Quantitative ^{18}F -FDG PET/CT in Esophageal Cancer and Vascular Graft Infections

Jorn Beukinga

Quantitative ^{18}F -FDG PET/CT in Esophageal Cancer and Vascular Graft Infections

Jorn Beukinga

UNIVERSITY OF TWENTE.

**QUANTITATIVE ^{18}F -FDG PET/CT IN ESOPHAGEAL
CANCER AND VASCULAR GRAFT INFECTIONS**

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. T.T.M. Palstra,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op woensdag 11 juli 2018 om 14.45 uur

door

Roelof Jorn Beukinga

geboren op 02 februari 1990
te Ruinen, Nederland

Dit proefschrift is goedgekeurd door:

Supervisors

prof. dr. R.H.J.A. Slart

prof. dr. J.Th.M. Plukker

Co-supervisor

prof. dr. ir. C.H. Slump

Beoordelingscommissie

prof. dr. R.H. Geelkerken

prof. dr. O.S. Hoekstra

prof. dr. J.M. Klaase

prof. dr. M.A.F.J. van de Laar

dr. R.J.H.M. Steenbakkers

prof. dr. ir. A. Stein

ISBN	978-90-365-4565-5
DOI	10.3990/1.9789036545655
URL	https://doi.org/10.3990/1.9789036545655
Cover	Hullo 3D design studio (www.hullo.nl)
Lay-out	R.J. Beukinga
Printing	Ipskamp Printing, Enschede

Copyright © R.J. Beukinga, 2018

All rights reserved. No part of this disseratation may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission from the author.

Paranimfen

T.B. Snuverink, MSc.

R. Wesselink, MSc.

TABLE OF CONTENTS

Chapter 1	General Introduction	9
Part I	Validation of ^{18}F-FDG PET Radiomic Features	
Chapter 2	Standardized Image Biomarkers for High-Throughput Extraction of 3D Images Submitted to <i>Med Image Anal.</i> 2018	23
Chapter 3	Reliability of ^{18}F -FDG PET Radiomic Features: a Phantom Study to Explore Sensitivity to Image Reconstruction Settings, Noise, and Delineation Method Submitted to <i>Med Image Anal.</i> 2018	33
Part II	Applications of ^{18}F-FDG PET/CT Radiomics in Oncology	
Chapter 4	Predicting Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer with Textural Features Derived from Pretreatment ^{18}F -FDG PET/CT Imaging <i>J Nucl Med.</i> 2017;58:723-729	53
Chapter 5	Prediction of Response to Neoadjuvant Chemoradiotherapy with Baseline and Restaging ^{18}F -FDG PET Image Biomarkers in Esophageal Cancer Patients <i>Radiology.</i> 2018;287:983-992	71
Chapter 6	HER2 Expression Integrated into a Clinico-Radiomic Model Enhances Prediction of Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer Patients Submitted to <i>Ann Surg Oncol.</i> 2018	89

Part III	Applications of ^{18}F-FDG PET Radiomics in Infectious Disorders	
Chapter 7	Textural Features of ^{18}F -Fluorodeoxyglucose Positron Emission Tomography Scanning in Diagnosing Aortic Prosthetic Graft Infection <i>Eur J Nucl Med Mol Imaging</i> . 2017;44:886-894	107
Chapter 8	Summary, Genenal Discussion, and Future Perspectives	125
Chapter 9	Nederlandse Samenvatting/Dutch Summary	137
Chapter 10	Author Affiliations	147
	Dankwoord	151
	Over de Auteur	157
	List of Publications	159
Appendix I	Supplemental Data	161

Chapter 1

General Introduction

ABBREVIATIONS

^{18}F -FDG PET	^{18}F -fluorodeoxyglucose positron emission tomography
AGI	aorto-iliac graft infections
CD44	cluster of differentiation 44
CROSS	ChemoRadiotherapy Oesophageal cancer followed by Surgery Study
CT	computed tomography
EC	esophageal cancer
GLCM	gray level co-occurrence matrix
GLRLM	gray level run-length matrix
HER2	human epidermal growth factor receptor 2
HIF1 α	hypoxia-inducible factor alpha
IBSI	image biomarker standardisation initiative
nCRT	neoadjuvant chemoradiotherapy
RECIST	Response Evaluation Criteria in Solid Tumors
PTCH1	patched homolog 1
SHH	Sonic Hedgehog
SUV	standardized uptake value
VOI	volume of interest

INTRODUCTION TO THE THESIS

Personalized treatment is one of the major challenges in modern medicine. To enable individually tailored treatments, medical imaging has become part of the standard diagnostic work-up, allowing a non-invasive anatomical and functional representation of organs. In the last decades, anatomy-based computed tomography (CT) and functional-based ^{18}F -fluorodeoxyglucose positron emission tomography (^{18}F -FDG PET) have been two corner stones in medical imaging. However, clinical evaluation of imaging data remains subject to intra-observer and inter-observer variation. Moreover, most imaging data contain subtle information reflecting underlying pathophysiological properties, which cannot be detected visually. Quantifying medical imaging improves reliability, precision, and speed of the assessment and may contribute to overcome such limitations.

The rapidly emerging field of radiomics generally quantifies large amounts of medical imaging data and applies a large number of quantitative image features to characterize these underlying pathophysiological properties. The majority of available research described radiomic features in oncological applications to support clinical decision-making (1-4), but radiomic features can also be applied in non-oncological disorders, such as infectious diseases. The set of radiomic features can be divided into a number of feature families, of which statistical, local intensity, morphological, and textural features are the most commonly used. In particular, quantifying heterogeneity with textural features seems to capture information which can be used for different clinical applications. The interpretation of radiomic features depends on the chosen image modality (usually CT or ^{18}F -FDG PET). Heterogeneity measured on CT reflects spatial variability in tissue density, resulting from variability in cellularity, vascularization or necrosis, whereas heterogeneity measured on ^{18}F -FDG PET reflects spatial variability in ^{18}F -FDG uptake, resulting from variability in metabolism, necrosis, and hypoxia. Calculation of radiomic features requires a segmentation mask identifying the voxels corresponding to a predefined volume of interest (VOI). Depending on the modality and feature family, different image pre-processing is required including data conversion, voxel interpolation to convert anisotropic to isotropic voxel dimensions, and discretization of gray levels into gray level bins.

Statistical features describe the distribution of gray levels within the VOI, while local intensity features consider gray levels within a defined neighborhood around a center voxel. Morphological and textural features quantify the geometric aspects and the spatial variation of gray-level intensity levels of a VOI, respectively. Textural features are being extracted from matrices expressing the distribution of discretized gray levels in a specified spatial relationship (5-8). The gray level co-occurrence matrix (GLCM) describes how often voxels with discretized gray level i co-occur adjacent to voxels with discretized gray level j (pairwise

arrangement of voxels) (5). In a 3 dimensional fashion, the direct neighborhood of a central voxel with a Chebyshev distance (or maximum norm) of 1 consists of 26 directly neighboring voxels, and thus 13 unique direction vectors Δ , i.e. (0; 0; 1), (0; 1; 0), (1; 0; 0), (0; 1; 1), (0; 1;-1), (1; 0; 1), (1; 0; -1), (1; 1; 0), (1;-1; 0), (1; 1; 1), (1; 1;-1), (1;-1; 1), and (1;-1;-1). To create a $GLCM_{\Delta}$ which is symmetric across its diagonal, the transpose of $GLCM$ is added to the original $GLCM$. Each $GLCM_{\Delta}$ is being normalized by the sum of its elements. To calculate a single feature value, either the feature values for each $GLCM_{\Delta}$ can be averaged or features can be extracted from a single merged matrix in which all $GLCM_{\Delta}$ are summed. The gray level run-length matrix (GLRLM) describes the occurrence of j aligned voxels with discretized gray level i in a direction specified by offset Δ (6). Equivalent to $GLCM$ -based textural feature extraction, single GLRLM feature values can be obtained by either averaging feature values extracted from each $GLRLM_{\Delta}$ or from a merged matrix. The gray level size-zone matrix counts the number of connected voxels j (size-zone) with the same discretized gray level i (7). The neighborhood gray-tone difference matrix describes the gray level difference between voxels with discretized gray level i and the average gray level over their surrounding neighborhoods (8).

The aim of this thesis was to apply quantitative radiomic features to objectively evaluate and adjust the clinical decision-making in the treatment of patients with esophageal cancer and vascular graft infections.

OUTLINE OF THE THESIS

Part I: Validation of ^{18}F -FDG PET Radiomic Features

In **Part I** of this thesis, we investigated the suitability of the most commonly used ^{18}F -FDG PET radiomic features as new potential biomarkers. Although radiomic features have been used for numerous clinical applications, these results are difficult to reproduce and validate. Key challenges in the field of radiomics are therefore the determination of consensus-based feature definitions and guidelines for methodological choices in the feature extraction workflow. Other important challenges facing radiomics are the establishment of feature measurement errors and the reduction of the multitude of features that are generated. Numerous features are likely to exhibit a high mutual correlation or feature redundancy because they are derived from only a few underlying mathematical families. By incorporating redundant features, prediction models may become unstable or overfitted. In overfitted prediction models, noise in the training data is learned by the model. Although this often leads to an improved apparent performance, the internal and external performances usually reduce. To avoid overfitting, the feature space should therefore be reduced to a non-redundant feature set.

In **Chapter 2**, an independent international collaboration was launched to address these issues. With the image biomarker standardisation initiative (IBSI), we provided nomenclature and definitions for the most commonly used radiomic features, an image processing scheme, benchmark data sets and associated values, and a set of reporting guidelines. In phase I of this study, radiomic features were extracted from a small 3D digital phantom, in absence of any additional image processing. The results of 21 participating teams were iteratively compared to reach consensus on radiomic feature definitions and their benchmark values. In phase II of this study, results were iteratively compared to reach consensus on image processing techniques by using a publicly available CT-based radiomics phantom data set (3, 9).

In **Chapter 3**, we investigated the sensitivity of PET radiomic features to several confounding factors in terms of reliability, which is an important measurement of error. We studied underlying data, image reconstruction methods and settings, noise, discretization method, and delineation method, as these have been identified to affect the majority of the radiomic features (10-15). A NEMA NU 2-2012 image quality phantom was used to provide data acquisition under controlled experimental conditions. The phantom consisted of six spheres with various sizes, which were filled with six different ^{18}F -FDG concentrations to simulate clinical data with different characteristics. Moreover, we have investigated the effect of these confounding factors to reduce the feature space in order to obtain a non-redundant set of features.

Part II: Applications of ^{18}F -FDG PET/CT Radiomics in Oncology

Part II and Part III of this thesis concern clinical applications of ^{18}F -FDG PET/CT radiomic features. In **Part II**, we focused on the use of ^{18}F -FDG PET/CT radiomic features to predict the pathologic response of neoadjuvant chemoradiotherapy (nCRT) in patients with locally advanced esophageal cancer (EC). EC patients require an extensive pre-operative imaging work-up for staging and are therefore particularly suitable for the application of radiomics.

EC is the 6th leading cancer type for estimated deaths worldwide in 2012 with 5% of all cancer specific deaths (16). It is an aggressive tumor with early nodal and distant metastases and early recurrences even after histologically confirmed radical surgery. At diagnosis, 20% of the patients have localized disease, 31% have locoregional disease, and 39% have distant metastases (17). The 5-year relative survival strongly depends on the cancer stage at diagnosis and is 42.9% in patients with localized disease, whereas patients diagnosed with regional and distant metastases have substantially lower rates of 23.4% and 4.6%, respectively (17).

The standard staging work-up for patients with EC consists of a diagnostic CT, ^{18}F -FDG PET, or integrated ^{18}F -FDG PET/CT, and endoscopic ultrasound with fine needle aspiration, whether standard or on indication (18-20). The current standard treatment for locally

advanced potentially resectable EC patients is nCRT followed by radical esophagectomy with curative intent. In the Netherlands, nCRT is based on a regimen according to the ChemoRadiotherapy Oesophageal cancer followed by Surgery Study (CROSS) consisting of weekly intravenous administered paclitaxel (50 mg/m²) and carboplatin (area under the plasma concentration curve of 2 mg·min·mL⁻¹) during 5 weeks with concurrent external radiotherapy (41.4 Gy in 23 fractions) (21). Compared to a surgery-alone approach, nCRT effectively leads to a considerable downstaging of the tumor, which improves the curative resectability from 69% to 92% and the 5-year overall survival rate from 34% to 47% (21, 22). However, a majority of these patients are not treated adequately with nCRT due to a considerable nCRT resistance. In fact, 29% of the patients treated in the CROSS trial achieved a pathologic complete response (pCR), 52% achieved a pathologic partial response, and even 18% achieved no pathologic tumor response (21). However, complete responders to nCRT may benefit more from an individualized treatment approach, including a “wait-and-see policy” after omitting initial surgery and potentially curable salvage surgery in the follow-up in selected cases of isolated locoregional recurrence. Consequently, possible severe peri- and postoperative mortality (approximately 4-7%) and complications can be avoided (23-25).

To enable this individualized treatment approach for complete responders, monitoring before, during, and after treatment is required. Response to nCRT is often objectively monitored by determining changes in tumor size on CT according to the Response Evaluation Criteria in Solid Tumors (RECIST) guideline version 1.1 (26). However, this guideline is hampered by several limitations, including that viable tumor tissue can hardly be differentiated from treatment-induced necrotic or fibrotic tissue and that tumors without volumetric shrinkage, but with a metabolic treatment response cannot be identified clearly (27). Moreover, visual assessment of the longest lesion diameter and volumetric tumor shrinkage is accompanied by a relatively high interobserver variability (28). Hence, RECIST has a limited value in the monitoring of treatment response in EC (29).

Monitoring of treatment response is also commonly based on the maximum standardized uptake value (SUV_{max}) on ¹⁸F-FDG PET, which addresses different biophysical tumor tissue properties than CT (glucose metabolism vs. tissue density) (30). However, the SUV_{max} is a single point estimation, which is susceptible to noise artifacts and ignores the intratumoral ¹⁸F-FDG spatial distribution (31). This may be one of the reasons why predictions based on the SUV_{max} yield an unsatisfactory low sensitivity and specificity of 67% and 68%, respectively (32). Consequently, an adequate method to predict pCR after nCRT has not yet been defined in EC patients.

In **Chapter 4**, we used ¹⁸F-FDG PET/CT radiomics to capture intratumoral ¹⁸F-FDG uptake heterogeneity in the prediction of pCR to nCRT in EC. Intratumoral ¹⁸F-FDG uptake

heterogeneity prior to nCRT is often induced by hypoxia (33) and is associated with tumor aggressiveness and treatment resistance (34, 35). Multiple promising ^{18}F -FDG PET/CT radiomic features have been reported to have higher discriminatory value than the SUV_{max} (36-40). We constructed several candidate multivariable prediction models which incorporated multiple complementary pre-treatment clinical factors and pre-treatment ^{18}F -FDG PET derived radiomic features. The prediction models were constructed based on a homogeneous group of locally advanced EC patients treated according to the CROSS regimen.

In **Chapter 5**, the image-based prediction strategy was expanded by analyzing radiomic features derived from both baseline (pre-nCRT) and restaging (post-nCRT) ^{18}F -FDG PET images. Post-nCRT imaging performed just prior to esophagectomy may provide a more appropriate representation of a patient's real response compared to analysis of pre-nCRT imaging alone. Recently, this has been substantiated by several studies with small sample sizes which found that relative radiomic features changes between pre-nCRT and post-nCRT ^{18}F -FDG PET showed promising results in predicting response in EC patients (36-41). In this chapter, the value of baseline and restaging ^{18}F -FDG PET radiomics was assessed.

In **Chapter 6**, the aim was to further extend the ^{18}F -FDG PET/CT clinico-radiomic based signatures, constructed in Chapter 3. Combining predictors derived from clinical, imaging, and basic (e.g. cell biology to determine characteristics of the cell and its components) research areas may provide complementary information and may therefore lead to further model optimization (42-45). We chose to study potentially useful biological tumor markers which have been associated with treatment resistance and tumor heterogeneity, but nonetheless have not yet been implemented in the current decision-making. The additional value of the expressions of the following biological tumor markers were determined: human epidermal growth factor receptor 2 (HER2) (46, 47), cluster of differentiation 44 (CD44) (48, 49), hypoxia-inducible factor alpha (HIF1 α) (50), patched homolog 1 (PTCH1) (50-52), and Sonic Hedgehog (SHH) (50-52).

Part III: Applications of ^{18}F -FDG PET Radiomics in Infectious Disorders

In **Part III**, ^{18}F -FDG PET radiomic features were applied to improve the non-invasive diagnosis of infections. Although there is still no evidence-based indication, ^{18}F -FDG PET/CT is a potential first-line non-invasive diagnostic tool and is suggested to play a role in obtaining a surrogate proof of infections (53). We focused on the application of PET radiomics in the diagnosis of aorto-iliac graft infections (AGI) after vascular graft reconstruction.

AGI is an uncommon complication of abdominal aorto-iliac aneurysm surgery and occlusive vascular disease with an incidence of 1-6%, depending on the location of the graft and the type of the procedure (54). However, the diagnosis and treatment of AGI is commonly

delayed due to aspecific early symptoms, which can result in life-threatening sepsis and/or severe or untreatable hemorrhage. Current selection in the treatment for AGI depends on the onset, extent, and type of AGI and includes: antimicrobial therapy alone, graft preservation, graft resection with in situ graft reconstruction, and graft removal with extra-anatomic bypass (55). The diagnosis of AGI is generally based on clinical findings (such as bioinflammatory markers and sepsis) and microbiological results. Recently, PET/CT imaging has increasingly been applied and is considered an accurate non-invasive diagnostic tool in the assessment of AGI (56). Positive bacterial cultures of puncture material are still considered as the gold standard (57-61), however, perigraft abscesses are not always present or suitable for puncture. However, the evaluation of PET images using traditional SUV metrics such as the SUV_{max} , the tissue-to-background ratio, and the visual grading scale lacks consensus in terms of interpretation (56, 62-65). Besides, non-infected grafts commonly present elevated diffuse ^{18}F -FDG uptake, which is possibly the result of a local sterile inflammatory process around the prosthesis due to a foreign body-related reaction. This leads to a number of false positive findings and complicates the exact diagnosis of infection (66). The ^{18}F -FDG uptake pattern has become of interest, as infections typically lead to a heterogeneous uptake pattern and foreign body reactions in general present as homogenous (62). In **Chapter 7**, we quantified the ^{18}F -FDG uptake pattern along the graft. The objective of this retrospective study was to examine whether ^{18}F -FDG PET textural features correlate with the standard clinical diagnosis of AGI.

Chapter 8 gives a summary, general discussion, and future perspectives of this work.

Chapter 9 contains a summary in Dutch.

REFERENCES

1. O'Connor JPB, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol*. 2017;14:169-186.
2. Yip SSF, Aerts HJW. Applications and limitations of radiomics. *Phys Med Biol*. 2016;61:150-166.
3. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14:749-762.
4. Aerts HJW, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
5. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern*. 1973;3:610-621.
6. Galloway MM. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*. 1975;4:172-179.
7. Thibault G, Angulo J, Meyer F. Advanced statistical matrices for texture characterization: application to cell classification. *IEEE Trans Biomed Eng*. 2014;61:630-637.
8. Amadasun M, King R. Textural features corresponding to textural properties. *IEEE Trans Syst Man Cybern*. 1989;19:1264-1274.

