR-radiomics in keelkanker patiënten te beoordelen werden voorspellende modellen ontwikkeld en gevalideerd in hoofdstuk 3 tot en met 5. In hoofdstuk 3 werden modellen ontwikkeld en getest die voorspellend zijn voor het behandelresultaat na chemoradiatie met behulp van een single-center cohort bestaande uit 177 keelkanker patiënten. Modellen werden gebouwd op basis van uitsluitend klinisch variabelen (bijv. leeftijd, geslacht), uitsluitend radiomics beeldeigenschappen (bijv. vorm, rondheid, heterogeneiteit in MR intensiteit) en de combinatie hiervan. Radiomics modellen waren in staat om de uitkomst van de behandeling te voorspellen, maar een model dat klinische variabelen én radiomics beeldeigenschappen combineert presteert beter dan modellen die uitsluitend zijn gebaseerd op klinische variabelen of radiomics beeldeigenschappen. Dit houdt in dat klinische variabelen en radiomics beeldeigenschappen onafhankelijke informatie bevatten voor het voorspellen van de behandeluitkomst. Sub-analyse bij patiënten met HPV-negatieve tumoren leverde geen consistente, betekenisvolle voorspellende eigenschappen op vanwege het lage aantal patiënten welke geanalyseerd zijn.

Medische beeldvorming tussen ziekenhuizen kan verschillen, door de grote variatie in merk van de scanner, type van de scanner en acquisitie protocols. Daarom is het niet vanzelfsprekend dat een voorspellend model ontwikkeld op data van patiënten afkomstig uit één specifiek ziekenhuis (mono-center) kan worden toegepast op patiënten afkomstig van andere ziekenhuizen. Externe validatie van deze modellen is daarom van groot belang om de klinische bruikbaarheid (generaliseerbaarheid) te bewijzen. Hoofdstuk 4 laat zien dat de prestatie van het model voorspellend voor behandeluitkomst licht afneemt wanneer het gevalideerd werd op data afkomstig van een onafhankelijke externe centrum. Om de reden hiervoor te achterhalen, werden verschillende factoren die de prestatie beïnvloeden bestudeerd, waarbij strategieën voor tumor segmentatie en slecht reproduceerbare radiomics beeldeigenschappen de generaliseerbaarheid negatief beïnvloeden. De generaliseerbaarheid nam toe wanneer het model werd gevalideerd op een subset van patiënten met overeenkomstige demografische karakteristieken of acquisitieparameters overeenkomstig met de getrainde dataset. Tevens verbetert de prestaties van het model wanneer de data van het externe centrum naar de data van het initiële centrum werd getransformeerd (harmoniseren).

Humaan papillomavirus (HPV) is een virus dat een rol kan spelen bij het ontstaan van keelkanker. Bewezen is dat patiënten met HPV-geïnfecteerde keelkanker (HPV positief) een betere kans op genezing hebben dan patiënten met keelkanker die niet HPV-geïnfecteerd (HPV negatief) is. In de praktijk wordt de HPV status van de tumor geanalyseerd door het uitvoeren van een biopsie. In hoofdstuk 5 werden modellen gebouwd op basis van klinische variabelen en/of radiomics beeldeigenschappen om de HPV-status van de tumor middels een niet-invasieve methode te voorspellen. Radiomics beeldeigenschappen laten zien dat HPV-positieve tumoren ronder zijn en een hogere homogeniteit in de tumor textuur bevatten, wat de tumorbiologie weerspiegelt van HPV-positieve tumoren die minder invasieve exofytische groei en niet-keratiniserende histopathologie omvat vergeleken met HPV-negatieve tumoren. Hoewel een model op basis van klinische en radiomics beeldeigenschappen het meest voorspellend is, zou een model dat uitsluitend is gebaseerd op klinische variabelen de voorkeursmethode zijn vanwege het gemak van implementatie. Deze bevindingen dienen extern te worden gevalideerd aangezien de single-center-methodiek toegepast in deze studie.

Deel III: Vereenvoudiging of automatisering van tumor segmentatie technieken om de klinische acceptatie van op MR-gebaseerde radiomics voor keelkanker patiënten te verbeteren

Een cruciale factor die de klinische acceptatie van radiomics belemmert is de benodigde manuele tumor segmentaties, die arbeidsintensief, tijdrovend en gebruikersafhankelijk zijn. Om deze reden zijn alternatieve segmentaties noodzakelijk, zoals segmentaties die al beschikbaar zijn of simpeler zijn om uit te voeren. Hoofdstuk 6 onderzoekt zes manuele segmentatie strategieën in radiomics modellen voorspellend voor HPV, waaronder tumor segmentaties uitgevoerd door een niet-ervaren waarnemer, een ervaren radioloog en een radiotherapeut (dit omvat de al aanwezige tumor delineatie gemaakt voor het radiotherapie behandelplan). Daarnaast werden "eenvoudige" tumor segmentaties geëvalueerd waarbij de tumor werd gesegmenteerd op de MRI-plak met de grootste axiale diameter (2D volume) of werd de tumor gesegmenteerd door het plaatsen van bolvormige volumes in het tumorgebied (met een diameter van 4 mm of de grootst mogelijke diameter van de tumor). Bevindingen tonen aan dat minder arbeidsintensieve, gemakkelijk toepasbare segmentaties de ervaren tumorvolume segmentaties kunnen vervangen, met name de 2D-segmentaties verkregen uit één MRI-plak. Aangezien de relevante radiomics beeldeigenschappen uniek zijn voor elke uitkomstparameter, hebben alternatieve tumor segmentaties invloed op de prestatie van de voorspelling van het model en kunnen deze bevindingen afwijkend zien in andere radiomic voorspellingsmodellen.

Zo zijn eenvoudiger segmentaties niet in staat om de deskundige tumor segmentaties te vervangen wanneer dezelfde alternatieve segmentatie strategieën werden toegepast in een radiomics model voorspellend voor behandelingsuitkomst (hoofdstuk 7). Deze tegenstrijdigheid kan het gevolg zijn van beeldinterpolatie. Vorm-gebaseerde beeldeigenschappen zullen grover zijn, maar nog steeds representatief voor tumor contour, als gevolg van beeldinterpolatie. Daarintegen zullen bij beeldeigenschappen die niet op vorm gebaseerd zijn de voxel-intensiteit waarden veranderen, met als gevolg dat de waarde van de beeldeigenschap ook verandert. Een model dat de uitkomst van de behandeling voorspelt, bestaat grotendeels uit vorm-gebaseerde beeldeigenschappen, terwijl HPV status van de tumor voornamelijk wordt voorspeld door beeldeigenschappen die niet op vorm gebaseerd zijn. Dit kan de verschillende bevindingen verklaren wanneer alternatieve tumor segmentaties worden toegepast op de beide modellen (pathologie classificatie versus behandelingsuitkomst). Een tweede verklaring kan liggen in informatie over het perifere oppervlak, zoals tumorinvasie. Deze informatie is relevanter bij het onderscheiden van het behandelresultaat in vergelijking met het bepalen van de HPV-status van de tumor. Mogelijk kan er onvoldoende informatie worden verkregen uit een enkele MRI plak om het behandelresultaat te voorspellen. De verschillende conclusies voor modellen die HPV status van de tumor en behandeluitkomst voorspellen, weerspiegelen de noodzaak om de impact van elke afzonderlijke factor van de radiomics-werkwijze op de variabiliteit van beeldeigenschappen te onderzoeken voordat radiomics in een klinische setting kan worden geïmplementeerd.