9. Lambin P. Radiomics Digital Phantom. Accessed: March 6, 2018. Available: <https://www.cancerdata.org>. DOI: 10.17195/candat.2016.08.1.
10. Frings V, van Velden FH, Velasquez LM, et al. Repeatability of metabolically active tumor volume measurements with FDG PET/CT in advanced gastrointestinal malignancies: a multicenter study. *Radiology*. 2014;273:539-548.
11. Leijenaar RTH, Carvalho S, Velazquez ER, et al. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol*. 2013;52:1391-1397.
12. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol*. 2010;49:1012-1016.
13. Yan J, Chu-Shern JL, Loi HY, et al. Impact of image reconstruction settings on texture features in ^{18}F -FDG PET. *J Nucl Med*. 2015;56:1667-1673.
14. Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A. The impact of image reconstruction settings on ^{18}F -FDG PET radiomic features: multi-scanner phantom and patient studies. *Eur Radiol*. 2017;27:4498-4509.
15. van Velden FH, Kramer GM, Frings V, et al. Repeatability of radiomic features in non-small-cell lung cancer [^{18}F]FDG-PET/CT studies: impact of reconstruction and delineation. *Mol Imaging Biol*. 2016;18:788-795.
16. Ferlay J, Soerjomataram I, Ervik M, et al. GLOBOCAN 2012 v1.0, Cancer incidence and mortality worldwide: IARC CancerBase No. 11. Accessed: October 3, 2017; Lyon, France: International Agency for Research on Cancer; 2013. Available: <http://globocan.iarc.fr>.
17. Surveillance, Epidemiology, and End Results (SEER) Program. Cancer of the esophagus: 5-year relative and period survival (percent) by race, sex, diagnosis year, stage and age. Accessed: October 10, 2017; Available: https://seer.cancer.gov/archive/csr/1975_2011/results_single/sect_08_table.08.pdf.
18. Berry MF. Esophageal cancer: staging system and guidelines for staging and treatment. *J Thorac Dis*. 2014;6 Suppl 3:289.
19. Parry K, Haverkamp L, Bruijnen RC, et al. Staging of adenocarcinoma of the gastroesophageal junction. *Eur J Surg Oncol*. 2016;42:400-406.
20. Oesofaguscarcinoom landelijke richtlijn, versie 3.1; Oncoline.
21. van Hagen P, Hulshof MC, van Lanschot JJ, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. *N Engl J Med*. 2012;366:2074-2084.
22. Oppedijk V, van der Gaast A, van Lanschot JJ, et al. Patterns of recurrence after surgery alone versus preoperative chemoradiotherapy and surgery in the CROSS trials. *J Clin Oncol*. 2014;32:385-391.
23. Dikken JL, Dassen AE, Lemmens VE, et al. Effect of hospital volume on postoperative mortality and survival after oesophageal and gastric cancer surgery in the Netherlands between 1989 and 2009. *Eur J Cancer*. 2012;48:1004-1013.
24. Bosch DJ, Pultrum BB, de Bock GH, Oosterhuis JK, Rodgers MG, Plukker JT. Comparison of different risk-adjustment models in assessing short-term surgical outcome after transthoracic esophagectomy in patients with esophageal cancer. *Am J Surg*. 2011;202:303-309.
25. Talsma AK, Lingsma HF, Steyerberg EW, Wijnhoven BP, Van Lanschot JJ. The 30-day versus in-hospital and 90-day mortality after esophagectomy as indicators for quality of care. *Ann Surg*. 2014;260:267-273.
26. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45:228-247.
27. Weber WA. Assessing tumor response to therapy. *J Nucl Med*. 2009;50 Suppl 1:10S.
28. Yoon SH, Kim KW, Goo JM, Kim DW, Hahn S. Observer variability in RECIST-based tumour burden measurements: a meta-analysis. *Eur J Cancer*. 2016;53:5-15.
29. Yanagawa M, Tatsumi M, Miyata H, et al. Evaluation of response to neoadjuvant chemotherapy for esophageal cancer: PET response criteria in solid tumors versus response evaluation criteria in solid tumors. *J Nucl Med*. 2012;53:872-880.
30. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors. *J Nucl Med*. 2009;50 Suppl 1:50S.
31. Lodge MA, Chaudhry MA, Wahl RL. Noise considerations for PET quantification using maximum and peak standardized uptake value. *J Nucl Med*. 2012;53:1041-1047.
32. Kwee RM. Prediction of tumor response to neoadjuvant therapy in patients with esophageal cancer with use of ^{18}F FDG PET: a systematic review. *Radiology*. 2010;254:707-717.

33. Harris AL. Hypoxia--a key regulatory factor in tumour growth. *Nat Rev Cancer*. 2002;2:38-47.
34. Dexter DL, Leith JT. Tumor heterogeneity and drug resistance. *J Clin Oncol*. 1986;4:244-257.
35. O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin Cancer Res*. 2015;21:249-257.
36. van Rossum PS, Fried DV, Zhang L, et al. The Incremental Value of Subjective and Quantitative Assessment of ^{18}F -FDG PET for the Prediction of Pathologic Complete Response to Preoperative Chemoradiotherapy in Esophageal Cancer. *J Nucl Med*. 2016;57:691-700.
37. Tan S, Kligerman S, Chen W, et al. Spatial-temporal [^{18}F]FDG-PET features for predicting pathologic response of esophageal cancer to neoadjuvant chemoradiation therapy. *Int J Radiat Oncol Biol Phys*. 2013;85:1375-1382.
38. Tan S, Zhang H, Zhang Y, Chen W, D'Souza WD, Lu W. Predicting pathologic tumor response to chemoradiotherapy with histogram distances characterizing longitudinal changes in ^{18}F -FDG uptake patterns. *Med Phys*. 2013;40:101707.
39. Zhang H, Tan S, Chen W, et al. Modeling pathologic response of esophageal cancer to chemoradiation therapy using spatial-temporal ^{18}F -FDG PET features, clinical parameters, and demographics. *Int J Radiat Oncol Biol Phys*. 2014;88:195-203.
40. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline ^{18}F -FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med*. 2011;52:369-378.
41. Yip SS, Coroller TP, Sanford NN, Mamon H, Aerts HJ, Berbeco RI. Relationship between the Temporal Changes in Positron-Emission-Tomography-Imaging-Based Textural Features and Pathologic Response and Survival in Esophageal Cancer Patients. *Front Oncol*. 2016;6:72.
42. Bollschweiler E, Holscher AH, Schmidt M, Warnecke-Eberz U. Neoadjuvant treatment for advanced esophageal cancer: response assessment before surgery and how to predict response to chemoradiation before starting treatment. *Chin J Cancer Res*. 2015;27:221-230.
43. Tao CJ, Lin G, Xu YP, Mao WM. Predicting the response of neoadjuvant therapy for patients with esophageal carcinoma: an in-depth literature review. *J Cancer*. 2015;6:1179-1186.
44. Vallbohmer D, Brabender J, Grimminger P, Schroder W, Holscher AH. Predicting response to neoadjuvant therapy in esophageal cancer. *Expert Rev Anticancer Ther*. 2011;11:1449-1455.
45. Forshaw MJ, Gossage JA, Mason RC. Neoadjuvant chemotherapy for oesophageal cancer: the need for accurate response prediction and evaluation. *Surgeon*. 2005;3:373-382.
46. Gowryshankar A, Nagaraja V, Eslick GD. HER2 status in Barrett's esophagus & esophageal cancer: a meta analysis. *J Gastrointest Oncol*. 2014;5:25-35.
47. Miller S, Hung M. Regulation of her2/neu gene-expression (review). *Oncol Rep*. 1995;2:497-503.
48. Zoller M. CD44: can a cancer-initiating cell profit from an abundantly expressed molecule? *Nat Rev Cancer*. 2011;11:254-267.
49. Smit JK, Faber H, Niemantsverdriet M, et al. Prediction of response to radiotherapy in the treatment of esophageal cancer using stem cell markers. *Radiother Oncol*. 2013;107:434-441.
50. Wang Y, Ding Q, Yen CJ, et al. The crosstalk of mTOR/S6K1 and Hedgehog pathways. *Cancer Cell*. 2012;21:374-387.
51. Yang Z, Cui Y, Ni W, Kim S, Xuan Y. Gli1, a potential regulator of esophageal cancer stem cell, is identified as an independent adverse prognostic factor in esophageal squamous cell carcinoma. *J Cancer Res Clin Oncol*. 2017;143:243-254.
52. Ma X, Sheng T, Zhang Y, et al. Hedgehog signaling is activated in subsets of esophageal cancers. *Int J Cancer*. 2006;118:139-148.
53. Jamar F, Buscombe J, Chiti A, et al. EANM/SNMMI guideline for ^{18}F -FDG use in inflammation and infection. *J Nucl Med*. 2013;54:647-658.
54. Seeger JM. Management of patients with prosthetic vascular graft infection. *Am Surg*. 2000;66:166-177.
55. Wilson WR, Bower TC, Creager MA, et al. Vascular Graft Infections, Mycotic Aneurysms, and Endovascular Infections: A Scientific Statement From the American Heart Association. *Circulation*. 2016;134:e460.
56. Keidar Z, Nitecki S. FDG-PET in prosthetic graft infections. *Semin Nucl Med*. 2013;43:396-402.
57. Valentine RJ. Diagnosis and management of aortic graft infection. *Semin Vasc Surg*. 2001;14:292-301.

58. O'Connor S, Andrew P, Batt M, Becquemin JP. A systematic review and meta-analysis of treatments for aortic graft infection. *J Vasc Surg.* 2006;44:38-45.
59. Saleem BR, Meerwaldt R, Tielliu IF, Verhoeven EL, van den Dungen, J J, Zeebregts CJ. Conservative treatment of vascular prosthetic graft infection is associated with high mortality. *Am J Surg.* 2010;200:47-52.
60. Legout L, D'Elia PV, Sarraz-Bournet B, et al. Diagnosis and management of prosthetic vascular graft infections. *Med Mal Infect.* 2012;42:102-109.
61. Perera GB, Fujitani RM, Kubaska SM. Aortic graft infection: update on management and treatment options. *Vasc Endovascular Surg.* 2006;40:1-10.
62. Fukuchi K, Ishida Y, Higashi M, et al. Detection of aortic graft infection by fluorodeoxyglucose positron emission tomography: comparison with computed tomographic findings. *J Vasc Surg.* 2005;42:919-925.
63. Keidar Z, Engel A, Hoffman A, Israel O, Nitecki S. Prosthetic vascular graft infection: the role of 18F-FDG PET/CT. *J Nucl Med.* 2007;48:1230-1236.
64. Bruggink JL, Glaudemans AW, Saleem BR, et al. Accuracy of FDG-PET-CT in the diagnostic work-up of vascular prosthetic graft infection. *Eur J Vasc Endovasc Surg.* 2010;40:348-354.
65. Saleem BR, Berger P, Vaartjes I, et al. Modest utility of quantitative measures in (18)F-fluorodeoxyglucose positron emission tomography scanning for the diagnosis of aortic prosthetic graft infection. *J Vasc Surg.* 2015;61:965-971.
66. Keidar Z, Pirmisashvili N, Leiderman M, Nitecki S, Israel O. ¹⁸F-FDG uptake in noninfected prosthetic vascular grafts: incidence, patterns, and changes over time. *J Nucl Med.* 2014;55:392-395.

Part I

Validation of ^{18}F -FDG PET Radiomic Features

Chapter 2

Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping

Image Biomarker Standardization Initiative*

Submitted to *Med Image Anal.* 2018

* Author names and affiliations are given on page 147

ABBREVIATIONS

IBSI	Image Biomarker Standardization Initiative
3D	three-dimensional
CT	computed tomography

ABSTRACT

Image biomarkers are increasingly employed for a diverse range of applications in medicine and health, but lack standardization. The Image Biomarker Standardization Initiative (IBSI) has assessed 172 image biomarkers for three-dimensional digital imaging and provides calculation benchmarks, definitions, data sets and reporting guidelines.

MAIN TEXT

Applications of quantitative image analysis in the life sciences and medicine are becoming increasingly diverse (1, 2). The use of high-throughput computer algorithms for quantitative image analysis has been reported to improve reliability, speed, and precision over what could be achieved by visual assessment or manual analyses (3-6). For example, in the context of clinical cancer research, high-throughput analysis of three-dimensional (3D) medical imaging (radiomics) is used to characterize tumor phenotypes, predict treatment outcomes, and assess tissue malignancy, among others (7-10). Additionally, biomarkers (11) from medical imaging were found to be linked to molecular expressions of the underlying tumor tissue (12, 13). Within quantitative cell biology, high-content screening combines automated microscopy with quantitative image analysis for the identification of phenotypes of interest (5, 6). Furthermore, quantitative 3D image analysis of large tissue samples and organoids is now possible due to recent advances in sample preparation (14), and is expected to lead to new fundamental insights (15, 16).

Despite many interesting and successful applications, studies involving high-throughput quantitative image analysis have been difficult to reproduce and validate (4, 8, 9, 17), and the field has been described as a “wild frontier” (6). Important challenges in this respect are the lack of consensus-based definitions and recommendations for calculating image biomarkers and the absence of reporting guidelines. The Image Biomarker Standardization Initiative (IBSI) was formed to address these challenges by fulfilling the following objectives: i) to establish a nomenclature and definitions for commonly used image biomarkers; ii) to establish an image processing scheme for calculation of image biomarkers from acquired imaging; iii) to provide benchmark data sets and associated values for verification and testing of software implementations for image processing and image biomarker calculation; and iv) to provide a set of reporting guidelines for studies involving high-throughput quantitative image analysis.

These objectives were reached through two distinct phases (see Fig. 1), which took place between June 2016 and February 2018. In phase I, the main objective was to reach consensus on image biomarker definitions and their benchmark values, in absence of any additional image processing. In phase II, which started in January 2017, the main objective was to reach consensus on image processing techniques.

In phase I, a set of 172 commonly used image biomarkers was first identified and documented. This set included biomarkers used to quantify the morphology, texture and first-order statistical aspects of regions of interest in a 3D image. Subsequently, software implementations of 21 participating teams were compared against a small 3D digital phantom

FIGURE 1 Quantifying a three-dimensional image with a set of image biomarkers requires image processing and subsequent image biomarker calculation. The objective of the IBSI is to provide guidelines and benchmarks for these steps. The benchmarks were derived in two distinct phases. In phase I, a specifically designed digital phantom was used to benchmark image biomarkers without the requirement of performing image processing. Subsequently, in phase II, a publicly available computed tomography (CT)-based radiomics phantom was used to benchmark image processing.

that was specifically designed for this study and did not require any image processing steps before biomarker calculation. During phase I, results were iteratively compared over time and divergences in definitions and implementations were resolved to achieve consensus on benchmark values. Agreement on a benchmark value was defined when 50% of the teams (minimum 3), weighted by their overall benchmarking accuracy, contributed the same value. Initially, agreement on benchmark values existed only for 23.2% of the image biomarkers, whereas, ultimately, agreement was reached for 99.4% of the image biomarkers (see Fig. 2).

In phase II, the above process was repeated to standardize image processing. A general image processing scheme was first defined based on frameworks reported in literature (17). Within this scheme, important image processing steps such as image interpolation and image intensity discretization were investigated in detail. Several alternative methods are available for most image processing steps. For example, interpolation of 3D images may be performed slice-by-slice (2D) or volumetrically, and by using different algorithms such as linear or cubic interpolation. Hence, five different image processing configurations were defined to cover a representative range of methods. Subsequently, image processing and image biomarker calculation were performed on a publicly available computed tomography (CT)-based radiomics phantom data set (9, 18) for each of the five configurations. Similarly to phase I, results from the different software implementations were iteratively compared, and divergence in definitions and implementations were resolved to reach consensus on benchmark values.

FIGURE 2 Development of consensus over time for phase I and phase II, as a percentage of the total number of image biomarkers. For phase II the average percentage over five image processing configurations is shown. Initial results were collected at time points 1 and 10 for phase I and II, respectively. Consensus was defined to exist for image biomarkers for which $\geq 50\%$ of contributing teams (minimum 3), weighted by overall accuracy, arrive at the same value. Additional image biomarkers were introduced between time points 4 and 5. Between time points 11 and 12, image interpolation was fully detailed.

In phase II, on average 27.7% (range: 12.6–46.5%) of the image biomarkers across the five configurations could be benchmarked initially (see Fig. 2). At the end of phase II, this number increased to 96.4% (range: 94.0–97.7%).

Our investigations resulted in an IBSI reference manual (see supplementary files [available at <https://arxiv.org/abs/1612.07003>]), containing a full description of the image processing scheme (chapter 2), the description of the 172 image biomarkers (chapter 3), identification of extensions to existing image biomarkers (chapter 4), reporting guidelines and biomarker nomenclature (chapter 5), a description of the benchmark data sets and instructions on how to use them (chapter 6) and the corresponding benchmark values (chapter 7).

Overall, IBSI has focused on creating a broadly applicable framework for calculating image biomarkers from 3D images. The framework is directly applicable to CT imaging. Specific imaging modalities such as positron-emission tomography and 3D microscopy, usually require additional image processing steps such as conversion into standardized uptake values (19) or field-of-view illumination correction (6, 20). Such steps, although discussed in the reference manual, were not assessed in detail in the current study. Moreover, we must

indicate that the set of 172 image biomarkers investigated in this work is not exhaustive. For example, fractals and image filters may likewise be used to analyze 3D images (2). The assessment of additional sets of image biomarkers is envisioned for the next installment of the IBSI. Furthermore, our work addresses only one part of the reproducibility challenge. Image acquisition protocols and equipment, segmentation of regions of interest, and modelling techniques, among others, constitute additional sources of variability in (high-throughput) image biomarker studies, and their harmonization is the focus of other consortia and professional societies, see e.g. O'Connor et al. (7).

In conclusion, the large initial discrepancy between image biomarker values across the different implementations of the participating teams of the IBSI clearly highlights the need for standardization of image biomarkers and image processing, in order to facilitate meta-analysis and reproducibility of published results. Fortunately, this work has also demonstrated the possibility of achieving methodological consensus for the calculation of quantitative image biomarkers. The reference manual, together with the benchmark data sets, values, and reporting guidelines produced by the IBSI constitute important tools available to the community to increase reproducibility of high-throughput image biomarker studies using 3D digital imaging.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the valuable support of Dr. Hesham Elhalawani, Dr. Jayashree Kalpathy-Cramer, Dr. Dennis Mackin, Ida A. Nissen and Dr. Maqsood Yacub. The authors would also like to acknowledge support from the German Federal Ministry of Education and Research (BMBF-0371N52), the French National Institute of Cancer (INCa project C14020NS), the French National Research Agency (Investing for the Future: ANR-10-LABX-07-01), the B-COM Institute of Research and Technologies in Rennes, France, the USA National Institutes of Health and National Cancer Institute (P30 CA008748; R01 CA198121; U01CA143062; U01 CA187947; U01 CA190234; U24 CA180918; U24 CA194354), the European Research Council (ERCADG-2015: 694812-Hypoximmuno; ERC StG-2013: 335367 bio-iRT), the QuIC-ConCePT project (IMI JU: 115151), the Dutch technology Foundation STW (10696-DuCAT; P14-19-Radiomics STRaTegy), the EU 7th framework program (ARTFORCE-257144; REQUITE-601826), SME Phase 2 (EU proposal 673780-RAIL), EUROSTARS (DART, eDECIDE), the European Program H2020-2015-17 (BD2Decide-PHC30-689715; ImmunoSABR-733008), Interreg V-A Euregio Meuse-Rhine (Euradiomics), the Netherlands Organisation for Health Research and Development (10-10400-98-14002), the Swiss National Science Foundation (PZ00P2-154891; 310030 173303),

the UK Engineering and Physical Sciences Research Council (EP/M507842/1) and the Clinical Research Priority Program Tumor Oxygenation of the University of Zurich.

REFERENCES

1. Myers G. Why bioimage informatics matters. *Nat Methods*. 2012;9:659-660.
2. Depeursinge A, Foncubierta-Rodriguez A, Van De Ville D, Müller H. Three-dimensional solid texture analysis in biomedical imaging: review and opportunities. *Med Image Anal*. 2014;18:176-196.
3. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48:441-446.
4. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278:563-577.
5. Usaj MM, Styles EB, Verster AJ, Friesen H, Boone C, Andrews BJ. High-content screening for quantitative cell biology. *Trends Cell Biol*. 2016;26:598-611.
6. Caicedo JC, Cooper S, Heigwer F, et al. Data-analysis strategies for image-based cell profiling. *Nat Methods*. 2017;14:849-863.
7. O'Connor JPB, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol*. 2017;14:169-186.
8. Yip SSF, Aerts HJW. Applications and limitations of radiomics. *Phys Med Biol*. 2016;61:150-166.
9. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14:749-762.
10. Aerts HJW, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
11. Buckler AJ, Bresolin L, Dunnick NR, et al. Quantitative imaging test approval and biomarker qualification: interrelated but distinct activities. *Radiology*. 2011;259:875-884.
12. Panth KM, Leijenaar RTH, Carvalho S, et al. Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. *Radiother Oncol*. 2015;116:462-466.
13. Grossmann P, Stringfield O, El-Hachem N, et al. Defining the biological basis of radiomic phenotypes in lung cancer. *Elife*. 2017;6:e23421.
14. Li W, Germain RN, Gerner MY. Multiplex, quantitative cellular analysis in large tissue volumes with clearing-enhanced 3D microscopy (Ce3D). *Proc Natl Acad Sci U S A*. 2017;114:7321-7330.
15. Coutu DL, Kokkalis KD, Kunz L, Schroeder T. Multicolor quantitative confocal imaging cytometry. *Nat Methods*. 2018;15:39-46.
16. Rios AC, Clevers H. Imaging organoids: a bright future ahead. *Nat Methods*. 2018;15:24-26.
17. Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present... any future? *Eur J Nucl Med Mol Imaging*. 2017;44:151-165.
18. Lambin P. Radiomics Digital Phantom. Accessed: March 6, 2018. Available: <https://www.cancerdata.org>. DOI: 10.17195/candat.2016.08.1. <https://www.cancerdata.org>. 2016.
19. Boellaard R, Delgado-Bolton R, Oyen WJG, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging*. 2015;42:328-354.
20. Smith K, Li Y, Piccinini F, et al. CIDRE: an illumination-correction method for optical microscopy. *Nat Methods*. 2015;12:404.

Chapter 3

Reliability of ^{18}F -FDG PET Radiomic Features: a Phantom Study to Explore Sensitivity to Image Reconstruction Settings, Noise, and Delineation Method

Elisabeth Pfaehler*
Roelof J. Beukinga*
Johan R. de Jong
Riemer H.J.A. Slart
Cornelis H. Slump
Rudi A.J.O. Dierckx
Ronald Boellaard

*Both authors contributed equally.

Submitted to *Med Image Anal.* 2018

ABBREVIATIONS

^{18}F -FDG	^{18}F -fluoro-2-deoxy-D-Glucose
EARL	European Association of Nuclear Medicine Research Ltd
ICC	intracorrelation coefficient
PET	positron emission tomography
PSF	point spread function
SBR	sphere-to-background ratios
SUV	standardized uptake value
TLG	total lesion glycolysis
VOI	volume of interest

ABSTRACT

Background: ^{18}F -fluoro-2-deoxy-D-Glucose positron emission tomography (^{18}F -FDG PET) radiomics has the potential to guide the clinical decision making, but validation is required before it can be implemented in the clinical setting. The aim of this study was to explore how feature space reduction and reliability of ^{18}F -FDG PET radiomic features are affected by various sources of variation such as underlying data (e.g. object size and contrast), image reconstruction methods and settings, noise, discretization method, and delineation method.

Methods: The NEMA image quality phantom was scanned with various sphere-to-background ratios, including cold spots. Data were reconstructed using various matrix sizes, reconstruction algorithms, and scan durations (noise). For every specific reconstruction and noise level, ten statistically equal replicates were generated. The spheres were delineated using CT and PET-based segmentation methods. A total of 246 radiomic features was extracted from each image dataset. Before textural features were calculated, the images were discretized with a fixed number of 64 bins (FBN) and a fixed bin width (FBW) of 0.25. In terms of feature reduction, we determined the impact of these factors on the composition of feature clusters, which were defined on the basis of Spearman's correlation matrices. To assess the reliability of the features, the intraclass correlation coefficient (ICC) of the features was calculated over the ten replicates. An $\text{ICC} > 0.8$ was considered to represent good reliability.

Results: In general, hot spots and larger spheres resulted in higher reliability compared to cold spots and smaller spheres. E.g. for an EARL-compliant reconstruction, larger and smaller hot spheres yielded good reliability for 35% and 27% of the features. For the cold spots this was the case for 22% and 20% of the features. Images reconstructed with point-spread-function (PSF) resulted in the highest reliability when compared with ordered subset expectation maximization (OSEM) or time-of-flight, e.g. 53%, 30%, and 32% of reliable features, respectively (for unsmoothed data, discretized with FBN, 300s scan duration). Reducing image noise (increasing scan duration and smoothing) and using CT-based segmentation for the cold spheres yielded increased reliability. FBW discretization resulted in higher reliability than FBN discretization, e.g. 89% and 35% of the features, respectively (for the EARL-compliant reconstruction and larger-hot spots).

Conclusion: The high sensitivity of PET radiomic features to image quality suggests that image acquisition and preprocessing need to be standardized in order to use ^{18}F -FDG PET radiomics as quantitative imaging biomarkers.

Key Words: ^{18}F -FDG PET/CT radiomic features; image reconstruction settings; delineation

INTRODUCTION

¹⁸F-fluoro-2-deoxy-D-Glucose (¹⁸F-FDG) positron emission tomography (PET) has become part of the routine oncological diagnostic workup and has been applied for treatment response monitoring and prognosis due to its ability to non-invasively visualize organs and lesions. Although qualitative visual image assessment remains important for these purposes, it has a limited capability to objectively quantify tracer uptake. The most widely used semi-quantitative measures are the maximum, mean, and peak standardized uptake value (SUV_{max} , SUV_{mean} , and SUV_{peak}) and morphologically-based imaging features, such as the metabolic tumor volume (MTV) or total lesion glycolysis (TLG) (1-3). However, these features ignore the intratumoral ¹⁸F-FDG spatial distribution (4). The rapidly emerging field of ‘radiomics’ computes a large number of quantitative image features to characterize this intratumoral distribution or other tumor phenotypes such as shape (5-7).

Even though radiomics has the potential to add valuable information to the visual image evaluation in various cancer types (8), several challenges need to be addressed before radiomics can safely be implemented in the clinical setting. One of the key problems with generating a multitude of features is the risk of false positive findings due to multiple testing. Moreover, numerous features may represent similar tracer uptake characteristics, and may therefore be correlated and redundant (9). As models composed of redundant features may become unstable and difficult to interpret, it is required to reduce the feature space to a degree that is manageable for clinical use without losing important information. However, the identification of non-redundant features is challenging. Possible solutions to reduce the feature space would be the use of principal component analysis or (hierarchical) clustering, based on correlation analysis or distance metrics (10). Another challenge facing radiomic features is the establishment of their measurement error (i.e. reproducibility, repeatability, and reliability). Several studies have shown that the majority of the ¹⁸F-FDG PET radiomic features are sensitive to numerous sources such as image acquisition, reconstruction protocols, or delineation method (11-16). However, in these studies it was not possible to ascribe feature variability to measurement errors caused by external sources of variability or to genuine differences in image characteristics which are adequately captured by the feature. Conversely, features with low variability are relatively robust to sources of variation, but may therefore be unable to capture true differences in data.

The aim of this study was to explore how dimensionality reduction and reliability of ¹⁸F-FDG PET radiomic features are affected by the underlying data, image reconstruction methods and settings, noise, discretization method, and delineation method.

MATERIALS AND METHODS

Phantom Experiments

To ensure that results were not biased by specific characteristics of clinical data, phantom experiments within a controlled environment were carried out. The NEMA NU 2-2012 image quality phantom was used consisting of a background volume of 9400 mL and six fillable spheres with inner diameters of 10, 13, 17, 22, 28 and 37 mm. The phantom was filled with different ^{18}F -FDG concentrations. Two hot and two cold spot scans were acquired with sphere-to-background ratios (SBRs) of about 10:1, 5:1, 0.5:1, and 0.25:1. The spheres were filled with 22.6, 10.87, 1.08, and 0.65 kBq/mL measured with a dosiscalibrator (Veenstra instruments, VDC 2.0.2), while the background was filled with 2.4, 2.26, 2.12, and 2.68 kBq/mL, respectively. All phantom scans were acquired as 70 minutes list-mode data on a PET/CT system (Biograph mCT-40 PET/CT, Siemens, Knoxville, TN, USA). The data were reconstructed to obtain a frame of 30, 60, 120, and 300s. For every scan duration, nine additional frames were reconstructed such that they contained the same amount of counts, taking into account the decay of the tracer. Each data set was reconstructed using iterative ordered subset expectation maximization (OSEM) algorithm (3 iterations, 24 subsets) and the vendor provided time-of-flight (TOF) iterative reconstruction method (3 iterations, 21 subsets). Furthermore, all scans were reconstructed with and without resolution modeling (or point spread function [PSF]). The data were reconstructed with an image matrix size of $256 \times 256 \times 111$ and a voxel size of $3.01 \times 3.01 \times 2$ mm. The TOF reconstructions with and without PSF were also reconstructed with a matrix size of $400 \times 400 \times 111$ leading to a voxel size of $2 \times 2 \times 2$ mm. A low dose CT scan (80 kV, 30 mAs, and 2 mm slice thickness) of the phantom was generated in order to calculate the attenuation map of the PET image. To obtain quantitative PET data, images were corrected for attenuation, scatter, random coincidences, and normalization. Images were smoothed with Gaussian filters of 0, 2, 4, 6, and 8 mm full width at half maximum (FWHM) and were converted to SUV so that the mean phantom background SUV was equal to 1 (17).

Segmentation

The spheres of the phantom were segmented using low dose CT- and PET-based delineation methods. The CT-based volume of interest (VOI) was generated by the manual placement of a sphere-shaped VOI with corresponding sphere diameter. The PET-based segmentations were generated using a region growing method using a connectivity of 26 voxels implemented in Matlab 2014b (Mathworks, Natick, MA, USA). For the hot spheres, the segmented region grew from the center voxel of the highest SUV_{peak} seed point till voxel intensities became less than 41% of this SUV_{peak} (17). Conversely, for the cold spheres the segmented region grew from the center voxel of the lowest SUV_{peak} seed point till the voxel

intensities became larger than a SUV of 0.59. To prevent excessive overestimation of the actual sphere volume, the PET-based segmentation was limited to a sphere volume of 300% of the CT-segmented sphere volume. As texture analysis in 3 dimensions requires the VOI to be specified in all 3 spatial dimensions, only those segmentations that eventually resulted in an actual 3D VOI were considered for feature extraction (i.e. segmentations of 1 or 2 voxels or those located in a single image plane were discarded).

Radiomic Feature Extraction

Image processing and feature extraction was performed using Matlab 2014b . For each VOI, 246 radiomic features were calculated, including 19 morphological features, 3 local intensity features, 18 statistical features, and 206 textural features (100 gray level co-occurrence based features, 64 gray level run length based features, 32 gray level size zone based features, and 10 neighborhood gray tone difference based features) (18). Textural features were extracted from discretized image stacks that reduced the continuous-scaled SUV to a countable number of intensity values. Image stacks were discretized using a fixed number of 64 bins (FBN) and a fixed bin width (FBW) of 0.25. Images were analyzed in both 2 and 3 dimensions with a connectivity of 8 and 26 voxels, respectively (using a Chebyshev norm of 1). Single feature values derived from the gray level co-occurrence and gray level run length matrices were calculated by both averaging the obtained feature values over all directions and by extracting the features values directly from a single merged matrix in which the gray level co-occurrence or gray level run length matrices over all directions were summed. We ensured that image processing and feature calculation matched publicly available benchmark values of digital phantom and patient test data (18).

Feature Space Reduction

To reduce the feature space, clusters of features with the same properties were identified using a Spearman correlation matrix of the CT-segmented features, evaluating the monotonic relationship between features. The correlation matrix was ordered by minimizing the mean correlation difference between neighboring features. A cluster was defined by features that had mutual Spearman's correlation coefficients of > 0.7 .

We have determined whether the composition of feature clusters was affected by discretization, reconstruction algorithm, sphere size, and activity uptake. For defining the correlation matrices we used the default settings: all activity uptakes and sphere sizes, a European Association of Nuclear Medicine Research Ltd (EARL) compliant reconstruction (OSEM, 4 mm FWHM, 120s scan duration) (19), matrix size $256 \times 256 \times 111$, CT-based segmentation, and FBW discretization. The clusters of this default correlation matrix were compared with the clusters of other correlation matrices which were composed on the basis

of different settings for discretization (FBW and FBN) and reconstruction (OSEM and PSF). Subsequently, the data of the default setting was divided into four sub-categories: larger (diameters of 37, 28 and 22 mm) – hot spheres, larger–cold spheres, smaller (diameters of 17, 13 and 10 mm) – hot spheres, and smaller–cold spheres. In this case, all clusters were compared against the clusters of the default correlation matrix of the larger-hot spheres. Moreover, we have compared the clusters of all statistically equal replicates using the default settings to ensure that all found differences in the composition of feature clusters could be ascribed to the sources of variation.

Reliability Analysis

Reliability was evaluated using the intraclass correlation coefficient (ICC), calculated with the irr package (version 0.84), available from the Comprehensive R Archive Network (<http://www.r-project.org>). A two-way single measure model was used to evaluate the consistency of the replicates of each setting. The ICC is the ratio of the inter-cluster variance and the sum of the intra-cluster and inter-cluster variability. Therefore, ICC values lie between 0 to 1, with 1 indicating perfect reliability.

Before extracting the ICCs, the data were split into the same four different underlying data sub-categories that were used for the redundancy analysis. The ICC was calculated for every combination of sub-category, matrix size, reconstruction algorithm, scan duration, Gaussian filter, discretization method, and segmentation method. Each sphere with a different size or SBR was considered a different subject. The equivalent replicates were regarded as the different raters. Features exhibiting an ICC > 0.8 were considered to represent good reliability. For each setting, the percentage of reliable features was obtained to identify trends in the data. Smaller subsets of features were analyzed in order to avoid that large groups of features with similar properties overrepresented and biased the analysis. For this purpose, we used a predefined set of uncorrelated radiomic features that was identified previously (Supplemental Data) (9).

To investigate the potential relationship between the reliability of radiomic features and image noise, a variance image of the statistically equal replicates was calculated for every studied setting. The image noise was measured by calculating the coefficient of variation over four different spherical VOIs defined in the phantom background of the variance image.

RESULTS

Feature Space Reduction

Fig. 1 and Fig. 2 demonstrate how the Spearman's correlation matrix was affected by reconstruction algorithm and discretization method (Fig. 1), as well as by sphere size and

activity uptake (Fig. 2). In order to illustrate the differences in correlation, the feature order and cluster composition of the default setting were used to display the correlation matrices of the other settings. The correlation matrix of this setting is displayed in the upper left corner of each figure. Changing the reconstruction algorithm to PSF had a minor impact on the correlation matrix. However, the increased number of clusters being composed of features with mutual Spearman's correlation coefficients of < 0.7 demonstrates that the impact of the discretization method was much larger. Similarly, Fig. 2 shows that sphere size and activity uptake both had a major impact on the correlation matrix. The correlation matrices of the statistically equal replicates showed to be similar, and therefore all found differences in the composition of feature clusters could be ascribed to the sources of variation.

FIGURE 1 Impact of discretization and reconstruction setting on the composition of feature clusters. Feature clusters (red rectangles) were defined based on Spearman's correlation matrices. The default setting in the upper-left corner consists of all activity uptakes and sphere sizes, matrix size $256 \times 256 \times 111$, OSEM reconstruction, 120s scan duration, FBW discretization, 4 mm FWHM, and CT-based segmentation. The feature order of this setting was also used to display the correlation matrices of the other settings.

Reliability Analysis

Reliability analysis was not performed for 15 geometry features derived from the CT-based segmentation, as they are a function of sphere size and hence exhibit an ICC of 1 by definition. Fig. 3 and Fig. 4 display how the reliability of radiomic features is affected by activity uptake, sphere size, discretization method, image noise, reconstruction algorithm, and matrix/voxel size for CT-based segmentations. The impact of the same sources of variation for PET-based segmentations are displayed in Supplemental Fig. 1 and Supplemental Fig. 2.

Influence of activity uptake and sphere size

In general, larger spheres and hot spots resulted in a higher number of reliable features than smaller spheres and cold spots. For longer scan durations and features discretized with

FIGURE 2 Impact of sphere size and activity uptake on the composition of feature clusters. Feature clusters (red rectangles) were defined based on Spearman's correlation matrices. The default setting in the upper-left corner consists of the data of larger-hot spheres, EARL reconstruction, matrix size $256 \times 256 \times 111$, 120s scan duration, FBW discretization, and CT-based segmentation. The feature order of this setting was also used to display the correlation matrices of the other settings.

FBN, there was more variability across different sphere sizes. In the hot spots, the difference between sphere sizes was more pronounced than in the cold spots.

Influence of discretization method

Especially for the hot spheres, FBW discretization led to higher reliability and to less variation across different reconstruction algorithms compared to FBN discretization. The larger-cold spheres only yielded higher reliability when less than 6 mm smoothing was applied. For the smaller cold spheres, FBN discretization generally resulted in higher reliability.

FIGURE 3 Percentage of reliable features discretized with FBW. Percentage of all features discretized with FBW and segmented based on CT exhibiting an ICC > 0.8 for all studied settings and underlying data categories.

Influence of image noise

Overall, image noise reduction by means of increasing the scan duration or increasing the smoothing factor yielded higher feature reliability. This inverse proportional relationship is illustrated in Fig. 5, which shows the number of reliable features discretized with FBN for the scan with SBR 1:10 for all reconstruction methods, scan durations, smoothing factors, and matrix sizes as function of noise. However, smoothing of larger-cold spheres discretized with FBW yielded lower reliability which leveled off at 6 mm FWHM. For the smaller-cold spheres discretized with FBW, there was almost no difference across smoothing factors. In general, FBW discretization had a lower impact across smoothing factors and scan durations than FBN discretization.