Een andere oplossing voor de gebruikers-afhankelijke, tijdrovende handmatige tumor segmentaties zijn automatische tumor segmentaties. Hoofdstuk 8 beschrijft een deep learning (3D UNet) architectuur om (semi-)automatische segmentaties van de primaire tumor van keelkanker patiënten te verkrijgen. Een architectuur ontworpen op meerdere MR-sequenties (T1w, T2w, T1w+contrast) vertoont de grootste spatiele overlap tussen de automatische tumor segmentatie en een tumor segmentatie verkregen van een humane waarnemer. Dit geeft aan dat elke MR-sequentie exclusieve kenmerken bezit met betrekking tot voxel-intensiteit om tumorweefsel te identificeren. Het handmatig plaatsen van een kubus die de tumor omvat (semi-automatische segmentatie) verbetert deze overeenstemming en vermindert de werklast van de clinicus aanzienlijk (30 min-2 uur versus 5 minuten), waardoor de haalbaarheid van zinvolle klinische kwantitatieve analyse (zoals radiomics) wordt vergroot.

IMPACT PARAGRAPH

RELEVANCE

Imagine that you are a medical doctor. Last year you have treated two patients with comparable presentation and identical complaints, namely pain in the throat, difficulties with chewing and a mass in the neck. After clinical and diagnostic assessment, both patients were diagnosed with locally advanced oropharyngeal squamous cell carcinoma (OPSCC) and followed a seven-week during trajectory of radiation therapy combined with chemotherapy. Now, a year later, one patient showed complete response, whereas, unfortunately, the other patient died five months after treatment from a local recurrence. Two comparable patients, two identical cases, although, the outcome of the treatment are extremely different. If we knew this on beforehand, we could identify which patients would benefit from intensification of treatment and for whom the treatment would do more harm than good. So actually, then we could anticipate by advising on more personalized medicine?

The selected treatment of OPSCC patients currently depends on patients factors as well as clinical and diagnostic evaluation of tumor characteristics such as TNM stage. Diagnostic evaluation with imaging includes the extraction of semantic features (i.e. shape, size, extent and metastases) and functional parameters (tumor diffusion/perfusion) from medical images (i.e. CT and/or MRI) by the radiologist. However, much more information might be hidden into medical images, which are not revealed by visual inspection. Radiomics is a quantitative method that enables extraction of a large number of “hidden” features, calculated from mathematical formulas (“data-characterization algorithms”). A unique combination of radiomic features (radiomic model) can be correlated to various clinical outcomes and might well play a role in personalized medicine.

This thesis shows the potential of magnetic resonance imaging (MRI)-based radiomics to increase reliably prediction if an OPSCC patient will have successful treatment and to classify if the tumor is infected with human papillomavirus (HPV). Despite its potential, the current format of radiomics still include some challenges hampering clinical implementation. Crucial steps to improve standardization, reproducibility, repeatability and generalizability are needed. The most important step is external validation of single-center radiomic models. Findings of this thesis proved already that our radiomic model, predictive of treatment outcome, could be applied on external data while maintaining good performance. Another challenge is to obtain a more standardized and user-independent radiomic workflow. This thesis investigates the possibilities of simplification and automation to substitute the manual observer-dependent time-consuming tumor delineations. We concluded that each individual radiomic model reacts differently on alternative tumor delineations. Automatic tumor delineations are feasible, although, manual adjustments are still required to optimize them.

When the challenges are overcome and radiomics shows to be reliable, radiomics can act as clinical decision support tool. In an ideal situation, radiomic models are implemented as software extension in current healthcare information systems. This software gives information complementary to clinicians' findings by combining parameters of clinical examinations, histopathology, genetic data and diagnostic imaging. The software visualizes the likelihood ratio for cancer response in the automatically delineated tumor region, where a cut-off value can be used to determine the best appropriate treatment for the patient.

TARGET POPULATION

The results of this thesis are relevant to several groups. Firstly, the scientific community investigating radiomics or other (quantitative) biomarkers in any cancer type or disease might obtain additional knowledge and new insights for future research. Especially a remark must be made for the finding that a certain approach may not be suitable for all radiomic models, suggesting that investigation of every unique model is necessary.

Secondly, healthcare professionals will benefit from radiomic models. The major profit will be for radiologists who are daily utilizing medical imaging to detect, diagnose and evaluate cancer progression. Radiomic models might assist in each of these steps, although, and also enable quantification of tumor characteristics or prediction of treatment outcome. The enhanced radiological information provided for the radiologist help further personalize treatment and optimize treatment outcomes. In addition, valuable time will be saved by assisting or eliminating time-consuming tasks. Extending the radiomic model with parameters originating from other disciplines will also aid other health professionals, such as the pathologist. Overall, information from radiomic models can be discussed during multidisciplinary meetings. While the radiomic model might well support clinicians, it will not replace them.

OPSCC patients are an important group who will benefit from radiomic models. The routinely non-invasive acquired patient images required for radiomics do not harm or ask for additional proceedings, while it might aid in the selection of the most appropriate treatment. This selected personalized treatment might well prevent treatment failure and unnecessary side effects as much as possible. Decreasing these factors, along with better expectation management, improves the spirit of the patient and quality of life.

Radiomics is able to support clinicians, reduce workload for radiologists, improve risk evaluations, improve patient management, prevent treatment failure and limit unnecessary side effects. All these factors might well contribute to make current healthcare more efficient and cost-effective. Therefore, hospitals and healthcare generally benefits from the implementation of radiomic models, which might also have an impact on the regular citizen (e.g. health insurance).

More internationally, the findings of this thesis might be useful for poor countries as well. We prove that radiomic models are able to predict HPV status of the tumor, without the need of invasive biopsy. This ability of radiomics will not replace the invasive biopsy in wealthy countries, however, it can substitute the costly polymerase chain reaction (PRC) immunohistopathology analysis when medical imaging is performed is poor countries. Additionally, retrospective analysis of the HPV status of the tumor is feasible when biopsy is not performed.