FIGURE 4 Percentage of all features discretized with FBN. Percentage of all features discretized with FBN and segmented based on CT exhibiting an ICC > 0.8 for all studied settings and underlying data categories.

FIGURE 5 Influence of noise on number of reliable features. Number of reliable features in the different sub-categories as function of noise.1

Influence of image reconstruction

The reconstruction algorithm especially affected the reliability for longer (120s and 300s) scan durations. For those scan durations, the reliability was lowest for images reconstructed with OSEM and increased by adding PSF. The additional use of TOF had almost no effect when applied to OSEM reconstructions while the effect was more pronounced when applied to PSF reconstructions. For hot spots discretized with FBW, almost no reliability differences were found across different reconstruction algorithms, while the use of PSF yielded lower reliability in the cold spot data with less than 6 mm FWHM smoothing.

Influence of matrix/voxel size

For most sub-categories, the effects of matrix/voxel size were small. Only in the smaller-cold spots with 300s scan duration, the smaller voxel size resulted in higher reliability for both discretization methods. This effect was also present in larger-cold spheres discretized with FBW, when less than 6 mm smoothing was applied.

Influence of segmentation method

The influence of segmentation method was particularly observed in the cold spots. For the bigger-cold spheres discretized with FBN, PET-based segmentations resulted in higher feature reliability than CT-based segmentations. However, for FBW discretization, CT-based segmentations resulted in higher reliability. For the smaller cold spots delineated with the PET-based segmentation, the PSF reconstruction yielded lower reliability compared to the other reconstruction settings.

DISCUSSION

This study demonstrated that both dimensionality reduction and reliability of ^{18}F -FDG PET radiomic features are sensitive to most sources of variation. In the subsequent sections, the underlying trends are described in more detail.

As described in several other studies (9,20), we found that many features were highly correlated. Discretization, sphere size, and activity uptake had a major impact on this correlation, while reconstruction method had less influence. To reduce the feature space, representative features should be chosen from each cluster. We showed that the composition of the correlation matrices was repeatable, but dependent on various factors such as image discretization, activity uptake, and sphere size. As a consequence, these correlation matrices yield different clusters of correlated features. Therefore, the representative features extracted from these clusters will differ across these matrices. Hence, the outcome of redundancy analyses are only generalizable among studies when these studies applied similar settings.

FBN discretization led to less reliability with a higher variation across reconstruction algorithms for the hot spot data. A reason for this might be that for this setting the bin width is sensitive to image noise and therefore every image is discretized with a different bin width. For the cold spots, FBW discretization resulted in lower reliability. This might be explained by the fact that the small intensity range in the cold spots resulted in too few bins. Decreasing the noise by using PSF (21) or increasing the smoothing also decreased this range and led to lower reliability. Some studies support these implications of different intensity ranges and reported better clinical applicability and repeatability of FBW discretization (18,22,23), but

up to now, both discretization methods are being used. Our results suggest that for positive activity uptakes, FBW discretization is the recommended method of choice, while for negative activity uptakes FBN discretization is recommended or a different fixed bin width should be used. Further research should be executed to explore the different settings of both discretization methods.

In general, increased smoothing resulted in a higher number of reliable features. This may seem counter-intuitive as for smoothing there is a trade-off between noise and spatial resolution; i.e. increased smoothing leads to less noise but lower spatial resolution. Hence in terms of reliability, reducing image noise was found to be more important than losing spatial resolution and textural information of the data. This may be due to PET textural features, capturing intensity differences between neighboring voxels, can be highly sensitive to stochastic image variation (12). As the reduction of noise leads to more homogeneous image texture, this may lead to more comparable textural matrices across the statistically equal replicates and hence to higher reliability.

In this study, radiomic features extracted from larger-hot spheres generally showed higher reliability than those extracted from smaller-cold spheres. In a clinical setting, the tracer uptake activity is affected by the tumor type and uptake mechanism. Further, signal depends on the used PET isotope. This can result in images with a poor signal-to-noise ratio (e.g. ^{89}Zr -antibodies in immunoPET studies) or even in cold spots. Therefore, it is not recommended to generalize results of radiomic studies in different tumor types and PET tracers.

Previous studies have shown that the segmentation method has a large impact on feature values and their repeatability (23-25). In this study, small reliability differences were found between both segmentation methods for hot spheres. However, for FBW discretized cold spots, the PET-based segmented data showed lower reliability, which was possibly caused by the fact that the threshold-based PET-segmentation yielded smaller intensity range in the segmentation. For the FBN discretized cold spots, the higher reliability for the PET-based segmentation may be biased by unsatisfactory over-segmentations of the 41% SUV_{peak} segmentation method. This finding may thus rather be driven by the enlarged VOIs (which previously showed to provide more reliable results) than by the segmentation method itself. This might also explain that cold spheres reconstructed with PSF resulted in lower reliability. PSF reconstructions result in sharper edges between sphere and background and may therefore yield more accurate segmentation results. Moreover, the reliability for smaller-cold spheres decreased when the smoothing was increased. A reason for this might be that smoothing degraded the signal, and led more frequently to a failure in segmentation.

Several studies showed that a large number of features exhibit high variability across various reconstruction settings (13,15,24). Our study showed that for all study conditions

the number of reliable features also depended on the used reconstruction algorithm. E.g. images reconstructed with PSF or PSF+TOF yielded higher reliability than OSEM or TOF reconstructions. The higher reliability found by using PSF is consistent with the fact that PSF decreases image noise (21) and with the previously reported finding that image noise and reliability have an inverse proportional relationship. Moreover, the additional use of TOF improves image quality and reduces image noise (26) and is therefore expected to increase feature reliability. However, our results showed the same reliability for images reconstructed with OSEM and images reconstructed with TOF. A comparison of image noise between these images showed that the TOF-effect on our scanner and for this phantom was small and had therefore also only a small effect on the reliability of features.

A drawback of our study is that the phantom data only contains ideal spheres that show little ^{18}F -FDG heterogeneity pattern, and may therefore not be representative for patient data. The results of this study will therefore be extended to more realistic phantoms which can simulate tumor heterogeneity in future work (27,28). Furthermore, simulated PET images will be used to explore the reliability of radiomic features for more complex shapes. However, features showing low reliability for these simple phantom inserts would likely also provide unreliable results for more complex phantoms or patient data.

CONCLUSION

This study reports on the impact of underlying data, image reconstruction methods and settings, noise, discretization method, and delineation method on the dimensionality reduction and reliability of ^{18}F -FDG PET radiomic features, which is an important measurement of error. Our data show that feature reduction is sensitive to discretization, sphere size, and activity uptake, and is therefore only generalizable among studies using the same settings. This study demonstrates that clinical PET studies and examinations need to be standardized in order to use ^{18}F -FDG PET radiomics as quantitative imaging biomarkers. Although this conclusion is not new for standard quantitative PET biomarkers, our study suggests that, in particular for radiomics features, efforts should focus on noise reduction even at the cost of spatial resolution and optimizing the choice for image reconstruction method, discretization method, and segmentation method. For every type of PET study, a separate validation is recommended as the results depend on the underlying data (i.e. activity uptake and size).

ACKNOWLEDGEMENTS

This work is part of the research program STRaTeGy with project number 14929, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). This work was (in part) financially supported by the Netherlands Organisation for Health Research and Development [grant 10-10400-98-14002].

REFERENCES

1. Zhang H, Wroblewski K, Liao S, et al. Prognostic value of metabolic tumor burden from ^{18}F -FDG PET in surgical patients with non-small-cell lung cancer. *Acad Radiol*. 2013;20:32-40.
2. Huang W, Fan M, Liu B, et al. Value of metabolic tumor volume on repeated ^{18}F -FDG PET/CT for early prediction of survival in locally advanced non-small cell lung cancer treated with concurrent chemoradiotherapy. *J Nucl Med*. 2014;55:1584-1590.
3. Hatt M, Majdoub M, Vallières M, et al. ^{18}F -FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J Nucl Med*. 2015;56:38.
4. O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin Cancer Res*. 2015;21:249-257.
5. Zhang Y, Oikonomou A, Wong A, Haider MA, Khalvati F. Radiomics-based prognosis analysis for non-small cell lung cancer. *Sci Rep*. 2017;7:46349.
6. Aerts HJW, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
7. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol*. 2015;60:5471-5496.
8. Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present... any future? *Eur J Nucl Med Mol Imaging*. 2017;44:151-165.
9. Orlhac F, Soussan M, Maisonneuve J, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in ^{18}F -FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *J Nucl Med*. 2014;55:414-422.
10. Virendra K, Yuhua G, Satrajit B, et al. Radiomics: the process and the challenges. *Magn Reson Imaging*. 2012;30:1234-1248.
11. Cortes-Rodicio J, Sanchez-Merino G, Garcia-Fidalgo MA, Tobalina-Larrea I. Identification of low variability textural features for heterogeneity quantification of ^{18}F -FDG PET/CT imaging. *Rev Esp Med Nucl Imagen Mol*. 2016;35:379-384.
12. Nyflot MJ, Yang F, Byrd D, Bowen SR, Sandison GA, Kinahan PE. Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards. *J Med Imaging (Bellingham)*. 2015;2:041002.
13. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol*. 2010;49:1012-1016.
14. Yan J, Chu-Stern JL, Loi HY, et al. Impact of image reconstruction settings on texture features in ^{18}F -FDG PET. *J Nucl Med*. 2015;56:1667-1673.
15. Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A. The impact of image reconstruction settings on ^{18}F -FDG PET radiomic features: multi-scanner phantom and patient studies. *Eur Radiol*. 2017;27:4498-4509.
16. Bailly C, Bodet-Milin C, Couespel S, et al. Revisiting the Robustness of PET-Based Textural Features in the Context of Multi-Centric Trials. *PLoS One*. 2016;11:e0159984.

17. Boellaard R, Delgado-Bolton R, Oyen WJG, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging*. 2015;42:328-354.
18. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative - feature definitions. *CoRR*. 2016;abs/1612.07003.
19. Boellaard R, O'Doherty MJ, Weber WA, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur J Nucl Med Mol Imaging*. 2010;37:181-200.
20. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep*. 2015;5:13087.
21. Arman R, Jinyi Q, Vesna S. Resolution modeling in PET imaging: theory, practice, benefits, and pitfalls. *Med Phys*. 2013;40:064301.
22. Leijenaar RT, Nalbantov G, Carvalho S, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. *Sci Rep*. 2015;5:11075.
23. van Velden FH, Kramer GM, Frings V, et al. Repeatability of radiomic features in non-small-cell lung cancer [^{18}F]FDG-PET/CT studies: impact of reconstruction and delineation. *Mol Imaging Biol*. 2016;18:788-795.
24. Altazi BA, Zhang GG, Fernandez DC, et al. Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms. *J Appl Clin Med Phys*. 2017;18:32-48.
25. Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour ^{18}F -FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *Eur J Nucl Med Mol Imaging*. 2013;40:1662-1671.
26. Vandenberghe S, Mikhaylova E, D'Hoe E, Mollet P, Karp JS. Recent developments in time-of-flight PET. *EJNMMI Phys*. 2016;3:3.
27. Carles M, Torres-Espallardo I, Alberich-Bayarri A, et al. Evaluation of PET texture features with heterogeneous phantoms: complementarity and effect of motion and segmentation method. *Phys Med Biol*. 2017;62:652-668.
28. Berthon B, Marshall C, Holmes R, Spezi E. A novel phantom technique for evaluating the performance of PET auto-segmentation methods in delineating heterogeneous and irregular lesions. *EJNMMI Phys*. 2015;2:13.

Part II

Applications of ^{18}F -FDG PET/CT Radiomics in Oncology

Chapter 4

Predicting Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer with Textural Features Derived from Pretreatment ^{18}F -FDG PET/CT Imaging

Roelof J. Beukinga
Jan Binne Hulshoff
Lisanne V. van Dijk
Christina T. Muijs
Johannes G.M. Burgerhof
Gursah Kats-Ugurlu
Riemer H.J.A. Slart
Cornelis H. Slump
Véronique E.M. Mul
John Th.M. Plukker

J Nucl Med. 2017;58:723-729

ABBREVIATIONS

^{18}F -FDG PET	^{18}F -Fluorodeoxyglucose positron emission tomography
AIC	Akaike information criterion
AUC	area under the receiver operating characteristics curve
CROSS	chemoradiotherapy for oesophageal cancer followed by surgery study
CT	computed tomography
EC	esophageal cancer
GLRLM	grey level run-length matrix
LRLGLE-PET	long run low gray level emphasis measured on PET
nCRT	neoadjuvant chemoradiotherapy
RP-CT	run percentage measured on CT
SUV	standardized uptake value
TRG	tumor regression grade
VOI	volume of interest

ABSTRACT

Background: Adequate prediction of tumor response to neoadjuvant chemoradiotherapy (nCRT) in esophageal cancer (EC) patients is important in a more personalized treatment. The current best clinical method to predict pathologic complete response is SUV_{max} in ^{18}F -FDG PET/CT imaging. To improve the prediction of response, we constructed a model to predict complete response to nCRT in EC based on pretreatment clinical parameters and ^{18}F -FDG PET/CT-derived textural features.

Methods: From a prospectively maintained single-institution database, we reviewed 97 consecutive patients with locally advanced EC and a pretreatment ^{18}F -FDG PET/CT scan between 2009 and 2015. All patients were treated with nCRT (carboplatin/paclitaxel/41.4 Gy) followed by esophagectomy. We analyzed clinical, geometric, and pretreatment textural features extracted from both ^{18}F -FDG PET and CT. The current most accurate prediction model with SUV_{max} as a predictor variable was compared with 6 different response prediction models constructed using least absolute shrinkage and selection operator regularized logistic regression. Internal validation was performed to estimate the model's performances. Pathologic response was defined as complete versus incomplete response (Mandard tumor regression grade system 1 vs. 2–5).

Results: Pathologic examination revealed 19 (19.6%) complete and 78 (80.4%) incomplete responders. Least absolute shrinkage and selection operator regularization selected the clinical parameters: histologic type and clinical T stage, the ^{18}F -FDG PET-derived textural feature long run low gray level emphasis, and the CT-derived textural feature run percentage. Introducing these variables to a logistic regression analysis showed areas under the receiver-operating-characteristic curve (AUCs) of 0.78 compared with 0.58 in the SUV_{max} model. The discrimination slopes were 0.17 compared with 0.01, respectively. After internal validation, the AUCs decreased to 0.74 and 0.54, respectively.

Conclusion: The predictive values of the constructed models were superior to the standard method (SUV_{max}). These results can be considered as an initial step in predicting tumor response to nCRT in locally advanced EC. Further research in refining the predictive value of these models is needed to justify omission of surgery.

Key Words: textural analysis; response prediction; esophageal cancer; ^{18}F -FDG PET/CT

INTRODUCTION

Esophageal cancer (EC) is one of the most aggressive tumors, with early recurrences even after radical surgery. The standard treatment in locally advanced ($T_1/N_{1-3}/M_0$ and $T_{2-4a}/N_{0-3}/M_0$) resectable EC is neoadjuvant chemoradiotherapy (nCRT) followed by a radical esophagectomy. In the Dutch CROSS (ChemoRadiotherapy for Oesophageal cancer followed by Surgery Study) trial, nCRT improved the 5-y overall survival rate from 34% to 47% (1). Not all patients benefit from nCRT; 29% of the patients in the CROSS trial had a complete response, 52% had a partial response, and even 18% had no tumor response (1). For complete responders, surgical intervention might not be beneficial and a wait-and-see policy might suffice. Hence, adequate response prediction is important in developing personalized treatment in EC. Moreover, accurate response prediction may be relevant in patient counseling in future clinical trial strategies based on personalized treatment. So far, response prediction showed only promising results with functional imaging of tumor viability with ^{18}F -FDG PET and recently with diffusion-weighted MRI (2–4). Traditional image-derived indices used in PET rely on quantification of lesion SUVs and overall tumor volume, which have been shown to be important factors for patient outcome and treatment response (5,6). Although useful, these parameters omit available information related to the spatial distribution and specific features regarding intratumor radiotracer accumulation. This may limit the possibility to further characterize the biologic behavior of the tumor, based on hypoxia-induced heterogeneity and genomic instability. Intratumoral heterogeneity is correlated with aggressive tumor behavior and a decreased response due to expression of specific receptors with high cellular proliferation and angiogenesis (7–9). Hence, even small tumor biopsies lack complete molecular characterization due to spatial heterogeneity. A novel approach is to quantify spatial heterogeneity of metabolism and tissue density characterized by ^{18}F -FDG uptake and Hounsfield units with textural features. The concept of textural analysis is based on the spatial arrangement of voxels in a predefined volume of interest (VOI). This spatial intratumoral heterogeneity can be depicted from different spatial interrelationships on ^{18}F -FDG PET/CT scans. Therefore, ^{18}F -FDG PET/CT textural features have been proposed to be valuable in response prediction (10–15). The aim of this study was to develop a model to predict complete response to nCRT in locally advanced EC based on pretreatment clinical predictors and ^{18}F -FDG PET/CT-derived textural features.

MATERIALS AND METHODS

Patients

In this retrospective study, potentially curatively resectable EC patients were consecutively selected who underwent nCRT followed by esophagectomy between December 2009 and March 2016. Patients with fewer than 4 courses of chemotherapy, with missing ^{18}F -FDG PET/CT or with incomplete medical records, were excluded, yielding a total of 97 patients. In line with the rules of the Dutch National Health Sciences, our Institutional Review Board approved this retrospective study, and the requirement to obtain informed consent was waived.

Data were obtained from a prospectively maintained single-institution database including patient characteristics, tumor- and treatment-related data, and follow-up data. All patients were clinically staged with esophagoscopy and biopsy, endoscopic ultrasonography with fine-needle aspiration if indicated, and whole-body integrated ^{18}F -FDG PET/CT. Patients were staged according to the seventh edition of the TNM system maintained by the American Joint Committee on Cancer (16) and discussed in the hospital's multidisciplinary esophageal tumor board.

Imaging

PET/CT imaging was performed with an integrated ^{18}F -FDG PET/CT system (Biograph mCT 4–64 PET/CT; Siemens). Patients fasted for at least 6 h before PET/CT, with no restrictions on drinking water. Serum glucose levels were measured before ^{18}F -FDG administration with a weight-based dose of 3 MBq/kg. Sixty minutes after tracer injection, patients were scanned in treatment position. An inspiration breath-hold low-dose CT for attenuation correction was performed, and PET acquisitions were obtained in the caudal–cranial direction with a field of view of $500 \times 500 \times 500$ mm, 3-dimensional setting, 2–3 min per bed position, matrices of 512×512 (0.98×0.98 mm pixel size), and 2-mm slice thickness. Image data were reconstructed according to guidelines of the European Association of Nuclear Medicine (17).

Radiotherapy treatment planning including target volume delineation and CT texture analysis was performed on a 16- or 64-multidetector row spiral CT machine (Somatom Sensation 16 or 64; Siemens Medical Systems). A CT thorax/abdomen scan was obtained in the cranial–caudal direction with matrices of 512×512 (0.98×0.98 mm pixel size) and a 3-mm slice thickness.

Treatment and Pathology

On the basis of the experiences and the good results of the CROSS study, in which our institute had participated, our multidisciplinary tumor board decided to continue

nCRT according to the CROSS schedule. This treatment consisted of weekly intravenously administered paclitaxel (50 mg/m²) and carboplatin (areas under the receiver-operating-characteristic curve [AUC], 2 mg·min·mL⁻¹) during 5 wk with concurrent external radiotherapy (41.4 Gy in 23 fractions, 5 d per wk) (1). Transthoracic esophagectomy with 2-field lymphadenectomy was performed within 6–8 wk after completion of nCRT. The resected specimens were examined according to a standard protocol (18). Resection margins were defined according to the definitions of the College of American Pathologists as microscopic tumor-free (R0: > 0 mm) or tumor-positive (R1). Pathologic response was assessed by 2 expert gastrointestinal pathologists according to the Mandard tumor regression grade (TRG) (19), ranging from complete response (TRG 1) without viable tumor cells left, to partial response (TRG 2–4) with viable tumor cells left, to no response at all (TRG 5).

VOI

Textural analysis was performed on a VOI incorporating the gross tumor volume for radiation treatment planning. Tumor delineation was performed manually with consensus between 3 experienced radiation oncologists on axial planes of the radiotherapy planning CT, to enclose 3-dimensional coverage of the entire tumor. Involved lymph nodes were not

FIGURE 1 A) Manual delineation of the gross tumor volume on planning CT; B) Co-registration of LD-CT (lava colormap) to the radiotherapy planning CT (grayscale colormap); C) Overlay of PET image (lava colormap) onto radiotherapy planning CT (grayscale colormap); D) Cropping of PET VOI; E) Feature extraction; I) Assessment of tumor shape by means of geometry features; II) Global assessment of tonal distribution by means of first order textural features; III) Assessment of pairwise arrangement of voxels by means of grey level co-occurrence matrix; IV) Assessment of alignment of voxels with the same intensity by means of grey level run-length matrix; V) Assessment of characteristics of homogenous zones by means of grey level size-zone matrix.

included into the VOI, because these lesions are too small ($<10 \text{ cm}^3$) for reliable textural analysis (20). The gross tumor volume was rigidly registered to the ^{18}F -FDG PET/CT data series (RTx Workstation 1.0; Mirada Medical). Erroneous registrations were manually adjusted after consensus of the collaborating investigators.

Tonal Discretization

^{18}F -FDG PET/CT imaging data and VOI delineations were loaded into Matlab 2014b (MathWorks; an interactive image processing environment) for processing and analyses. The SUV, for semiquantitative analysis of metabolism, was corrected for individual variations in serum glucose level and was discretized to reduce the continuous scale to a finite set of values and to reduce noise throughout the entire study in increments of 0.5 g/mL according to Doane's optimal bin width (21). Similarly, the Hounsfield unit scale for quantitative analysis of tumor density was discretized in increments of 30 Hounsfield units for textural analysis.

Candidate Predictors

For each patient, a total of 88 parameters were evaluated, including 7 clinical parameters; 16 geometry features; the glycolytic volume based on tumor volume and SUVmean; and 19 first-order, 24 second-order, and 22 higher order textural features extracted from ^{18}F -FDG PET and CT (supplemental materials [available at <http://jnm.snmjournals.org>]). First-order textural features are statistics based on the gray level distribution of the image but do not consider relative positions of gray levels. Second- and higher order textural features do consider relative positions of gray levels and therefore allow quantification of heterogeneity. For various spatial interrelationships, frequency distributions (Fig. 1) were obtained—that is, the gray level co-occurrence (spatial dependence) matrix for pairwise arrangement of voxels (extracted with a pixel-to-pixel distance equal to 1) (22), the gray level run-length matrix for alignment of voxels with the same intensity (23), and the gray level size-zone matrix for characteristics of homogeneous zones (24). Directional voxel analysis was performed in 3 dimensions with a connectivity of 26 voxels and analysis in 13 angular directions. All second- and higher order textural features are weighted averages of these matrices to express the relative importance of their properties. All extracted textural features were normalized to the range $[0,1]$.

Statistical Analysis

Statistical analysis was performed with R 3.2.2 open-source software using the glmnet package (version 2.0–2) and the rms package (version 4.4–0), available from the Comprehensive R Archive Network (<http://www.r-project.org>).

Because textural feature values may be subject to interobserver variability in the delineation of the tumor, the original delineations were uniformly eroded by ball-shaped structuring elements with radii of 1 and 2 voxels. For each delineation, textural features were extracted and the stability of each feature was evaluated with the intraclass correlation. Only stable features (intraclass correlation > 0.7) were considered for further analysis. Predictors were then selected by a univariable logistic model with a response variable labeling complete (Mandard TRG 1) and incomplete response (Mandard TRG 2–5). All potential predictors that met the Akaike information criterion (AIC) were considered significant. To discourage

TABLE 1
Patient Characteristics

Characteristic	<i>n</i>	%
Sex		
Male	82	84.5
Female	15	15.5
Age		
<70 y	78	80.4
≥ 70 y	19	19.6
Histology		
AC	88	90.7
SCC	9	9.3
Tumor grade		
Missing	6	6.2
G1	49	50.5
G3	42	43.3
EUS tumor length		
<5 cm	37	40.2
≥5 cm	58	59.8
Localization		
Mid	4	4.1
Distal	62	63.9
GEJ	31	32.0
nCRT cycles		
4	18	92.8
5	79	7.2
Circumferential resection margin*		
R0	90	99.0
R1	7	1.0
Mandard TRG		
1	19	19.6
2	23	23.7
3	37	38.1
4	15	15.5
5	3	3.1

* Tumor-free (R0) resection margin defined according to criteria of the CAP as >0mm.

Abbreviations: AC: adenocarcinoma; SCC: squamous cell carcinoma; EUS: endoscopic ultrasonography; GEJ: gastroesophageal junction; nCRT: neoadjuvant chemoradiotherapy; TRG: tumor regression grade; cT/N: clinical tumor/nodal stage; ypT/N: pathologic tumor/nodal stage after nCRT; and CAP: College of American Pathologists.

overfitting, the AIC is based on rewarding goodness of fit and penalizing the complexity of the model. The AIC requires $\chi^2 > 2 \cdot \text{df}$ —that is, when considering a predictor with one degree of freedom df, this implies an $\alpha = P(\chi^2 \geq 2) = 0.157$ (25).

Significant predictors were used to construct 6 multivariable logistic regression models for comparison with current most accurate prediction model with $\text{SUV}_{\max}$ as predictor variable (model 1). These models were constructed by introducing clinical parameters (model 2); clinical parameters and geometry features (model 3); clinical parameters, geometry features, and PET textural features (model 4); clinical parameters, geometry features, and CT textural features (model 5); and clinical parameters, geometry features, and PET/CT textural features (model 6) to a least absolute shrinkage and selection operator, a technique for L1-norm regularization. By increasing the shrinkage parameter λ , the regularization shrinks the estimated coefficients and excludes variables when they become zero. The λ -value that minimized the 10-fold cross-validated mean squared error was repeatedly determined with 100 repetitions. The optimal λ -value was robustly determined by averaging over these obtained λ -values. The selected variables were fitted to the data with a logistic regression.

The model's calibration was evaluated using visual inspection of calibration plots and the Hosmer–Lemeshow test. The model's performance was quantified in terms of discrimination with the AUC and the discrimination slope. The goodness of fit was evaluated with the -2

TABLE 2

Staging According to Seventh Edition of TNM System Maintained by American Joint Committee on Cancer (16)

Characteristic	n	%
cT stage		
cT1	2	2.1
cT2	16	16.5
cT3	74	76.3
cT4a	5	5.2
cN stage		
cN0	21	21.7
cN1	45	46.4
cN2	28	28.9
cN3	3	3.1
ypT stage		
ypT0	19	19.6
ypT1	13	13.4
ypT2	13	13.4
ypT3	52	53.6
ypN stage		
ypN0	63	65.0
ypN1	18	18.6
ypN2	10	10.3
ypN3	6	6.2

Abbreviations: cT/N = clinical tumor/nodal stage; ypT/N = pathologic tumor/nodal stage after nCRT.

TABLE 3

Estimated Regression Coefficients of Prediction Models for Pathologic Complete Response Without Optimism Correction

Variable	Model 1			Model 2 – 3			Model 4			Model 5			Model 6		
	Coef	S.E.	P	Coef	S.E.	P	Coef	S.E.	P	Coef	S.E.	P	Coef	S.E.	P
Intercept	-0.88	0.50	0.08	-0.42	0.50	0.39	-0.86	0.65	0.19	-1.15	0.72	0.11	-1.83	0.91	0.04
SUV _{max}	-1.72	1.45	0.24												
Histology															
SCC				1.00			1.00			1.00			1.00		
AC				-1.70	0.61	0.01	-1.47	0.65	0.02	-1.89	0.64	0.00	-1.63	0.67	0.02
cT stage															
cT1 & cT2				1.00			1.00			1.00			1.00		
cT3 & cT4a				2.03	0.78	0.01	1.98	0.80	0.01	2.27	0.81	0.00	2.23	0.83	0.01
LRLGLe-PET							0.56	0.55	0.31				0.71	0.59	0.23
RP-CT										0.01	0.01	0.15	0.02	0.01	0.10

Abbreviations: Coef: Coefficient; LRLGLe-PET: long run low gray level emphasis measured on PET; and RP-CT: run percentage measured on CT.

loglikelihood and the Nagelkerke R^2 . The model was internally validated by a bootstrap approach with 2,000 repetitions. Bootstrapping allowed for obtaining the optimism-corrected measures for model performance and for shrinkage of the estimated regression coefficients using the optimism-corrected slope.

RESULTS

Patients and Treatment

Patients characteristics are shown in Tables 1 and 2. Seventy-nine patients (81.4%) received the complete nCRT regimen (all patients received the full radiotherapy dose). Resection with curative intent was performed within a mean time of 56 (SD, 14) d after completion of nCRT. R0 resection was achieved in 90 (92.8%) patients and R1 resection in 7 (7.2%) patients, all with positive circumferential resection margins and one with a positive proximal resection margin. Pathologic findings revealed complete response in 19 patients (19.6%) and incomplete response in 78 patients (80.4%).

Model Development

For the preselection, 144 of the 147 (97.3%) variables were found to be robust for contour variations. These variables were introduced to univariable logistic regression analysis, resulting in 24 significant variables predictive for response, including 4 clinical parameters; 0 geometry features; 1 first-order, 8 second-order, and 5 higher order PET textural features; and 1 first-order, 1 second-order, and 4 higher order CT textural features. All constructed

TABLE 4
Estimates of Model Performance for the Prediction Models

	Goodness-of-fit			Discrimination		Calibration			Validation	
	-2LLH	AIC	R ²	AUC	DS	Intercept	Slope	HLp	R ² boot	AUC boot
Model 1	94.46	98.46	0.02	0.58	0.01	2.09	2.46	0.75	0.00	0.54
Model 2 - 3	81.80	87.80	0.22	0.71	0.14	-0.04	0.94	1.00	0.17	0.70
Model 4	80.59	88.59	0.23	0.71	0.15	-0.15	0.87	0.45	0.17	0.69
Model 5	79.82	87.82	0.24	0.79	0.16	-0.14	0.86	0.42	0.18	0.76
Model 6	78.03	88.03	0.27	0.78	0.17	-0.22	0.81	0.46	0.18	0.74

Abbreviations: -2LLH = -2 loglikelihood; R² = Nagelkerke R²; DS = discrimination slope; HLp = Hosmer-Lemeshow P value; boot = internal validated with bootstrapping.

prediction models performed significantly better than model 1 (based on SUV_{max}). The introduction of only significant clinical parameters to the least absolute shrinkage and selection operator regularization process resulted in the selection of histologic type and clinical T-stage (model 2). These variables were selected in each subsequently constructed model. Compared with model 1, the AUC improved from 0.58 to 0.71, the discrimination slope improved from 0.01 to 0.14, and the AIC decreased ($\Delta AIC = 10.66$). For model 3, no additional variables were selected compared with model 2, because no geometry features were significant at the univariable logistic regression analysis. For model 4, the PET textural feature long run low gray level emphasis (LRLGLE-PET) was selected. Adding this variable did slightly improve the discrimination and the likelihood compared with models 2 and 3, but resulted in a higher AIC ($\Delta AIC = -0.79$). After internal validation, the AUC was equal to 0.69. For model 5, the CT textural feature run percentage (RP-CT) was selected. Although the AIC was almost equal compared with models 2 and 3 ($\Delta AIC = -0.02$), adding this variable improved the discrimination slope to 0.16 and the AUC remarkably to 0.79. This also persisted after internal validation (AUC = 0.76). Finally, entering all variables to the modeling process resulted in the selection of all above-mentioned variables (model 6). Model 6 had the best goodness of fit, but not the lowest AIC ($\Delta AIC = -0.23, 0.56$, and -0.21 compared with models 2-3, 4, and 5, respectively). The AUC slightly decreased to 0.78, whereas the discrimination slope was increased to 0.17. After internal validation, the AUC decreased to 0.74. The model regression coefficients and the corresponding model performance measures are shown in Tables 3 and 4, respectively. Fig. 2 gives the values of the selected textural features and their corresponding frequency distributions for a complete and a nonresponder. For the selected textural features, the range of values to reproduce the normalization process and the found intraclass correlations for quantifying contouring robustness are given in the supplemental materials.

DISCUSSION

An adequate method to predict pathologic complete response after nCRT has not yet been defined in EC patients. In personalized treatment, accurate response prediction will lead to a paradigm shift with omission of surgical treatment in complete responders or prevention of unnecessary nCRT in nonresponders. Response evaluation of nCRT is commonly based on tumor metabolic response measured by SUV_{max} with ^{18}F -FDG PET, but with a low sensitivity and specificity of 67% and 68%, respectively (26). The current study is, to our knowledge, the first in predicting complete response with ^{18}F -FDG PET/CT-derived textural features in a homogeneous group of EC patients treated according to the CROSS regimen. We demonstrated that all constructed prediction models showed significant improvement

FIGURE 2 Example of the values of the selected textural features and their corresponding frequency distributions for a complete and a non-responder.

Abbreviations: GLRLM: gray level run-length matrix; LRLGLE-PET: long run low gray level emphasis measured on PET; and RP-CT: run percentage measured on CT.

compared with predictions based on SUV_{max} alone and may therefore be considered as an initial step in predicting response.

In this study, the most predictive textural features were LRLGLE-PET and RP-CT. LRLGLE-PET depends on long runs (coarse texture) with low gray levels and was higher (i.e., low and homogeneous ^{18}F -FDG uptake) for complete responders and lower (i.e., high and heterogeneous ^{18}F -FDG uptake) for incomplete responders, possibly due to tumor hypoxia and necrosis. RP-CT measures the homogeneity of runs (fine texture) and was higher in complete responders. In univariable logistic regression, high LRLGLE-PET and RP-CT values were associated with squamous cell carcinoma ($P = 0.12$ and 0.13 , respectively), confirming the higher complete response rates in squamous cell carcinoma (1).

The clinical value of SUV_{max} was limited, possibly because it is extracted from a single voxel and does not characterize the total ^{18}F -FDG uptake. This causes a high dependency on the quality of the PET images (including noise) and the voxel size, which induces a low reproducibility.

Several studies focused on response prediction in EC using ^{18}F -FDG PET/CT-derived textural features (Table 5). Van Rossum et al. concluded that ^{18}F -FDG PET-derived textural features provide statistical value (14), but this does not translate into a clinically relevant benefit, which is in line with our findings. Van Rossum et al. performed only ^{18}F -FDG PET textural analysis, whereas this study demonstrates the additional value of CT textural analysis. Other studies demonstrated promising findings but are hampered by several limitations, including small patient cohorts with heterogeneous treatment schedules, lack of multivariable analyses, and a substantial chance of model overfitting due to the lack of optimism correction (11–13,15).

A limitation of this study is the absence of external validation, which is essential for implementation into clinical practice. Moreover, the following factors that affect textural analysis should be considered. Changing the bin width influences the quantization noise and has a crucial effect on textural features (27). Although only an indication, we used Doane's optimal bin width to discretize the SUVs and Hounsfield units (21). Moreover, respiratory gated PET/CT acquisitions could be considered to reduce respiration-induced smearing and contrast degradation (28).

The constructed prediction model may serve as a basic model, which can be extended with new features for usage for other applications. The current constructed model might be helpful toward a safe decision in postponing a burdensome surgical procedure in patients with a doubtful adequate physical condition after nCRT. In patients treated with definitive chemoradiotherapy, detection of nonresponders might allow additional treatments when

TABLE 5
Current Literature Describing ^{18}F -FDG PET Texture Analysis in Response Prediction in EC

Study	<i>n</i>	Type	nCRT	Timing PET/CT	Outcome	Reported entered variables
Tixier et al. (13)	41	AC, SCC	60 Gy + cisplatin or carboplatin/fluorouracil	Pre-nCRT	CR vs. PR vs. NR (32)	Pre-local homogeneity, pre-local entropy, pre-coarseness, pre-intensity variability, pre-size-zone features
Hatt et al (11)	50	AC, SCC	60 Gy + cisplatin/fluorouracil	Pre-nCRT	CR + PR vs. NR (32)	Pre-MATV, pre-entropy, pre-homogeneity, pre-dissimilarity, pre-intensity variability, pre-zone percentage
Tan et al (12)	20	AC, SCC	50.4 Gy + cisplatin/fluorouracil	Pre- and post-nCRT	TRG 1 + 2 vs. 3-5 (19)	$\Delta\text{SUV}_{\text{max}}$, SUV_{max} ratio, $\Delta\text{SUV}_{\text{mean}}$, pre-skewness, post-inertia, post-correlation, post-cluster prominence
Zhang et al (15)	20	AC, SCC	50.4 Gy + cisplatin or carboplatin	Pre- and post-nCRT	TRG 1 + 2 vs. 3-5 (19)	Post-orientation, tumor involves GEJ, $\Delta\text{Inertia}$, post-energy, post-entropy, $\Delta\text{skewness}$
Van Rossum et al (14)	217	AC	45 or 50.5 Gy + fluoropyrimidine with either a platinum compound or a taxane	Pre- and post-nCRT	TRG 1 vs. 2-4 (33)	EUS tumor length, cT stage, induction chemotherapy, post-nCRT endoscopic biopsy, subjective PET assessment, post-nCRT TLG, pre-cluster shade, Δrun percentage, ΔGLCM entropy, post-nCRT roundness
Current study	97	AC, SCC	41.4 Gy + carboplatin/paclitaxel	Pre-nCRT	TRG 1 vs. 2-5 (19)	Histology, cT stage, pre-LRLGLE-PET, pre-RP-CT

Abbreviations: AC = adenocarcinoma; SCC = squamous cell carcinoma; CR = complete response; PR = partial response; NR = no response; MATV = metabolically active tumor volume; GEJ = gastroesophageal junction; EUS = endoscopic ultrasonography; cT = clinical T stage; TLG = total lesion glycolysis; GLCM = gray level co-occurrence matrix.

available, whereas in complete responders an adjusted follow-up might be justified, to identify candidates for salvage surgery.