ACTIVITIES

The results of this thesis have been presented at multiple (inter)national conferences and published in peer-reviewed international journals. Lessons learned from this thesis can be applied in future research to optimize radiomics. Follow-up research projects in our institution are now investigating the potential of integrating genomic data in a radiomic model to predict treatment outcome in head and neck cancer patients.

This thesis shows that radiomics for OPSCC patients has potential, but the current format strongly requires optimization to make it applicable as clinical decision-support tool. Importantly, collaboration of research teams, centres and countries is recommended to work together towards a trustable, reliable and representative tool that is able to support clinicians by providing complementary information.

LIST OF PUBLICATIONS

Bos P, van den Brekel MWM, Taghavi M, Gouw ZAR, Al-Mamgani A, Waktola S, Aerts HJWL, Beets-Tan RGH, Castelijns JA, Jasperse B. Largest diameter delineations can substitute 3D tumor volume delineations for radiomics prediction of human papillomavirus status on MRI's of oropharyngeal cancer. *Physica Medica.* Accepted July 2022.

Bos P, van den Brekel MWM, Taghavi M, Gouw ZAR, Al-Mamgani A, Waktola S, Aerts HJWL, Beets-Tan RGH, Castelijns JA, Jasperse B. Simple delineations cannot substitute full 3D delineations for MR-based radiomics prediction of locoregional control in oropharyngeal cancer. *Eur J Radiol*. 2022;148:110167. doi:10.1016/j.ejrad.2022.110167

Bos P, van der Hulst HJ, van den Brekel MWM, Schats W, Jasperse B, Beets-Tan RGH, Castelijns JA. Prognostic functional MR imaging parameters in head and neck squamous cell carcinoma: A systematic review. *Eur J Radiol*. 2021;144:109952. doi:10.1016/j.ejrad.2021.109952

Rodriguez Outeiral R, **Bos P**, Al-Mamgani A, Jasperse B, Simoes R, van der Heide UA. Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning. *Phys Imaging Radiat Oncol*. 2021;19:39-44. doi:10.1016/j.phro.2021.06.005

Bos P, van den Brekel MWM, Gouw ZAR, Al-Mamgani A, Taghavi M, Waktola S, Aerts HJWL, Castelijns JA, Beets-Tan RGH, Jasperse B. Improved outcome prediction of oropharyngeal cancer by combining clinical and MRI features in machine learning models. *Eur J Radiol*. 2021;139:109701. Doi:10.1016/j.ejrad.2021.109701

Bos P, van den Brekel MWM, Gouw ZAR, Al-Mamgani A, Waktola S, Aerts HJWL, Beets-Tan RGH, Castelijns JA, Jasperse B. Clinical variables and magnetic resonance imaging-based radiomics predict human papillomavirus status of oropharyngeal cancer. *Head Neck*. 2021;43:485-495. doi:10.1002/hed.26505

Min LA, Vacher YJL, Dewit L, Donker M, Sofia C, van Triest B, **Bos P**, van Griethuysen JJW, Maas M, Beets-Tan RGH, Lambregts DMJ. Gross tumour volume delineation in anal cancer on T2-weighted and diffusion-weighted MRI – Reproducibility between radiologists and radiation oncologists and impact of reader experience level and DWI image quality. *Radiother Oncol*. 2020;150:81-88. doi:10.1016/j.

radonc.2020.06.012

Van Leeuwen KG, **Bos P**, Trebeschi S, van Alphen MJA, Voskuilen L, Smeele LE, van der Heijden F, van Son RJJH. CNN-based pheneme classifier from vocal tract MRI learns embedding consistent with articulatory topology. *Proc Interspeech* 2019;909-913. doi:10.21437/Interspeech.2019-1173

Van ‘t Hullenaar CDP, **Bos P**, Broeders IAMJ. Ergonomic assessment of the first assistant during robot-assisted surgery. *J Robot Surg*. 2019;13:283-288. doi:10.1007/s11701-018-0851-0

Bos P, Martens RM, de Graaf P, Jasperse B, van Griethuysen JJM, Boellaard R, Leemans CR, Beets-Tan RGH, van de Wiel MA, van den Brekel MWM, Castelijns JA. External validation of an MR-based radiomic model predictive of locoregional control in oropharyngeal cancer. *In submission*

Rodriguez Outeiral R, **Bos P**, van derHulst HJ, Al-Mamgani A, Jasperse B, Simoes R, van der Heide UA. Strategies for tackling the class imbalance problem of oropharyngeal primary tumor segmentation on Magnetic Resonance Images. *In submission*

DANKWOORD

Een promotietraject die twee afdelingen combineert: de radiologie en hoofd-hals oncologie. Een ontzettend toffe combinatie, waar ik onwijs veel plezier en energie uit heb gehaald. De afgelopen jaren hebben mij doen beseffen dat een promotietraject veel meer is dan onderzoek doen. Naast het ontwikkelen van diverse skills benodigd voor gedegen onderzoek (zoals statistiek, programmeren, klinische achtergrond en samenwerken) komen lessen op persoonlijk vlak ook om de hoek kijken. De persoonlijke groei die ik heb doorgemaakt op al deze vlakken was dan ook niet mogelijk geweest zonder de feedback, hulp en steun van vele lieve mensen om mij heen. Gezegend en dankbaar wil ik daarom eenieder in het zonnetje zeten en hun te bedanken voor hun waardevolle bijdrage. Zonder hun hulp was dit proefschrift er niet geweest.

Allereerst wil ik mijn promotieteam, bestaande uit **prof. dr. Regina Beets-Tan, prof. dr. Michiel van den Brekel, dr. Bas Jasperse** en **prof. dr. Jonas Castelijns**, bedanken voor de kans die zij mij hebben geboden om in het Antoni van Leeuwenhoek ziekenhuis een PhD traject te vervaardigen. Uit ervaring kan ik zeggen dat het een zeer prettig ziekenhuis is om in te werken, waar iedereen onderzoek ambieert en mogelijk maakt.

Beste **Regina**, “We choose to go to the moon in this decade and do the other things, not because they are easy, but because they are hard, because that goal will serve to organize and measure the best of our energies and skills” (J.F. Kennedy). Dit citaat zat in je speech tijdens een diner met alle onderzoekers van het Tuinhuis, met als les dat je moet leren van de tegenslagen tijdens een promotietraject. Mijn PhD ging niet zonder slag of stoot, maar jij wist mij altijd vertrouwen te geven en te kijken naar de mogelijkheden in plaats van beperkingen. Met je onbevangen gedrevenheid en hart voor onderzoek, mag je terecht trots zijn op je werk en team. Als ambitieuze, innovatieve radioloog ben jij altijd bezig met de toekomst van de radiologie. Zet deze inspiratiebron voort. Bedankt voor het delen van je ambitie en je feedback tijdens mijn onderzoek.

Eigenlijk moet ik deze alinea zo kort mogelijk houden, om zo min mogelijk tijd van je volle schema te vragen. Maar dan doe ik tekort aan de steun die ik van je heb ontvangen. Beste **Michiel**, altijd in razend tempo verplaatste (of vloog kun je wel zeggen!) jij van de ene naar de andere afspraak. Maar eenmaal op locatie, was je de rust zelfde en nam je juist alle tijd. Zo ook tijdens onze besprekingen, waar jij altijd met een positieve, maar realistische, blik keek naar mijn onderzoek en voortgang.

Vooral in dat laatste was het meermaals cruciaal dat jij knopen doorhakte, zodat ik niet te veel zij-projecten op mijn hals haalde. Door jouw enorme kennis en ervaring knoopte je diverse onderzoeken aan elkaar en zorgde je voor handige nieuwe connecties en samenwerkingen. Als bescheiden, maar sociale hoofd-hals chirurg bleef je ook tijdens de pandemie betrokken bij je PhD studenten. Het gegeven kunstdoek hangt steevast in mijn werkkamer, waar ik regelmatig naar kijk als ik even moet ontspannen en de rust moet herpakken (‘Breathe in, breathe out, repeat’). Goede dingen hebben immers tijd nodig. Ontzettend bedankt voor je vertrouwen en de structuur die je mijn onderzoek gaf.

Lieve Bos’ Boss **Bas**, ik weet niet of je het zelf weet, maar zo werd je bij ons op kantoor ook wel genoemd. Puur omdat het lekker in de mond lag, want van hiërarchie was (gelukkig!) geen sprake. Een echte knuffelbeer en altijd in voor een gezellig praatje. Praten over het weekend of andere sociale dingen was vaak iets te gezellig, waardoor de tijd voor het bespreken van het onderzoek werd beperkt. Naast sociaal, zijn we ook allebei eigenwijs, wat niet altijd hielp in de samenwerking. Met name in het begin moest ik hieraan wennen, maar uiteindelijk zag ik in dat er een gegronde redenatie aan ten grondslag lag wanneer de plannen en ideeën (weer eens) compleet werden omgegooid. Je radiologische en klinische kennis, je kritische blik en helikopterview hebben mij zowel op wetenschappelijk, maar zeker ook op persoonlijk vlak, laten ontwikkelen. Manuscripten kwamen volledig rood terug, waar complete alinea’s waren geherformuleerd. Maar jouw snoeiwerk van het manuscript zorgde er juist voor dat het artikel meer ging bloeien. Inmiddels heb ik je schrijfstijl overgenomen, waarvoor ik je zeer dankbaar ben. Bas, bedankt voor je gezelligheid, lessen, geduld en je inzet. Want pfoe, wat hebben we een hoop tumoren samen ingetekend!

Beste **Jonas**, in de laatste fase van mijn promotietraject werd jij toegevoegd aan mijn promotiecommissie. En daar kwam ik gelijk in aanraking met een aanpak die mij nog onbekend was. Je zat continue achter mijn broek aan, met als resultaat dat mijn planning beter werd nagestreefd. Toen mijn contract werd verlengd met 9 maanden, wist jij mij te vertellen dat er heel veel baby’s zijn geboren in 9 maanden, een beeldspraak voor de vele publicaties die zouden komen in deze 9 maanden. Misschien heeft het iets langer geduurd dan we gewild hadden, maar dat lag niet aan jouw snelle reacties. Wanneer ik een mail stuurde met een deadline, belde jij mij gelijk op dat je er morgen geen tijd voor had, maar wel overmorgen. En dat terwijl de deadline pas over 2 weken was. Vol trots praat je over je promovendi en hoe leuk het is als we elkaar zouden ontmoeten. Het etentje bij jou en Hafina thuis met al jouw promovendi en hun begeleiders was dan ook zeer geslaagd. Bedankt voor je begeleiding en jullie gastvrijheid.

Daarnaast wil ik de leden van de beoordelingscommissie bedanken: **prof. dr. Bernd Kremer, prof. dr. Remco de Bree, prof. dr. ir. Andre L.A.J. Dekker, dr. Frank J.P. Hoebers** en **dr. Stefan Steens**. Bedankt dat jullie waardevolle tijd wilden vrijmaken voor het doorlezen en beoordelen van dit proefschrift.

Als onafhankelijk deskundigen hebben de leden van mijn OOA-commissie mijn promotietraject altijd gevolgd. **Prof. dr. Marcel Stokkel, dr. Neeltje Steeghs** en **dr. Abrahim Al-Mamgani,** bedankt voor jullie betrokkenheid.

Elk artikel is echt een team effort, wat zonder de betrokkenheid van **de vele co-auteurs** niet tot stand had kunnen komen. Jullie onmisbare waardevolle bijdrage in de vorm van kennis, ervaring en tijd zorgde ervoor dat de artikelen werden voorzien van klinische, technische of statistische onderbouwingen. Bedankt voor de goede samenwerking. **Winnie**, jouw enthousiasme is aanstekelijk. Bedankt voor het meedenken en uitvoeren van de systematische search. Zonder jou was ik nog verstrikt in MeSH termen. **Hedda,** de tweede reviewer voor de systematische review. Wat een hoop abstracts en artikelen hebben wij gescand, gelezen en beoordeeld. Bedankt voor je hulp! **Hugo**, bedankt voor het delen van je expertise op het gebied van radiomics. Een waar genoegen om met jou samen te werken. **Zeno** en **Abrahim**, ik ben dankbaar dat ik gebruik mocht maken van een cohort die jullie hebben samengesteld. Er zijn al vele publicaties verschenen met dit cohort, die alleen mogelijk zijn gemaakt door jullie harde werk om alle relevante klinische gegevens te verzamelen. **Selam**, a real helpline when a difficult technical issue came up. Your innovative ideas and discipline to keep searching to obtain more clarity about the data were of great value. **Marjaneh and Joost,** no programming error was too much for you. You were always willing to help me out. **Roland, Pim, Ronald, René** en **Mark,** wat kijk ik trots terug op een geweldige multicenter samenwerking. Eentje die zeer vlot is verlopen door de adequate samenwerking en vlotte feedback. **Roland**, binnen één week zullen we beiden promoveren. Wat zal Jonas trots zijn. **Mark,** hoewel je tijd gering was wist je toch wat tijd vrij te maken om mij wat handvaten te geven die mij hielpen om de data beter te bekijken, te analyseren en te begrijpen. Bedankt voor je statistische expertise. **Roque,** the PhD life is not always easy. We struggled, we fell and we had to motivate ourselves to find the discipline to convert the feeling of failure into success and stand up again. Therefore, the acceptance of a manuscript is costly, which has to give you enough discipline and motivation to go for the next one. I really thank you for our grateful collaboration, but also for being able to share our feelings, coffee moments and of course the nice dancing during the OOA retreat. **Rita** en **Uulke**, sparsessies met jullie zorgde ervoor dat er weer nieuwe invalshoeken aan het licht kwamen, die nieuwe inzichten en ideeën meebrachten. A special thanks to **Richard Golding** to correct and improve the manuscripts on English grammar.