Up to now, the authors do not consider the predictive value of the constructed model high enough to justify the omission of surgery after nCRT in EC. A potential approach to improve the constructed basic prediction models could be (a) adding interim- or posttreatment textural analysis. Studies investigating both pre- and posttreatment textural analysis mainly reported posttreatment textural features to be associated with response (12,14,15). We performed a posttreatment textural analysis in patients with a posttreatment PET/CT scan ($n = 20$) and found 21 significant textural features for response in univariable regression analysis. However, posttreatment textural analysis suffers from radiation-induced esophagitis, which complicates delineation of the primary tumor and creates difficulties with tumor delineation in complete responders. (b) Texture could be characterized with more specific PET tracers such as ^{18}F -fluoroerythronitroimidazole (quantifying hypoxia) (29) or ^{18}F -fluorothymidine (targeting cellular proliferation) (30) or by other functional imaging modalities including the apparent diffusion coefficient in diffusion-weighted MRI (4). (c) Moreover, biologic markers

have shown to be potential molecular markers in individualizing EC treatment and may be incorporated to improve prediction models (31).

CONCLUSION

The constructed models are a valuable initial step in predicting response to nCRT in locally advanced EC. Adding the ^{18}F -FDG PET-derived textural feature LRLGLE-PET and RP-CT to a model with the clinical parameter histologic type and clinical T-stage is potentially predictive and was more accurate than response prediction based on SUV_{max} . These models may serve as basic models in determining clinical complete responders and can be extended with new features for usage for other applications.

REFERENCES

1. van Hagen P, Hulshof MC, van Lanschot JJ, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. *N Engl J Med*. 2012;366:2074–2084.
2. Chen YM, Pan XF, Tong LJ, Shi YP, Chen T. Can ^{18}F -fluorodeoxyglucose positron emission tomography predict responses to neoadjuvant therapy in oesophageal cancer patients? A meta-analysis. *Nucl Med Commun*. 2011;32:1005–1010.
3. Ott K, Weber WA, Lordick F, et al. Metabolic imaging predicts response, survival, and recurrence in adenocarcinomas of the esophagogastric junction. *J Clin Oncol*. 2006;24:4692–4698.
4. van Rossum PS, van Lier AL, van Vulpen M, et al. Diffusion-weighted magnetic resonance imaging for the prediction of pathologic response to neoadjuvant chemoradiotherapy in esophageal cancer. *Radiother Oncol*. 2015;115:163–170.
5. Boggs DH, Hanna A, Burrows W, Horiba N, Suntharalingam M. Primary gross tumor volume is an important prognostic factor in locally advanced esophageal cancer patients treated with trimodality therapy. *J Gastrointest Cancer*. 2015;46:131–137.
6. Pan L, Gu P, Huang G, Xue H, Wu S. Prognostic significance of SUV on PET/CT in patients with esophageal cancer: a systematic review and meta-analysis. *Eur J Gastroenterol Hepatol*. 2009;21:1008–1015.
7. Diaz-Cano SJ. Tumor heterogeneity: mechanisms and bases for a reliable application of molecular marker design. *Int J Mol Sci*. 2012;13:1951–2011.
8. Harris AL. Hypoxia: a key regulatory factor in tumour growth. *Nat Rev Cancer*. 2002;2:38–47.
9. Rajendran JG, Schwartz DL, O'Sullivan J, et al. Tumor hypoxia imaging with $[\text{F}-18]$ fluoromisonidazole positron emission tomography in head and neck cancer. *Clin Cancer Res*. 2006;12:5435–5441.
10. Chicklore S, Goh V, Siddique M, Roy A, Marsden PK, Cook GJ. Quantifying tumour heterogeneity in ^{18}F -FDG PET/CT imaging by texture analysis. *Eur J Nucl Med Mol Imaging*. 2013;40:133–140.
11. Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour ^{18}F -FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *Eur J Nucl Med Mol Imaging*. 2013;40:1662–1671.
12. Tan S, Kligerman S, Chen W, et al. Spatial-temporal $[\text{F}-18]$ FDG-PET features for predicting pathologic response of esophageal cancer to neoadjuvant chemoradiation therapy. *Int J Radiat Oncol Biol Phys*. 2013;85:1375–1382.
13. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline ^{18}F -FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med*. 2011;52:369–378.

14. van Rossum PS, Fried DV, Zhang L, et al. The incremental value of subjective and quantitative assessment of ^{18}F -FDG PET for the prediction of pathologic complete response to preoperative chemoradiotherapy in esophageal cancer. *J Nucl Med*. 2016;57:691–700.
15. Zhang H, Tan S, Chen W, et al. Modeling pathologic response of esophageal cancer to chemoradiation therapy using spatial-temporal ^{18}F -FDG PET features, clinical parameters, and demographics. *Int J Radiat Oncol Biol Phys*. 2014;88:195–203.
16. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th ed. of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol*. 2010;17:1471–1474.
17. Boellaard R, O'Doherty MJ, Weber WA, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur J Nucl Med Mol Imaging*. 2010;37:181–200.
18. Muijs C, Smit J, Karrenbeld A, et al. Residual tumor after neoadjuvant chemoradiation outside the radiation therapy target volume: a new prognostic factor for survival in esophageal cancer. *Int J Radiat Oncol Biol Phys*. 2014;88:845–852.
19. Mandard AM, Dalibard F, Mandard JC, et al. Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma: clinicopathologic correlations. *Cancer*. 1994;73:2680–2686.
20. Hatt M, Majdoub M, Vallieres M, et al. ^{18}F -FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J Nucl Med*. 2015;56:38–44.
21. Doane DP. Aesthetic frequency classifications. *Am Stat*. 1976;30:181–183.
22. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern*. 1973;3:610–621.
23. Galloway MM. Texture analysis using gray level run lengths. *Comput Graph Image Process*. 1975;4:172–179.
24. Thibault G, Angulo J, Meyer F. Advanced statistical matrices for texture characterization: application to cell classification. *IEEE Trans Biomed Eng*. 2014;61:630–637.
25. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr*. 1974;19:716–723.
26. Kwee RM. Prediction of tumor response to neoadjuvant therapy in patients with esophageal cancer with use of ^{18}F FDG PET: a systematic review. *Radiology*. 2010;254:707–717.
27. Leijenaar RT, Nalbantov G, Carvalho S, et al. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. *Sci Rep*. 2015;5:11075.
28. Sindoni A, Minutoli F, Pontoriero A, Iati G, Baldari S, Pergolizzi S. Usefulness of four dimensional (4D) PET/CT imaging in the evaluation of thoracic lesions and in radiotherapy planning: Review of the literature. *Lung Cancer*. 2016;96:78–86.
29. Yue J, Yang Y, Cabrera AR, et al. Measuring tumor hypoxia with ^{18}F -FETNIM PET in esophageal squamous cell carcinoma: a pilot clinical study. *Dis Esophagus*. 2012;25:54–61.
30. Yue J, Chen L, Cabrera AR, et al. Measuring tumor cell proliferation with ^{18}F -FLT PET during radiotherapy of esophageal squamous cell carcinoma: a pilot clinical study. *J Nucl Med*. 2010;51:528–534.
31. Honing J, Pavlov KV, Mul VE, et al. CD44, SHH and SOX2 as novel biomarkers in esophageal cancer patients treated with neoadjuvant chemoradiotherapy. *Radiother Oncol*. 2015;117:152–158.
32. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J Natl Cancer Inst*. 2000;92:205–216.
33. Chirieac LR, Swisher SG, Ajani JA, et al. Posttherapy pathologic stage predicts survival in patients with esophageal carcinoma receiving preoperative chemoradiation. *Cancer*. 2005;103:1347–1355.

Chapter 5

^{18}F -FDG PET Radiomics for Prediction of Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer

Roelof J. Beukinga
Jan Binne Hulshoff
Véronique E.M. Mul
Walter Noordzij
Gursah Kats-Ugurlu
Riemer H.J.A. Slart
John Th.M. Plukker

Radiology. 2018;287:983-992

ABBREVIATIONS

^{18}F -FDG PET	^{18}F -Fluorodeoxyglucose positron emission tomography
AUC	area under the receiver operating characteristics curve
CROSS	chemoradiotherapy for oesophageal cancer followed by surgery study
CT	computed tomography
LASSO	least absolute shrinkage and selection operator
nCRT	neoadjuvant chemoradiotherapy
SUV	standardized uptake value

ABSTRACT

Background: To assess the value of baseline and restaging ^{18}F -Fluorodeoxyglucose positron emission tomography (^{18}F -FDG PET) radiomics in predicting pathologic complete response to neoadjuvant chemoradiotherapy (nCRT) in patients with locally advanced esophageal cancer.

Methods: In this retrospective study, 73 patients with histologically confirmed $\text{T}_1/\text{N}_{1-3}/\text{M}_0$ or $\text{T}_{2-4a}/\text{N}_{0-3}/\text{M}_0$ esophageal cancer were treated with nCRT followed by surgery (CROSS regimen) between October 2014 and August 2017. Clinical variables and radiomic features from baseline and restaging ^{18}F -FDG PET were selected by univariable logistic regression and Least Absolute Shrinkage and Selection Operator (LASSO). The selected variables were used to fit a multivariable logistic regression model, which was internally validated using bootstrap resampling with 20,000 replicates. The performance of this model was compared with reference prediction models composed of SUV_{max} metrics, clinical variables, and pre-nCRT radiomic features. Outcome was defined as complete versus incomplete pathologic response (tumor regression grade 1 versus 2–5 according to the Mandard classification).

Results: Pathologic response was complete in 16 (21.9%) patients and incomplete in 57 (78.1%) patients. A prediction model combining clinical T-stage and post-nCRT joint maximum (quantifying image orderliness) yielded an optimism-corrected area under the receiver operating characteristics curve (AUC) of 0.81. Post-nCRT joint maximum was replaceable with 5 other redundant post-nCRT radiomic features which provided equal model performance. All reference prediction models exhibited substantially lower discriminatory accuracy.

Conclusion: The combination of clinical T-staging and quantitative assessment of post-nCRT ^{18}F -FDG PET orderliness (joint maximum) provides high discriminatory accuracy in predicting pathologic complete response in patients with esophageal cancer.

Key Words: esophageal cancer, nCRT, response prediction, radiomics, PET

INTRODUCTION

Neoadjuvant chemoradiotherapy (nCRT) followed by esophagectomy is the common standard treatment of resectable locally advanced esophageal cancer (1,2). Pathologic complete response after nCRT is generally achieved in 25-42% of esophageal cancer patients and is accompanied with a lower rate of recurrence and longer survival (1,3-6). Some patients without gross residual tumor on restaging evaluation (i.e. clinical complete response) may benefit from a 'wait-and-see policy'. However, improving the identification of complete responders is still required in order to omit surgery after nCRT on an individualized manner. The current standard method to predict response to nCRT is the semi-quantitative measurement of temporal change in maximum standardized uptake value (SUV_{max}) between baseline (pre-nCRT) and restaging (post-nCRT) ^{18}F -Fluorodeoxyglucose positron emission tomography (^{18}F -FDG PET) (7), however this value still yields an insufficient sensitivity and specificity of 67% and 68%, respectively (8). One of the reasons for this inadequate performance may be that SUV_{max} is a single voxel representation which is susceptible to noise artifacts (9). Moreover, SUV_{max} ignores the intratumoral ^{18}F -FDG spatial distribution and does not represent the overall tumor burden. Intratumoral heterogeneity is present in nearly all esophageal tumors and it has been hypothesized that high intratumoral ^{18}F -FDG uptake heterogeneity prior to nCRT, usually due to hypoxia, is associated with an impaired tumor response to nCRT (10).

Radiomics extracts a large number of quantitative imaging features from medical images and may improve image interpretation by acquiring in-vivo tumor information. In earlier studies, ^{18}F -FDG PET radiomic features quantifying geometric, intensity, and textural (^{18}F -FDG spatial distribution) characteristics of tumors, have exhibited higher diagnostic accuracies than the SUV_{max} in predicting response in esophageal cancer patients (11-17). Acquiring both pre-nCRT and post-nCRT ^{18}F -FDG PET enables close follow-up of tumor response during treatment. Changes in radiomic features during treatment may reflect changes in intratumoral ^{18}F -FDG heterogeneity and tumor phenotype. Changes in radiomic features have shown promising results in predicting response in esophageal cancer patients (11-13,16,18).

The aim of this study was to assess the value of radiomic features extracted from baseline and restaging ^{18}F -FDG PET scans in predicting pathologic complete response to nCRT in patients with locally advanced EC.

MATERIALS AND METHODS

Patients

This retrospective study was conducted according to the national Dutch guidelines for retrospective studies and rules of the local Institutional Ethical Board and the need to obtain informed consent was waived. Patients were eligible for inclusion if they had locally advanced ($T_1N_{1-3}M_0$ or $T_{2-4a}N_{0-3}M_0$) esophageal cancer and were treated with nCRT (ChemoRadiotherapy for Oesophageal cancer followed by Surgery Study [CROSS] regimen) followed by esophagectomy with a two-field lymph node dissection in the University Medical Center Groningen between October 2014 and August 2017 (1). Excluded from the analyses were patients with missing data, those with ^{18}F -FDG PET/computed tomography (CT) scans made in other medical centers, those with non- ^{18}F -FDG-avid tumors, and those with distant metastases found prior to or during surgery. Overall, 73 patients (mean age, 64.4 years $\pm$ 8.3; range, 42–83 years) were consecutively included: 58 men (mean age, 64.0 years $\pm$ 7.8; range, 51–81 years) and 15 women (mean age, 65.7 years $\pm$ 9.9; range, 42–83 years). Clinical data were obtained from a prospectively maintained database (Table 1). Twenty-two of these 73 patients have been reported in an earlier published article (15). The present manuscript can be considered as a continuation of this prior article in which only radiomic features derived from baseline ^{18}F -FDG PET/CT scans were analyzed (15), whereas in this manuscript we analyzed radiomic features derived from both baseline and restaging ^{18}F -FDG PET scans.

Staging, Treatment, and Pathology

Staging consisted of 64-multidetector CT thorax/abdomen, ^{18}F -FDG PET/CT, and endoscopic ultrasonography with fine needle aspiration, if indicated. After staging, all patients were discussed in our multidisciplinary upper gastrointestinal tumor board. All patients were treated with nCRT according to the CROSS regimen consisting of carboplatin (2 mg·min·mL⁻¹) and paclitaxel (50 mg/m²) in 5 cycles combined with 41.4 Gy in 23 fractions. All patients were restaged with ^{18}F -FDG PET/CT approximately 6–8 weeks after nCRT. Surgery consisted of either open or minimally invasive transthoracic esophagectomy, combined with a two-field lymph node dissection. Two experienced gastrointestinal pathologists determined tumor response to nCRT according to the Mandard tumor regression grade, which was considered the gold standard (19). This five-point scoring system classifies the percentage of residual vital tumor cells and the degree of nCRT-induced fibrosis. Response was categorized into pathologic complete response (tumor regression grade 1) versus pathologic incomplete response (tumor regression grade 2–5). As the clinical relevance of categorizing response into pathologic good response (tumor regression grade 1–2) versus poor response (tumor regression grade 3–5) remains unclear, this distinction was not made.

¹⁸F-FDG PET/CT Imaging

Integrated baseline and restaging ¹⁸F-FDG PET/CT (Biograph mCT-64 PET/CT, Siemens, Knoxville, TN, USA) scans were performed. Patients were instructed to fast except for the consumption of water for at least 6 h prior to administration of 3 MBq/kg ¹⁸F-FDG. Serum glucose levels were evaluated just before tracer injection. Sixty minutes after tracer injection, continuous breathing low-dose CT (80–120 kV, 20–35 mAs, and 5 mm slice thickness) for visualization of anatomical structures, and PET images were acquired with the patient positioned in radiation treatment planning position. PET acquisitions were obtained

TABLE 1

Patient and Tumor Characteristics

Characteristics	(n = 73)
Sex	
Male	58 (79.5)
Female	15 (20.5)
Age, median (IQR), years	63.0 (10.0)
Histology	
Adenocarcinoma	65 (89.0)
Squamous cell carcinoma	8 (11.0)
Tumor location	
Middle esophagus	9 (12.3)
Distal esophagus	64 (87.7)
Tumor length, median (IQR), cm	6.0 (4.0)
Clinical T-stage	
T2	9 (12.3)
T3	59 (80.8)
T4a	5 (6.8)
Clinical N-stage	
N0	15 (20.5)
N1	32 (43.8)
N2	22 (30.1)
N3	4 (5.5)
Number of chemotherapy cycles	
2	1 (1.4)
3	1 (1.4)
4	12 (16.4)
5	59 (80.8)
Radicality	
R0	70 (95.9)
R1	3 (4.1)
Mandard TRG	
1	16 (21.9)
2	19 (26.0)
3	23 (31.5)
4	14 (19.2)
5	1 (1.4)

Abbreviations: IQR: interquartile range; GEJ: gastroesophageal junction; R0: microscopically radical resection; R1: microscopically irradiated resection/circumferential resection margin > 1 mm; and TRG: tumor regression grade.

with 2–3 min per bed position in 3D setting. Images were reconstructed according to the European Association of Nuclear Medicine guidelines (20), using a time of flight iterative reconstruction method (3 iterations, 21 subsets, and voxel-size $3.1819 \times 3.1819 \times 2$ mm) with point spread function correction. Images were corrected for random coincidences, scatter, and attenuation and were smoothed with a Gaussian filter of 6.5 mm in full width at half maximum.

Volume of Interest Delineation

Based on the radiotherapeutic gross tumor volume, which was manually delineated by an expert radiation oncologist in gastrointestinal malignancies, the volume of interest was defined. The gross tumor volume was rigidly co-registered to the CT component of the baseline and restaging PET/CT images (RTx Workstation 1.0, Mirada Medical, Oxford, UK). Because registration errors could occur, the volume of interest was manually corrected after consensus of the collaborating investigators. The post-nCRT delineation included the pre-nCRT localization of the primary tumor, and was adjusted manually to compensate for regression of the tumor size.

Set of Predictors and Radiomic Feature Extraction

We analyzed the clinical parameters sex, age, histology, tumor location, tumor length, clinical T-stage, and clinical N-stage. Moreover, we extracted a total of 113 radiomic features from both pre-nCRT and post-nCRT PET images. Software was developed in-house with Matlab 2014b for feature extraction and image processing (21). We extracted 19 morphological features, 2 local intensity features, 18 statistical features, and 62 textural features (25 gray level co-occurrence based features, 16 gray level run-length based features, 16 gray level size-zone based features, and 5 neighborhood gray tone difference based features). Among these radiomic features were the 5 traditional radiomic features volume, SUV_{max} , SUV_{peak} , SUV_{mean} , and total lesion glycolysis (tumor volume multiplied by SUV_{mean}). For each of these radiomic features, the relative difference between pre-nCRT and post-nCRT ^{18}F -FDG PET (Δ -nCRT) was calculated according to:

$$\Delta - \text{nCRT radiomic feature} = \frac{\text{postnCRT radiomic feature} - \text{prenCRT radiomic feature}}{\text{prenCRT radiomic feature}} \times 100\%$$

Furthermore, 8 bin-to-bin histogram-distances and 4 cross-bin histogram-distances were selected which quantified the perceived similarity in intensity distribution between the pre-nCRT and post-nCRT PET images.