Lieve **Jorrita**, Beste **Marion**, Beste **Evelyn**, jullie waren er altijd om mij uit de brand te helpen met praktische zaken. Bedankt voor het inplannen van afspraken en geven van informatie die mij nog niet bekend was. **Jorrita**, je positiviteit, vrolijkheid en oprechte interesse maakten je een waardevolle gesprekspartner die mijn tijd in het Tuinhuis zeker kleur (Jorrita droeg altijd vrolijke gekleurde kleding) hebben gegeven! Beste **Minke De Haan**, halverwege mijn PhD bleken 'to do' boekjes een grote uitkomst om mijn dagelijkse planning strakker te laten verlopen. Bedankt voor het bestellen van deze boekjes en indirect bijdragen aan mijn planning. Beste **Carine Sondermeijer**, bedankt voor het regelen van al de benodigde papieren voor de multicenter studie. Ik ben dankbaar dat ik deze hoeveelheid papierwerk met vertrouwen uit handen kon geven.

Lieve **kamergenootjes** van 'de Geekroom'. Eén van de kamers in het Tuinhuis waar je kunt voelen wanneer er een tram passeert. Of was het toch de wiebelende voet van Niels? Ik hoop dat mijn valse zangkunsten en groene medekamergenoten niet te veel voor afleiding hebben gezorgd. Dear **Stefano**, we shared the office before you left it for 'room 10'. As Italian you tried to teach me the meanings behind these hand gestures, which were quite a few and therefore hard to remember. Sorry, I am still not able to do them. But I had a lot of fun with it during the ISMRM conference in Paris. Maybe also due to the nice view from the roof terrace, the good drinks, delicious food, our dance moves and, especially, the closing party in Museé des arts forains with the carousel. Thank you for the international influences and good time together. Lieve **Joost**, als wandelende encyclopedie was je een grote informatiebron voor iedereen uit het Tuinhuis. Er ging geen dag voorbij zonder dat er iemand aan jouw bureau zat die jij met alle liefde hielp met een probleem. Zo heb je mij ook diverse keren geholpen met het oplossen van een error uit mijn programmeercode of door mij je code te laten begrijpen door deze regel voor regel uit te leggen. Als programmerende dokter weet ik zeker dat je veel zal bijdragen aan medische innovaties in de toekomst. Dear **Marjaneh**, as the only two girls in the 'Geekroom', we were strong together. We showed this girl power by sticking the poster "We can do it" on the door and give the room some pink touches when the guys were off to New York. I think these touches are still there, keeping our spirit alive. Lieve **Niels**, een hardwerkende positivo die altijd rustig bleef, zelfs onder hoge druk. Maar ook een echte babbelaar, waarbij het keer op keer genieten was als je vol trots zat te vertellen over je oma, je dates (in het begin van je PhD) of je crossfit workout. Daarnaast was je ook een luisterend oor en baken van advies wanneer je zag dat ik het even kon gebruiken. Met je onuitputtelijke enthousiasme en energie liet je graag (onzinnige!) memes of filmpjes zien, PPAP (Pen-Pineapple-Apple-Pen) zal mij altijd bijblijven! Lieve **Kay**, van stille muis ontpopte je in een echte prater. Je perfectionisme en kritische blik zorgden ervoor dat je veel wist over de technische kant van de MRI, resulterend in adviezen en ideeën voor mijn onderzoek waar ik zelf nog niet aan gedacht had. Leer te geloven in je eigen kunnen, want jij bent de specialist op jouw vakgebied. Sorry voor de windows-update grap, maar vergeet deze vooral niet bij nieuwe collega's uit te proberen. Beste **Najim,** ik dacht dat ik altijd vroeg op kantoor was, maar jij spande echt de kroon als vroege vogel. Hierdoor stond de deur van ons kantoor al als een warm welkom open wanneer ik aan een nieuwe werkdag begon. Door COVID hebben we maar een korte tijd samen doorgebracht, maar die tijd heb ik zeker als prettig ervaren. Geeks, allen bedankt voor de gezellige mooie tijd samen!

Natuurlijk gaat onderzoek gepaard met een hele hoop leuke en mooie herinneringen. Daarvoor wil ik **alle collega's van het Tuinhuis** bedanken. Zowel binnen de muren van het Tuinhuis met de Sinterklaas viering, het oplossen van de AIVD kerstpuzzel, potluck diners, gezellige lunchmomenten, spelletjes avonden en de vele koffie momenten (soms met taart, want publiceren = trakteren!). Maar ook buiten het Tuinhuis wisten we een hoop plezier te maken tijdens de bezoekjes aan de markt, de escape room, NKI Summer party's, PhD diners, medewerkersfeest, congressen en de OOA retreat in Renesse. Een aantal wil ik in het bijzonder bedanken. **Lisa,** het grote brein in het oplossen van de AIVD puzzel 2016! Met veel plezier hebben we aan deze puzzel gewerkt, waarbij elke ochtend begon met nieuwe inzichten die we de avond ervoor thuis hadden gevonden. Ik schreef met alle plezier code voor je, zodat handmatig uitzoekwerk je bespaard bleef. Bedankt voor de samenwerking en alle gezelligheid! **Femke,** vrolijke energiebom, wat heb ik van jou genoten! Op feesten konden wij helemaal losgaan (vooral in de silent disco). Van tevoren spraken we dan af dat we de laatste trein gingen pakken, om elkaar op het feest aan te kijken en te weten dat we nog niet naar huis gingen. Naast de hoeveelheid plezier, gebabbel, danspassen en hoop borrels die we samen hebben beleefd, ben ik je ook altijd dankbaar voor jouw ongezouten mening op de layout van mijn figuren en presentaties. Ik heb er een echte vriendin bij! **Hedda**, we leerden elkaar kennen toen je als student bij radiotherapie mijn hulp vroeg, niet wetend dat we een jaar later collega's zouden worden in het Tuinhuis. Inmiddels hebben we in onze samenwerking veel gedachten uitgewisseld, waarbij we elkaar onbewust motiveerden. Dit resulteerde in een systematic review die zeer spoedig gepubliceerd is. Bedankt voor je gezelligheid en prettige samenwerking. Een speciaal bedankje ook naar **Daphne**, **Judith OH** en **Judith** voor de gezellige avonden bij elkaar thuis, voorzien van een heerlijke maaltijd (niet roeren in de pan!), onophoudelijk geklets en fanatisme tijdens spelletjes. En natuurlijk niet te vergeten, de vele (gekke) danspasjes en momenten waarin we de slappe lach hadden tijdens de OOA retreat, NKI summer party en het medewerkersfeest. Wanneer ik even mijn ei kwijt moest, kon ik altijd bij jullie terecht voor goede adviezen om mijn focus weer te vinden. Een PhD is immers een sinus van motivatie en effectiviteit (aldus Judith OH), wat door deze gesprekken met julie weer helder werd, resulterend in een teruggevonden motivatie. Bedankt voor de fijne momenten en natuurlijk ook de Disney piano afspeellijsten.

Ook wil ik mijn dank betuigen aan **alle collega’s van hoofd-hals**. De inspirerende OIO onderwijs momenten tijdens de lunch op vrijdag, de informatieve meetings op dinsdagochtend en niet te vergeten de vele koffie (oké thee) momenten. **Luuk**, **Kilian**, **Bence**, **Maarten**, **Rob**, **Kicky, Martijn, Klaske** en **Maartje**, bedankt voor de gezellige momenten en natuurlijk het onderkomen tijdens de brandmelding in het Tuinhuis, wat stiekem direct een mooi moment was om even bij te kletsen;). De variatie in onderzoeksvelden maakte het iedere keer weer een leerzaam feest om met jullie te sparren en samen tot mooie onderzoeksvraagstukken/oplossingen te komen. Ik droeg met liefde dan ook bij aan deze onderzoeken, ook al moesten er daarvoor stikkers op mijn tong geplakt worden of moest ik in een MRI liggen met een nieuwe MRI spoel.

Ook dank aan **alle collega’s van de radiologie, radiotherapie, nucleaire geneeskunde, genetica en pathologie** met wie ik overleggen heb gevoerd en heb samengewerkt in projecten die niet in dit proefschrift beschreven zijn. In het bijzonder **Petra, Arjan, Cees, Leon, Laura, Conchita** en **Wouter.**

.نم یرتکد هرود هارمه و تسود ،ابیز و زیزع هناجرم
رتفد دراو نیشنلد دنخبل و دنلب ریخب حبص کی اب هشیمه ؟مکن عورش اجک زا
مفادها زا یکی ؛مهاتوک یگداوناخ مان یفالت یارب .یتخاسیم ور مزور و یدشیم راک
قفوم هک مگب تسرد)هدازیوضر یوقت(ور وت یگداوناخ مان منوتب هک دوب نیا
.دوب نم یارب یسراف یریگدای عورش نیا .مدش
یاه هتسپ نم هک یدروایم هدش رکف و ابیز یایاده ،ناریا زا تشگرب زا دعب
یارب ناوارف هزیگنا نتشاد رانک رد .متشاد تسود یلیخ هژیو هب ور هزمشوخ
یبوخ زور یتقو .یدوب بوخ یاه هملاکم و یخوش هدامآ هراومه ،ینف لیاسم مهف
هب ور نم .دوب نم کمک هشیمه تاه لغب و اونش شوگ ،تمرگ تیصخش ،متشادن
هظحل مه راک رتفد زا جراخ .شخبب راک رتفد یوت ای هملد لفلف ندز زاگ رطاخ
ماش ؛میدرکیم راک هناخ رد مه اب انورک هرود رد یتقو .میتشاد یبوخ یلیخ یاه
هب تریذپ فاطعنا یاه تسد و اهاپ و رمک اب .میتفر یم ینامهم هب ای میدروخیم
یقیسوم یتقو هژیو هب) یدرکیم توهبم ور همه و یتفر یم یقیسوم لابقتسا
ناتسود یرتکد هرود رد ندوب هارمه رطاخ هب هک نونمم ،زیزع هناجرم .(دوب اریکش زا
.سوب سوب .هدنیآ رد بوخ یلیخ تاظحل دیما هب .میدش هم یارب یبوخ

مسا نیا ایوگ یلو درکیم باطخ ناج ور وت هناجرم ارچ مدشیمن هجوتم ؟ناج ای بارهس
هب لگ نداد هیده اب نومرادید نیلوا رد .دش لح هلاسم .دوب ت یعقاو مسا یاج هب
ور هظحل نوا تقو چیه .یداد رارق ریثات تحت ور قاتا همه نیاتنلو زور رد هناجرم
هناجرم .نوت هناصلاخ یزاون نامهم و هقالع ،اه دنخبل زا منونمم .منک یمن شومارف
.سوب سوب .نوتاهاب ییانشآ زا متخبشوخ بارهس و

Translation: Dear and pretty **Marjaneh**, my PhD buddy and paranymph! Where should I start? You always entered the room with a loud “Good Morning” and big smile on your face. It made my day! To compensate for my own short last name, I set a goal to be able to pronounce your full last name (Taghavirazavizadeh), which I succeeded. It was the beginning of learning Farsi. After a visit to Iran, you came back with thoughtful gifts, where I especially appreciated the delicious pistachios. Besides your tireless motivation and drive to really understand the technical problem, you were always up for jokes or good conversations. Your warm personality, listening ear and the hugs were always encouraging when I didn’t have my day. Sorry for the noise when I ate another bell pepper at the office. Besides the office, we shared even more fun moments, where we worked together at home, dined, ran or went to a party. There, of course, you always stole the show with your flexible hips, feet and hands that moved gracefully to the music (especially when Shakira was played). Lovely Marjaneh, thank you for the great time sharing our PhDs, a true friendship is born. I look forward to good moments in the future. Boos Boos (kiss in Farsi)! **Sohrab**, or John? I didn’t understand why Marjaneh always picked up the phone with “John”, but apparently it is an in place of the name. Problem solved! We first met when you brought flowers to Marjaneh on Valentine day, you impressed the whole room and I will never forget that moment. Thank you for your sincere smile, interest and hospitality. Marjaneh and Sohrab, I am blessed to have met you. Boos boos!