^{18}F -FDG uptake was converted to SUV and was corrected for the serum glucose level (20). Original voxel dimensions were up-sampled to isotropic voxel-dimensions of $2 \times 2 \times 2$ mm using trilinear spline interpolation. Voxels, enclosed for $\geq 50\%$ coverage, were included to

the up-sampled volume of interest. Textural features were extracted from discretized image stacks to reduce the continuous-scaled SUV to a limited number of gray levels and to reduce image noise. Voxels were discretized in 0.5 g/mL increments starting at a minimum of 0 g/mL. Images were analyzed in 3 dimensions with a connectivity of 26 voxels (13 angular directions and a pixel-to-pixel distance of 1). The 13 different gray level co-occurrence matrices and gray level run-length matrices along each angular direction were merged into combined matrices before feature extraction.

TABLE 3
Estimated Regression Coefficients of Prediction Models M_{1-12}

Variable	M_7			M_8			M_9			M_{10}			M_{11}			M_{12}		
	Coef	S.E.	P	Coef	S.E.	P	Coef	S.E.	P	Coef	S.E.	P	Coef	S.E.	P	Coef	S.E.	P
Intercept	-1.28	0.28	0.00	-1.27	0.28	0.00	-1.27	0.29	0.00	0.85	0.84	0.31	0.88	0.85	0.30	0.90	0.87	0.30
Clinical T-stage																		
T2	1.00		<0.01	1.00		<0.01	1.00		<0.01	1.00		<0.01	1.00		<0.01	1.00		<0.01
T3 & T4a	-2.81	0.95		-2.78	0.90		-2.76	0.91		-2.70	0.92		-2.71	0.93		-2.92	0.94	
Joint maximum	0.83	0.38	0.03															
Median absolute deviation				-0.74	0.68	0.27												
Joint entropy							-0.68	0.44	0.12									
Sum entropy										-0.68	0.41	0.09						
Angular second moment													0.74	0.37	0.05			
Inverse variance																-0.71	0.38	0.06

All reported IBMs were measured at post-nCRT PET imaging.
Abbreviations: Coef = regression coefficient estimate; SE = standard error.

FIGURE 1 Correlation matrix of all radiomic features which positively (Spearman's $\rho > 0.9$) or negatively (Spearman's $\rho < -0.9$) correlated with post-nCRT joint maximum. All mentioned radiomic features were extracted from post-nCRT imaging.

Statistical analysis

Statistical analysis was performed with Matlab 2014b (Mathworks, Natick, MA, USA) and R 3.2.2 open-source software available from the Comprehensive R Archive Network (<http://www.r-project.org>). Radiomic features were tested for their robustness to delineation variations. Therefore, two additional segmentations were created by morphological dilation with two ball-shaped structuring elements with radii of 1 and 2 voxels. The reliability of ratings was measured by the intraclass correlation coefficient. Only radiomic features having an excellent reliability (intraclass correlation coefficient > 0.75) at both the baseline and the restaging measurements were considered robust. Furthermore, all predictors were tested in a univariable logistic regression with a significance level α of 0.157 according to the Akaike Information Criterion requiring $\chi^2 > 2 \cdot df$ with degrees of freedom df .

Only robust predictors and predictors significant on univariable logistic regression were introduced to a Least Absolute Shrinkage and Selection Operator (LASSO) for variable selection. LASSO shrinks estimated regression coefficients and excludes variables by forcing certain coefficients to become zero in order to reduce overfitting (Supplementary Appendix A). Logistic regression model were fitted with the selected variables.

FIGURE 2 Representative transaxial pre-nCRT and post-nCRT ^{18}F -FDG PET/CT images and corresponding values of 6 radiomic features of a pathologic complete responder (Mandard TRG 1) and a pathologic incomplete responder (Mandard TRG 5). At diagnosis, the pathologic complete responder had a $\text{T}_2\text{N}_0\text{M}_0$ adenocarcinoma of the distal esophagus and the pathologic incomplete responder had a $\text{T}_3\text{N}_0\text{M}_0$ adenocarcinoma of the distal esophagus. For the pathologic complete responder, joint maximum (JM) and angular second moment (ASM) initially showed high values which increased even more after nCRT, while median absolute deviation (MAD), joint entropy (JE), sum entropy (SE), and inverse variance (IV) initially showed low values which decreased even more after nCRT. An inverse trend applied for the pathologic incomplete responder.

The final model was tested for multicollinearity, as quantified by a variance inflation factor > 4 . The goodness-of-fit of each model was evaluated with the -2 log-likelihood, the Akaike Information Criterion, and the Nagelkerke R^2 . Each model was quantified in terms of discrimination with the area under the receiver operating characteristic curve (AUC) and the discrimination slope and quantified in terms of calibration using the slope and intercept of calibration plots and the Hosmer–Lemeshow test. Because the models are trained and tested on the same dataset, these performance measures may potentially be optimistic. To adjust for this optimism, the models were internally validated by bootstrap resampling with 20,000

TABLE 3

Estimates of Model Performance for the Prediction Models M_{1-12}

	Goodness-of-fit			Discrimination		Calibration			Validation	
	-2LLH	AIC	R ²	AUC	DS	Intercept	Slope	HLp	R ² boot	AUC boot
M_1	76.53	80.53	0.01	0.50	-0.01	-0.42	0.59	0.05	0.00	0.47
M_2	76.66	80.66	0.00	0.34	-0.01	-3.40	-1.59	#N/B	-0.06	0.31
M_3	73.83	77.83	0.06	0.43	0.03	0.93	2.08	0.17	0.01	0.41
M_4	61.52	65.52	0.29	0.70	0.20	-0.53	0.79	1.00	0.19	0.67
M_5	57.83	63.83	0.35	0.75	0.25	-0.19	0.86	1.00	0.30	0.74
M_6	52.69	60.69	0.43	0.73	0.32	-0.26	0.80	0.29	0.33	0.70
M_7	56.10	62.10	0.38	0.82	0.27	-0.17	0.87	0.78	0.33	0.81
M_8	59.88	65.88	0.32	0.78	0.22	-0.18	0.86	0.68	0.26	0.76
M_9	58.79	64.79	0.34	0.78	0.23	-0.17	0.88	0.92	0.29	0.77
M_{10}	58.34	64.34	0.34	0.78	0.24	-0.16	0.88	0.92	0.30	0.77
M_{11}	57.14	63.14	0.36	0.82	0.26	-0.17	0.87	0.77	0.32	0.81
M_{12}	57.58	63.58	0.36	0.81	0.25	-0.18	0.85	0.32	0.30	0.80

M_1 : pre-nCRT SUV_{max} ; M_2 : post-nCRT SUV_{max} ; M_3 : Δ -nCRT SUV_{max} ; M_4 : clinical T-stage; M_5 : clinical T-stage + histology; M_6 : clinical T-stage + histology + pre-nCRT long run low gray level emphasis; M_7 : clinical T-stage + post-nCRT joint maximum; M_8 : clinical T-stage + post-nCRT median absolute deviation; M_9 : clinical T-stage + post-nCRT joint entropy; M_{10} : clinical T-stage + post-nCRT sum entropy; M_{11} : clinical T-stage + post-nCRT angular second moment; and M_{12} : clinical T-stage + post-nCRT inverse variance.

Abbreviations: -2LLH: -2 log-likelihood; AIC: Akaike information criterion; R²: Nagelkerke R²; AUC: area under the receiver operating characteristic; DS: discrimination slope; Intrcpt: intercept; HLp: Hosmer–Lemeshow *P* value; R²_b: internal validated Nagelkerke R²; and AUC_b: internal validated AUC.

replicates, yielding optimism-corrected Nagelkerke R² and AUC. The constructed prediction model was compared on these terms with six reference prediction models. Based on an earlier published article (15), six reference prediction models were constructed composed of pre-nCRT SUV_{max} (M_1); post-nCRT SUV_{max} (M_2); Δ -nCRT SUV_{max} (M_3); clinical T-stage (M_4); clinical T-stage and histology (M_5); and clinical T-stage, histology, and pre-nCRT long run low gray level emphasis (M_6).

RESULTS

All patients received the full dose of radiotherapy and 72 patients (97.2%) received 4 or 5 cycles of chemotherapy (Table 1). All volumes of interest had a minimum volume of 10 cm³ as was required for textural analysis to provide valuable complementary information (22). The median overall tumor volume before treatment was 52.1 cm³ (interquartile range, 43.8 cm³) and decreased to 41.8 cm³ (interquartile range, 25.6 cm³) on the restaging PET scan. Of all patients, 16 (21.9%) had a pathologic complete response, while 57 (78.1%) patients had a pathologic incomplete response.

FIGURE 3 Calibration plots of M_{7-12} referring to the agreement between the predicted probability of pathologic complete response by M_{7-12} and the true (observed) probability of pathologic complete response.

Model Construction

Of all studied radiomic features, 86 were robust for delineation variations, including all traditional radiomic features (exhibiting intraclass correlation coefficients > 0.94). Of these robust radiomic features, 22 pre-nCRT radiomic features, 45 post-nCRT radiomic features, 34 Δ -nCRT radiomic features, and 11 histogram distances were significant on univariable logistic regression analysis (Supplementary Table 1). Of the studied clinical parameters, we found a significant association between pathologic complete response and histology ($P = 0.01$), clinical N-stage ($P < 0.01$), and clinical T-stage ($P = 0.10$). Of the traditional radiomic features, overall tumor volume was significant on both pre-nCRT ($P = 0.03$) and post-nCRT ($P = 0.03$) ^{18}F -FDG PET, while total lesion glycolysis was only significant at post-nCRT ^{18}F -FDG PET ($P = 0.01$). These variables were introduced to the LASSO regularization process. LASSO selected clinical T-stage and post-nCRT joint maximum, derived from the gray level co-occurrence matrix. The joint maximum is the probability corresponding to the most common gray level co-occurrence in the gray level co-occurrence matrix. Clinical T-stage and post-nCRT joint maximum were introduced to a logistic regression analysis

(M_7). Minimal effects of multicollinearity within the model were found, as quantified by variance inflation factors of 1.01 and 1.01 for the variables clinical T-stage and post-nCRT joint maximum, respectively.

Model Performance

The performance measures of M_{7-12} are shown in Table 3. M_7 exhibited the following goodness-of-fit metrics: a -2 log-likelihood of 56.10, an Akaike Information Criterion of 62.10, and a Nagelkerke R^2 of 0.32. The model had a good discriminatory accuracy with an AUC of 0.82 and a discrimination slope of 0.27. The model was rather good calibrated with an intercept of -0.17, a slope of 0.87, and a Hosmer–Lemeshow P value of 0.78 (Fig. 3). After internal validation, the Nagelkerke R^2 and AUC slightly decreased to 0.33 and 0.81, respectively. M_{1-5} all exhibited worse goodness-of-fit, discrimination, and calibration compared to M_7 . M_6 had a better goodness-of-fit than M_7 , however this was not reflected in a higher discriminatory accuracy and a better calibration. The model performances of M_{7-12} were relatively consistent, though M_7 did exhibit the best goodness-of-fit and discrimination, which also persisted after internal validation. Although M_{9-10} had a slightly better calibration than M_7 , the calibrations of M_{7-12} were also relatively consistent.

DISCUSSION

Clinical evaluation of response to nCRT is important for patients with potentially curable locally advanced esophageal cancer for exploring future personalized treatment and in predicting prognostic outcome. Recently, response has increasingly been defined on the basis of tumor biology rather than on tumor morphology using RECIST criteria (23,24). Image quantification with radiomics is an emerging noninvasive approach to predict survival outcome and response in cancer treatment. In the present study, we constructed a prediction model based on clinical evaluation of tumor invasion depth (T-stage) and post-nCRT joint maximum. Joint maximum is a measure of orderliness, quantifying the systematic arrangement of voxel intensity differences. Joint entropy and angular second moment are other radiomic features that are measures of image orderliness. High values of joint maximum and angular second moment occur in orderly images, while large joint entropy values indicate disorder. For each increase in joint maximum, the odds of pathologic complete response increases by a factor of $e^{\text{Coeff}_{\text{joint maximum}}} = e^{0.83} = 2.3$. In accordance with our expectations, this would mean that higher post-nCRT ^{18}F -FDG PET orderliness increases the probability for achieving pathologic complete response.

The final prediction model was superior compared to the reference prediction models composed of conventional SUV_{max} measurements (M_{1-3}) and clinical parameters (M_4).

5). Moreover, it demonstrated a higher discriminatory accuracy than the reference model consisting of clinical parameters along with a pre-nCRT radiomic feature (M_e), suggesting the higher potential of post-nCRT radiomic features in predicting pathologic complete response in esophageal cancer patients. This higher discriminatory accuracy was retained when post-nCRT joint maximum was exchanged by one of the correlated radiomic features. Predictions based on relative differences of radiomic features and histogram distances were clinically not relevant.

Numerous different radiomic features have been reported as promising over the years. To validate whether these radiomic features also provide complementary clinical value in external cohorts, pooling of those results is required. In a systematical MEDLINE database search, we found 10 relevant studies investigating the value of PET/CT radiomic features in the prediction of therapy response in esophageal cancer (11-18,25,26). The study by Van Rossum et al. was the only study which performed quantitative ^{18}F -FDG PET from pre-nCRT and post-nCRT ^{18}F -FDG PET, classified response as complete (absence of viable tumor tissue) versus incomplete (any grade of residual tumor tissue), and performed a univariable logistic regression analysis (11). Consistent with our results, Van Rossum et al. reported that the post-nCRT radiomic features joint maximum, joint entropy, sum entropy, and angular second moment were significant on univariable logistic regression analysis. However, inverse variance was reported as non-significant and median absolute deviation was not analyzed. They proposed four multivariable prediction models, of which the prediction model exhibiting the highest discriminatory accuracy consisted of 10 different clinical and (subjective and quantitative) ^{18}F -FDG PET parameters. Their model contains substantially more variables and hence is more complex, their model exhibited a slightly lower optimism-corrected AUC of 0.77 compared to 0.81 of the current proposed prediction model.

Though promising, the use of radiomics in selecting patients to different treatment strategies based on identification of response is still not practicable due to numerous unsolved problems and the complexity of data. There is a significant lack of standardization including the usage of different scanners and image acquisition protocols among hospitals and different feature extraction work-flows, which could lead to a large variability in radiomic features (27). Moreover, not all textural features are deemed stable enough with respect to the impact of different devices and delineation variation. The study by Hatt et al. found that robust radiomic features for delineation variation included: joint entropy, inverse difference, dissimilarity, and zone percentage (28), however they only assessed a limited number of radiomic features. Therefore, the current study tested the impact of delineation variation on all studied radiomic features by creating multiple artificial tumor delineations. Our results were generally consistent with the study by Hatt et al. Post-nCRT tumor delineation is in any case complicated in complete responders, due to the absence of viable tumor tissue. In

those patients, the baseline tumor delineation was registered with the restaging PET, and the original tumor dimensions were retained. Moreover, localization of the tumor on the restaging PET can be difficult as the metabolically active area is often contaminated with nCRT-induced esophagitis.

A shortcoming of the present study was the lack of a patient cohort for external validation. However, verification of the proposed prediction models requires large cohorts of homogeneously staged and treated patients. Another key factor which hampers clinical implementation of the proposed prediction model is that there is still limited information regarding the repeatability of the proposed radiomic features. In one of the few studies regarding this subject, Tixier et al. compared textural features of double baseline ^{18}F -FDG PET scans and found that joint entropy exhibited a high reproducibility, whereas angular second moment was characterized by lower reproducibility (29). Joint maximum, median absolute deviation, sum entropy, and inverse variance were not evaluated.

In summary, higher post-nCRT ^{18}F -FDG PET orderliness measured by joint maximum, increased the probability for achieving pathologic complete response after nCRT in patients with esophageal cancer. In this study, a prediction model composed of clinical T-stage and post-nCRT joint maximum seems to add important information to the visual PET/CT evaluation to guide eventual omission of surgery for patients achieving a pathologic complete response.

REFERENCES

1. van Hagen P, Hulshof MC, van Lanschot JJ, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. *N Engl J Med*. 2012;366:2074-2084.
2. Shapiro J, van Lanschot JJ, Hulshof MC, et al. Neoadjuvant chemoradiotherapy plus surgery versus surgery alone for oesophageal or junctional cancer (CROSS): long-term results of a randomised controlled trial. *Lancet Oncol*. 2015;16:1090-1098.
3. van Hagen P, Wijnhoven BP, Nafteux P, et al. Recurrence pattern in patients with a pathologically complete response after neoadjuvant chemoradiotherapy and surgery for oesophageal cancer. *Br J Surg*. 2013;100:267-273.
4. Meguid RA, Hooker CM, Taylor JT, et al. Recurrence after neoadjuvant chemoradiation and surgery for esophageal cancer: does the pattern of recurrence differ for patients with complete response and those with partial or no response? *J Thorac Cardiovasc Surg*. 2009;138:1309-1317.
5. Smit JK, Guler S, Beukema JC, et al. Different recurrence pattern after neoadjuvant chemoradiotherapy compared to surgery alone in esophageal cancer patients. *Ann Surg Oncol*. 2013;20:4008-4015.
6. Zaroni A, Verlato G, Giacopuzzi S, et al. Neoadjuvant concurrent chemoradiotherapy for locally advanced esophageal cancer in a single high-volume center. *Ann Surg Oncol*. 2013;20:1993-1999.
7. Lordick F, Ott K, Krause BJ, et al. PET to assess early metabolic response and to guide treatment of adenocarcinoma of the oesophagogastric junction: the MUNICON phase II trial. *Lancet Oncol*. 2007;8:797-805.
8. Kwee RM. Prediction of tumor response to neoadjuvant therapy in patients with esophageal cancer with use of ^{18}F FDG PET: a systematic review. *Radiology*. 2010;254:707-717.