Lieve **Maud**, toen ik je vroeg of je mijn paranimf wilde zijn, zei je volmondig ja. Zelfs zonder dat je wist wat die rol precies inhoudt. Op het moment van schrijven heb je het idee dat paranimfen tijdens de verdediging trots met het proefschrift in de hand achter de promovendi staan te stralen. En die taak is jou op het lijf geschreven. Maar… er komt meer bij kijken (sorry voor het nog niet eerder vertellen). Ik hoop dat je nog steeds mijn paranimf wilt zijn, want als enthousiaste, leergierige, sociale en vooral vrolijke meid weet ik zeker dat die andere taken jou met veel gemak af zullen gaan. Ik zal dan ervoor zorgen dat de vragen van de oppositie beantwoord worden. Ik zal namelijk nooit vergeten dat jij mijn promotieonderzoek eens vergeleek met een zak M&M’s gevuld met verschillende smaken. Van de buitenkant zien de M&M’s er nagenoeg hetzelfde uit, maar door juist de kleine verschillen te analyseren (die we niet met het menselijk oog zien), kun je voorspellen welke inhoud de M&M heeft (pinda, chocolade of crispy). Een beknopte versie van mijn onderzoek in een chocoladelaagje. Gelukkig maak je deze vergelijkingen niet in je eigen onderzoeken tijdens je (master)studie, waar je goed gebruik wist te maken van mijn kennis en ervaring om een artikel op te zoeken, te vertalen, statistiek te bekijken of je verslagen/posters door te lezen.

Toppers! Ofwel **al mijn lieve vriend(inn)en.** Ook jullie zijn bijna van mijn promotieverhalen af =P**.** De afgelopen jaren hebben jullie vaak interesse getoond in mijn promotieonderzoek, waarin ik zowel de hoogte als dieptepunten met jullie kon delen. Werk is leuk, maar momenten met waardevolle vriendschappen natuurlijk altijd leuker! Bedankt daarom voor de welkome afleiding en ontzettend fijne momenten die we met elkaar gedeeld hebben. Ik waardeer jullie allemaal enorm. Lieve **Anneke, Carmita, Elyse, Joyce** en **Nienke**, onze jarenlange vriendschap door dik en dun is gewoon goud! Wat hebben wij al vele mooie momenten beleefd, en de maat(beker) is nog lang niet vol! Samen gezellig lunchen, weekenden weg, escape rooms of spelletjes spelen, feesten en natuurlijk vele drankjes op het terras. Bedankt lieve meiden voor de vele keren waarop de tranen in onze ogen stonden van de slappe lach, al het gebabbel en ontelbare danspasjes! Lieve **Anne, Anouk, Denise, Kim** en **Suzanne,** mijn relax (maar ook thee, eet & feest) maatjes. Soms waren we iets te relaxt, vooral tijdens een weekend op de Finca la Pajera (Maella, Spanje), een plek in de middle of nowhere zonder enige prikkel van de buitenwereld. Op deze primitieve plek vonden we rust, waren we één met de natuur en mochten we ook nog legaal brand stichten. Het besef van tijd verdween volledig tijdens dit weekend. Toen aan het einde van het weekend weer strikte tijden om de hoek kwamen kijken, vluchtten we dan ook in de Mac Donalds in plaats van het vliegtuig. Mijn duurste hamburger ooit. Meiden, bedankt voor alle PhD detox (k) uren! Lieve **Feike** en **Vincent**, met het stellen van kritische vragen lieten jullie mij nadenken over de problemen waar ik tegenaan liep. Deze sparsessies en goede adviezen waren een inspiratiebron voor nieuwe inzichten. Dat dit gepaard ging met heerlijke maaltijden of versgebakken brood of baksels was zeker geen straf, keer op keer was het weer een genot om nieuwe recepten te mogen proeven. Feike, jij bent mij al voorgegaan en hebt het goede voorbeeld al gegeven. Nu ik nog! Bedankt voor jullie deur die altijd open staat. Lieve **Ronald,** wat hebben wij samen veel gedeeld en meegemaakt. De hoogtepunten werden natuurlijk groots samen gevierd, maar tijdens dieptepunten bood jij mij de schouder aan die ik nodig had. Je luisterend oor, je vertrouwen in mijn kunnen, de vele carpoolritten waar creatieve ideeën ontstonden, het vergroten van mijn sociale netwerk en niet te vergeten de bak aan plezier die we beleefd hebben zorgden allemaal voor waardevolle momenten die dienden als goede afleiding voor mijn PhD. Bedankt voor het zijn van een goede vriend die altijd voor een ander klaarstaat. Lieve **Renee**, het leek altijd wel of de tijd twee keer zo snel ging als wij aan het ouwehoeren waren. De vele adviezen, maar ook gesprekken over koetjes en kalfjes waren een heuse afleiding en bron van inspiratie. Ik heb waardering voor hoe je altijd alles weet te combineren, onvermoeibaar ga jij altijd door. Ik ben dan ook met alle liefde een animatieteam als jullie kinderen hierom vragen, door mij in hun te verplaatsen wordt mijn creatieve brein immers geprikkeld voor nieuwe ideeën en invalshoeken. Lieve **Laurien** en **Petra**, dat wandelingen niet saai zijn bewezen we elke keer maar weer. Aangezien we vaak aan het eind van het rondje nog niet uitgepraat waren, werd het wandelrondje verlengd. Maar zelfs dat mocht niet altijd baten, waardoor we uiteindelijk in de kroeg belandden. Gelukkig waren kroegen ons niet vreemd, vooral tijdens de carnavalsperiode wisten we met enige regelmaat hier een koude versnapering te nuttigen waarbij de voeten toch wel zeker een beetje van de vloer gingen. Zeg ik nou een beetje? Volledig van de vloer bedoel ik natuurlijk! Vol energie, dwars door de hele zaal, van muur 1 naar muur 3 en weer terug, ZOIGE! Bedankt voor deze hilarische momenten van afleiding.

Lieve **vrienden van de muziek,** altijd een vrolijke noot met jullie! Geen muziek? Geen probleem, onbewust begint iemand te trommelen op attributen resulterend in een heel klankspel. Muziek is, mede dankzij de gezelligheid, een belangrijk onderdeel in mijn leven. Van samen musiceren (en knipogen) tot de borrels achteraf. En van (straat)optredens tot onwijs veel plezier met elkaar beleven buiten de muziekvereniging. Zo waren de sportieve weekenden weg een ultieme compensatie voor de wekelijkse zit achter mijn computer. Lieve **Bram** en **Marco**, elke mijlpaal was een reden om een champagnefles te poppen! De kurken vlogen dan ook vaak op donderdagavond tijdens onze eetclub avond door de lucht. Hoewel ik soms optimistisch was en dacht nog iets af te kunnen maken tijdens deze avonden, bleek dit keer op keer een desillusie (voor mijn werk). De gezelligheid overheerste en overwon het productieve. Ook naast deze eetdates waren jullie een fijn gezelschap. Als klankbord kon ik altijd mijn verhaal bij jullie kwijt en leerden jullie mij dat ik moest leren vertrouwen in mijn kunnen, een PhD is namelijk “in veel landen de hoogste academische graad” (aldus Wikipedia, goede bron ook!). Ik kijk uit naar nog mooie en hilarische momenten samen. Dus kom maar op met die wintersport, weekendjes weg, hardlopensessies, feestjes en andere activiteiten! **Tycho** (en de Brains), de ware pubquiz master, winnen was immers toch écht belangrijker dan meedoen. Met jouw bewonderingswaardig hoge niveau van algemene kennis, werd ook mijn kennis buiten het promotieonderwerp uitgebreid. Altijd handig en leerzaam! Maar ook de gezelligheid die gepaard ging met deze avonden zorgden voor hilarische momenten. **Tim**, designen kan ik wel aan jou overlaten. Met je scherpe blik zag je direct als de uitlijning niet klopte. Hoewel ik het design al uitgewerkt had, maakte jij het af door het aanbrengen van details.

Speciaal voor jou heb ik dan ook extra gele tinten in de cover verwerkt, je bent er immers dol op! Stiekem wil ik ook een klein bedankje brengen naar **Hans Zimmer** en **John Williams**. Twee fantastische (film)muziek componisten die keer op keer ervoor zorgden dat ik mij compleet kon afsluiten van de buitenwereld en zo mij volledig kon concentreren op mijn werk. Ik heb met jullie muziek misschien wel de meeste tijd van mijn promotieonderzoek doorgebracht. Bedankt voor de vele afspeellijsten.

Ik ben ongelofelijk trots en dankbaar voor **mijn lieve familie** die altijd achter mij staan, ongeacht de keuze die ik maak. Onze hechte familieband is enorm waardevol, ik kan mij geen betere familie wensen. Lieve **Papa en Mama**, al van jongs af aan hebben jullie altijd voor een stabiele en veilige basis gezorgd die mij mogelijkheden bood om mezelf te ontwikkelen. Ondanks dat het onderwerp van mijn promotie abracadabra was voor jullie, boden jullie mij altijd een luisterend oor en goede adviezen aan, met als boodschap dat ik dicht bij mezelf moest blijven. Het vertrouwen dat jullie in mij hebben zorgt ervoor dat niks onmogelijk is. **Pap**, het afgelopen jaar zijn we vaak samen op pad geweest, lekker samen klussen of tuinieren (jazeker, een dochter met groene vingers!). Deze praktische klussen gaven mij de tijd om mijn gedachten te verzetten en vrij te komen van mijn PhD. **Mam**, samen maakten we er een gezellig dagje uit van toen ik in het AvL HR-zaken moest regelen, want zo wordt zo'n regelding immers toch veel leuker! Wanneer mijn perfectionisme weer eens iets te veel van mij vroeg, wist je mij altijd gerust te stellen met je opmerking "meer dan je best kun je niet doen". Iets wat ik vast nog vaker van je zal horen. Langzaamaan begin ik zelf ook (eindelijk) in te zien dat ik soms te veel hooi op mijn vork neem en moet accepteren dat genoeg genoeg is. Maar daarvoor zal ik zeker nog een aantal keer je hulp nodig hebben! Lieve **Tessa**, mijn oudste zusje. Dat jij kansen aangrijpt waar ze liggen heb je de afgelopen twee jaren wel bewezen. Door te verhuizen naar Sint Maarten liet je alles achter en ging je een nieuwe uitdaging aan. Het bracht vele mooie dingen met zich mee, waaronder **Kess**, mijn nichtje **Emmay** en een leuk vakantieadres. Laat deze gewaagde stap een inspiratiebron voor anderen zijn die altijd in hun vertrouwde omgeving (Hoogland) blijven. Lieve **Maud**, mijn jongste zusje en ja, een 2^{e} alinea. Een echte spring in 't veld. Met je energie voor 10 houd jij altijd alle ballen in de lucht; studie, werk, vrienden, sporten, je vriendje **Niels** en ook nog klussen in jullie nieuwe huis. Ik bewonder hoe jij dit allemaal combineert, waarbij je alles voor de volle 100% blijft doen. Lieve familie, ik houd ontzettend veel van jullie. Dikke (tweezijdige) knuffel!

Lieve lieve **Martijn**, mijn trotse hobbyboer! Wat vind ik het fijn om samen met jou te zijn. Hoewel ik in de laatste periode van mijn proefschrift niet altijd te genieten was (sorry hiervoor!), steunde jij mij onvoorwaardelijk. Zelfs wanneer je voor werk in het buitenland was, wist je een moment te vinden om mij te bellen en te vragen hoe het met mij ging. Als echt maatje was je altijd betrokken bij mijn onderzoek en durfde je eerlijk je mening te delen (ook al moest ik eerst de vele variaties 'huhm' ontcijferen). Daarnaast kon ik mijn onderzoek ook gemakkelijk loslaten tijdens onze vele fietstochten, kampeervakanties of als we een lammetje moesten vangen. Ik kijk enorm uit om nog veel meer mooie avonturen met jou samen te beleven.

Hoe zorgvuldig zo'n dankwoord ook is geschreven, het is altijd mogelijk dat er onbedoeld iemand is vergeten. Om deze reden wil ik iedereen die onder deze noemer valt, ontzettend bedanken voor zijn/haar inzet.

CURRICULUM VITAE

Paula Bos was born on January 16th 1990, in Amersfoort, the Netherlands. She graduated from secondary school ('t Hooghe Landt College, Amersfoort) in 2008, after which she started studying Technical Medicine at the University of Twente in Enschede. During this study, she performed five clinical internships, where she mastered the different aspects of clinical research. Paula received her Master degree in 2016. Inspired by innovative possibilities in healthcare, Paula started working as a PhD-candidate at the department of radiology and head and neck oncology and surgery of the Antoni van Leeuwenhoek hospital in October 2016, under the supervision of prof. dr. Beets-Tan, prof. dr. van den Brekel, dr. Jasperse. Later, in 2020, prof. dr. Castelijns joined her supervision committee. The focus of the thesis was to develop machine learning models, based on clinical variables and/or magnetic resonance imaging radiomic features, to predict treatment outcome in patients with oropharyngeal cancers. Paula has presented her research results at national and international conferences, such as the Symposium Extranodal Spread in Head & Neck Cancer, European Congress of Radiology (ECR) and International Society for Magnetic Resonance in Medicine (ISMRM). Since April 2022, Paula is working as medical physicist in nuclear medicine at MILabs, Houten.