9. Lodge MA, Chaudhry MA, Wahl RL. Noise considerations for PET quantification using maximum and peak standardized uptake value. *J Nucl Med.* 2012;53:1041-1047.
10. O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin Cancer Res.* 2015;21:249-257.
11. van Rossum PS, Fried DV, Zhang L, et al. The Incremental Value of Subjective and Quantitative Assessment of ^{18}F -FDG PET for the Prediction of Pathologic Complete Response to Preoperative Chemoradiotherapy in Esophageal Cancer. *J Nucl Med.* 2016;57:691-700.
12. Tan S, Kligerman S, Chen W, et al. Spatial-temporal [^{18}F]FDG-PET features for predicting pathologic response of esophageal cancer to neoadjuvant chemoradiation therapy. *Int J Radiat Oncol Biol Phys.* 2013;85:1375-1382.
13. Tan S, Zhang H, Zhang Y, Chen W, D'Souza WD, Lu W. Predicting pathologic tumor response to chemoradiotherapy with histogram distances characterizing longitudinal changes in ^{18}F -FDG uptake patterns. *Med Phys.* 2013;40:101707.
14. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline ^{18}F -FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med.* 2011;52:369-378.
15. Beukinga RJ, Hulshoff JB, van Dijk LV, et al. Predicting response to neoadjuvant chemoradiotherapy in esophageal cancer with textural features derived from pretreatment ^{18}F -FDG PET/CT imaging. *J Nucl Med.* 2017;58:723-729.
16. Yip SS, Coroller TP, Sanford NN, Mamon H, Aerts HJ, Berbeco RI. Relationship between the Temporal Changes in Positron-Emission-Tomography-Imaging-Based Textural Features and Pathologic Response and Survival in Esophageal Cancer Patients. *Front Oncol.* 2016;6:72.
17. Nakajo M, Jinguji M, Nakabeppu Y, et al. Texture analysis of ^{18}F -FDG PET/CT to predict tumour response and prognosis of patients with esophageal cancer treated by chemoradiotherapy. *Eur J Nucl Med Mol Imaging.* 2017;44:206-214.
18. Zhang H, Tan S, Chen W, et al. Modeling pathologic response of esophageal cancer to chemoradiation therapy using spatial-temporal ^{18}F -FDG PET features, clinical parameters, and demographics. *Int J Radiat Oncol Biol Phys.* 2014;88:195-203.
19. Mandard AM, Dalibard F, Mandard JC, et al. Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations. *Cancer.* 1994;73:2680-2686.
20. Boellaard R, Delgado-Bolton R, Oyen WJ, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging.* 2015;42:328-354.
21. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative - feature definitions. *CoRR.* 2016;abs/1612.07003.
22. Hatt M, Majdoub M, Vallieres M, et al. ^{18}F -FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *J Nucl Med.* 2015;56:38-44.
23. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer.* 2009;45:228-247.
24. Oxnard GR, Schwartz LH. Response phenotype as a predictive biomarker to guide treatment with targeted therapies. *J Clin Oncol.* 2013;31:3739-3741.
25. Desbordes P, Ruan S, Modzelewski R, et al. Predictive value of initial FDG-PET features for treatment response and survival in esophageal cancer patients treated with chemo-radiation therapy using a random forest classifier. *PLoS One.* 2017;12:e0173208.
26. Ypsilantis PP, Siddique M, Sohn HM, et al. Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks. *PLoS One.* 2015;10:e0137036.
27. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol.* 2010;49:1012-1016.
28. Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour ^{18}F -FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *Eur J Nucl Med Mol Imaging.* 2013;40:1662-1671.
29. Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in ^{18}F -FDG PET. *J Nucl Med.* 2012;53:693-700.

Chapter 6

HER2 Expression Integrated into Clinico-Radiomic Models Enhances Prediction of Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer Patients

Roelof J. Beukinga*
Da Wang*
Willemieke P.M. Dijksterhuis
Hette Faber
Johannes G.M. Burgerhof
Arend Karrenbeld
Véronique E.M. Mul
Riemer H.J.A. Slart
Rob P. Coppes
John Th.M. Plukker

*Both authors contributed equally.

Submitted to *Ann Surg Oncol*. 2018

ABBREVIATIONS

¹⁸ F-FDG PET	¹⁸ F-Fluorodeoxyglucose positron emission tomography
AIC	Akaike information criterion
AUC	area under the receiver operating characteristics curve
CROSS	chemoradiotherapy for oesophageal cancer followed by surgery study
CT	computed tomography
EC	esophageal cancer
HER2	human epidermal growth factor receptor 2
HIF1 α	hypoxia-inducible factor alpha
IRS	immuno-reactivity score
LRT	likelihood ratio test
nCRT	neoadjuvant chemoradiotherapy
PTCH1	patched1
SHH	sonic hedgehog
SUV	standardized uptake value

ABSTRACT

Background: To assess the complementary value of biological tumor markers to pre-defined clinico-radiomic models in predicting complete response to neoadjuvant chemoradiotherapy (nCRT) in esophageal cancer (EC) patients.

Methods: Expression of human epidermal growth factor receptor 2 (HER2), cluster of differentiation 44 (CD44), hypoxia-inducible factor alpha (HIF1 α), protein patched homolog 1 (PTCH1), and Sonic Hedgehog (SHH) was assessed by immunohistochemistry in pre-treatment tumor biopsies of 42 patients with locally advanced EC. After staging with ^{18}F -fluorodeoxyglucose positron emission tomography (^{18}F -FDG PET)/computed tomography (CT), all patients received at least four cycles of chemotherapy and full radiotherapy followed by esophagectomy. Pathologic response was categorized into complete versus incomplete (Mandard tumor regression grade 1 vs. 2-5). Incorporation of the immunohistochemistry scores into 3 pre-defined clinico-radiomic prediction models was evaluated by internal validation through bootstrapping.

Results: Pathologic examination revealed 8 (19.0%) complete and 34 (81.0%) incomplete responders. HER2 and CD44 were univariably associated with pathologic response. Incorporation of HER2 into the 3 pre-defined clinico-radiomic prediction models significantly improved the goodness-of-fit (likelihood ratio test [LRT] P values of 0.02, 0.01, and 0.02, respectively), minimized the information loss (Δ Akaike information criterions [Δ AICs] of 7.11, 7.61, and 7.08, respectively), and increased the discriminative ability (areas under the receiver operating characteristic curves [AUCs] of 0.82, 0.71, and 0.77, respectively). All remaining studied biological tumor markers did not yield model improvement.

Conclusion: Incorporation of HER2 into 3 pre-defined clinico-radiomic prediction models improved response prediction of nCRT in EC patients. Further model improvement may be obtained by integration of multimodal biomarkers.

Key Words: Biological tumor markers, ^{18}F -FDG PET/CT textural analysis, response prediction esophageal cancer

INTRODUCTION

Neoadjuvant chemoradiotherapy (nCRT) followed by surgery is the preferred treatment for locally advanced ($T1/N_{1-3}/M_0$; $T_{2-4a}/N_{0-3}/M_0$) resectable esophageal cancer (EC). Complete response to nCRT is achieved in 29% of the patients, whereas 18% will not respond at all (1). For complete responders, an active surveillance or “wait-and-see” policy instead of an esophagectomy may lead to equivalent results and may avoid the risk of subsequent morbidity and even mortality. The overall tumor volume measured on computed tomography (CT) and the maximum standard uptake value (SUV_{max}) measured on ^{18}F -fluorodeoxyglucose positron emission tomography (^{18}F -FDG PET) are important predictors of the clinical evaluation of tumor response. However, these predictors have rather unsatisfactory predictive values and hence no clinical relevant conclusions can be based on these features (2).

Intratumoral ^{18}F -FDG uptake heterogeneity prior to nCRT plays a key role in the acquired treatment resistance (3, 4). Knowledge of intratumoral heterogeneity may result in more robust predictive parameters as it is associated with tumor aggressiveness and impaired response to nCRT due to hypoxia (5). Textural characteristics of tumors potentially depict intratumoral ^{18}F -FDG spatial distribution. Recently, we developed prediction models based on clinical and ^{18}F -FDG PET/CT textural features which showed considerable higher discriminative abilities compared to the traditional SUV_{max} model (6). Although promising, further optimization is needed, which can be achieved by incorporating combinations of putative predictors derived from clinical, imaging, and basic research areas (7-10). Over the years, extensive research has been performed on identifying treatment resistance mechanisms to find novel molecular targets to increase sensitivity and response in EC (11). Based on solid data of biomarkers that have been associated with nCRT response, we included the following potentially useful biological tumor markers for further model optimization.

The Hedgehog pathway, with its receptor protein patched homolog 1 (PTCH1) and ligand Sonic Hedgehog (SHH), has been implicated in treatment resistance (12-14). In addition, the transcription factor hypoxia-inducible factor alpha (HIF1 α) is associated with tumor heterogeneity (5, 13). The cell surface marker cluster of differentiation 44 (CD44) has been associated with cancer stem cell-like cells (15, 16) and the human epidermal growth factor receptor 2 (HER2), have also gained predictive importance in EC (17, 18). Although explored extensively, none of these candidate predictive tumor markers have been incorporated as major predictors of nCRT response in clinical practice.

The aim of this study was to assess the complementary value of potential biological tumor markers to pre-defined clinico-radiomic prediction models in predicting complete response in EC patients.

MATERIALS AND METHODS

Patients

Patients with a histologically confirmed locally advanced (T1/N1-3/M0; T2-4a/N0-3/M0) EC (according to the seventh tumor-node-metastasis classification system) (19) were eligible for inclusion if pre-treatment biopsies and PET/CT were available. All patients had received at least four cycles of chemotherapy and full concomitant radiotherapy followed by esophagectomy with curative intent in the University Medical Center Groningen. Between March 2010 and February 2016, 66 consecutive patients had sufficient amounts of pre-treatment biopsy material available to be enrolled in this study. Of these 66 patients, 43 had sufficient tumor tissues, including one with squamous cell carcinoma. In this patient group, it seemed justified to focus the analysis only on the 42 patients with adenocarcinoma. All related data were collected from a prospectively maintained database. The aim of the current study was to improve our existing prediction model rather than to validate it, which justified the inclusion of 40 patients of our prior study (6). The Institutional Ethical Board approved this retrospective study and the requirement to obtain informed consent was waived.

Treatment and Pathology

The nCRT was given according to the Chemoradiotherapy for Oesophageal cancer followed by Surgery Study (CROSS) regimen, including carboplatin (area under the receiver operating characteristic curve [AUC] $2 \text{ mg} \cdot \text{min} \cdot \text{mL}^{-1}$) and paclitaxel (50 mg/m^2) with concurrent radiotherapy (41.4 Gy in 23 fractions) (1). A curative intended transthoracic open or minimal invasive esophagectomy with two-field mediastinal and upper abdominal lymphadenectomy was performed 6-8 weeks after completion of nCRT. The primary outcome was pathologic response according to the Mandard tumor regression grade (20). This grading system classifies the ratio of residual vital tumor cells and the degree of nCRT-induced fibrosis and defines complete response (Mandard tumor regression grade 1) as no residual vital tumor cells and non-response (Mandard tumor regression grade 5) as no tumor regression at all.

Clinico-Radiomic Prediction Models

Here, we aimed to improve 3 pre-defined clinico-radiomic prediction models (6). Model 1 was composed of the variables histologic type, clinical T-stage, and the PET-derived textural feature “long run low grey level emphasis” for quantifying long runs (coarse texture) with low grey levels. Model 2 was constructed of histologic type, clinical T-stage, and the CT-derived textural feature “run percentage” for quantifying homogeneity of runs (fine texture). Model 3 was defined as a full model incorporating all these variables. PET/CT imaging, volume of

interest delineation, and textural feature extraction were performed as described previously (6).

Immunohistochemistry and Analyses of Biological Tumor Markers

Immunohistochemistry staining was performed on 5 μ m tissue sections from archival biopsies using primary antibodies against CD44 (1:100, Biolegend), HIF1 α (1:100, ABCAM), PTCH1 (1:100, ABCAM), and SHH (1:100, ABCAM). HER2 was scored by a pathologist, specialized in upper gastrointestinal cancer, according to standardized methods in a continuous scale of 0, 1+, 2+ and 3+ when at least 5 clustered tumor cells were stained (21, 22). De-paraffinized tissue sections were immersed in PBS 2% hydrogen peroxidase to block endogenous peroxidase activity. Antigen-retrieval was performed and the sections were incubated overnight at 4°C with primary antibodies. Tissue sections were then incubated with biotinylated secondary antibodies at 1:300 dilutions for 1 hour. The ABC complex was formed using the Vectastain Elite ABC HRP kit (Vector Laboratories). This complex was visualized with SIGMA FAST 3,3'-Diaminobenzidine tablets (Sigma-Aldrich). Finally, sections were counterstained with haematoxylin and scored. HIF1 α was scored separately as it is expressed either in the cytosol (normoxia) or in the nucleus (hypoxia). CD44, HIF1 α _{cytosol} and PTCH1 were scored using the fifteen-point immuno-reactivity score (IRS), calculated by multiplying the percentage of positive cells with the stain intensity score (23). The percentage of positive cells was categorized into five levels by labeling 0% as 0, >0% to <5% as 1, $\geq$ 5% to <25% as 2, $\geq$ 25% to <50% as 3, $\geq$ 50% to <75% as 4, and $\geq$ 75% to 100% as 5. Stain intensity was categorized into 4 levels by labeling negative as 0, weak as 1, medium as 2, and strong as 3 (23). A dichotomous immunoscore was created by categorizing IRS into two levels by labeling absent/low expression (IRS 0-5) as 0 and moderate/high expression (IRS 6-15) as 1 (23). SHH demonstrated the maximum percentage of positive cells ($\geq$ 75% to 100%) in the majority of the cases. Therefore, SHH intensity was only scored dichotomously by categorizing stain intensity into absent/low expression (0-1 intensity) and moderate/high expression (2-3 intensity) (23). HIF1 α _{nucleus} was categorized as either present or absent. Scoring was blinded and was carried out by two researchers independently. Discordant cases and random slides of each marker were validated by the same experienced pathologist.

Statistical Analysis

Statistical analysis was performed with R 3.2.2 open-source software using the regression modeling strategies package (version 4.4-0), available from the Comprehensive R Archive Network (<http://www.r-project.org>). All textural features were normalized into the range [0,1].

The outcome variable, categorized as pathologic complete response versus pathologic incomplete response (Mandard tumor regression grade 1 vs. 2-5), was modeled using logistic regression. The biological tumor markers were introduced to a univariable logistic regression analysis. As has been stated earlier, dichotomizing HER2 substantially increases the chance of false-positive and false-negative cases (24, 25). Therefore, HER2 was analyzed as a continuous variable to retain the variation captured by the immunohistochemistry scoring system and to optimize the statistical power. All remaining tumor markers were analyzed as dichotomous immunoscores. Predictors were considered to fit the data significantly better in case of a likelihood ratio test (LRT) $P < 0.2$. Subsequently, the biological tumor markers were separately entered to models 1-3. Only those models that explained the data better than models 1-3 (LRT $P < 0.05$) and that minimized the loss of information were left for further consideration. The

TABLE 1
Patient and tumor characteristics.

Characteristics	pCR (<i>n</i> = 8)	Non-pCR (<i>n</i> = 34)
Sex		
Male	7 (87.5)	29 (85.3)
Female	1 (12.5)	5 (14.7)
Age, median (IQR), years	64.5 (9.0)	62.5 (10.0)
Histology		
Adenocarcinoma	8 (100.0)	34 (100.0)
Tumor location		
Distal esophagus/GEJ	8 (100.0)	34 (100.0)
Tumor length, median (IQR), cm	3.5 (7.5)	6.0 (4.0)
Clinical T-stage		
T1	0 (0.0)	1 (2.9)
T2	3 (37.5)	2 (5.9)
T3	5 (62.5)	28 (82.4)
T4a	0 (0.0)	3 (8.8)
Clinical N-stage		
N0	3 (37.5)	6 (17.6)
N1	4 (50.0)	16 (47.1)
N2	1 (12.5)	10 (29.4)
N3	0 (0.0)	2 (5.9)
Number of chemotherapy cycles		
4	2 (25.0)	8 (23.5)
5	6 (75.0)	26 (76.5)
Mandard tumor regression grade		
1	8 (100.0)	0 (0.0)
2	0 (0.0)	13 (38.2)
3	0 (0.0)	16 (47.1)
4	0 (0.0)	4 (11.8)
5	0 (0.0)	1 (2.9)

Data are numbers, with percentages in parentheses.

Abbreviations: pCR = pathologic complete response; IQR = interquartile range; and GEJ = gastroesophageal junction.

TABLE 2

The distribution of the immunohistochemistry scores.

Tumor markers	pCR (n = 9)	Non-pCR (n = 34)
Her2neu		
0	7 (87.5)	20 (58.8)
1+	0 (0.0)	6 (17.6)
2+	0 (0.0)	1 (2.9)
3+	0 (0.0)	7 (20.6)
Missing	1 (12.5)	0 (0.0)
CD44		
Negative	1 (12.5)	15 (44.1)
Positive	7 (87.5)	19 (55.9)
HIF1 α _{nucleus}		
Negative	8 (100.0)	31 (91.2)
Positive	0 (0.0)	3 (8.8)
HIF1 α _{cytosol}		
Negative	2 (25.0)	13 (38.2)
Positive	6 (75.0)	21 (61.8)
Ptch1		
Negative	2 (25.0)	4 (11.8)
Positive	6 (75.0)	30 (88.2)
SHH		
Negative	0 (0.0)	3 (8.8)
Positive	8 (100.0)	31 (91.2)

Data are numbers with percentages in parentheses.

Abbreviations: pCR = pathologic complete response; HER2 = human epidermal growth factor receptor 2; CD44 = cluster of differentiation 44; HIF1 α = hypoxia-inducible factor alpha; PTCH1 = protein patched homolog 1; and SHH = Sonic Hedgehog.

information loss was measured by the Akaike information criterion (AIC), which rewards goodness-of-fit, but penalizes models by the number of degrees of freedom. Models were considered to exhibit excessive multicollinearity in case of variance inflation factors > 4. Each model was quantified in terms of discrimination with the AUC. The models were internally validated by a bootstrap approach with 2000 repetitions, yielding the optimism-corrected AUC and 95% bootstrap percentile confidence intervals.

RESULTS

Patient Characteristics

Table 1 displays the patient and tumor characteristics. All patients received the full radiotherapy dose, whereas 32 patients (76.2%) received all 5 cycles of chemotherapy and 10 patients (23.8%) received 4 cycles of chemotherapy. Pathologic response was assessed as complete in 8 patients (19.0%) and as incomplete in 34 patients (81.0%).

FIGURE 1 Representative immunohistochemistry staining of 2 patients who were scored as immunohistochemistry 0 (A) and 2 patients who were scored as immunohistochemistry 3+ (B) for HER2 expression.

Complementary Value of the Biological Tumor Markers

HER2 staining failed in one patient and this observation was therefore scored as missing. Immunohistochemistry staining for all remaining biological tumor markers was successfully completed. The distribution of the immunohistochemistry scores are provided in Table 2. Of all examined biological tumor markers, only HER2 and CD44 were found to be significant at univariable logistic regression analysis with LRT *P* values of 0.01 and 0.08, respectively (Table 3). The addition of the biological tumor markers to models 1-3 revealed that HER2 was the only marker that significantly improved the goodness-of-fit (LRT *P* values of 0.02, 0.01, and 0.02, respectively), minimized the information loss (Δ AICs of 7.11, 7.61, and 7.08, respectively), and increased the discriminative ability (AUCs of 0.82, 0.71, and 0.77, respectively). Incorporating HER2 into model 1 exhibited the lowest AIC and the highest discriminative ability. High HER2 expression was found to be more prevalent in impaired responders. Figure 1 demonstrates representative pictures of low (0) and high (3+) HER2 expression patterns. Incorporating HER2 into models 1-3 exhibited minimal effects of multicollinearity with variance inflation factors of ≤ 1.4 . Although the incorporation of CD44, HIF1 α_{nucleus} , and SHH into models 1-3 slightly improved the discriminative ability,

TABLE 3

Goodness-of-fit and discriminative ability of all combinations of biological tumor markers incorporated into the pre-defined clinico-radiomic prediction models.

	none	model 1	model 2	model 3
none				
LRT <i>P</i>				
AIC		40.63	43.19	42.58
AUC (95% CI)		0.67 (0.44-0.94)	0.62 (0.41-0.88)	0.68 (0.47-0.96)
HER2				
LRT <i>P</i>	0.01	0.02	0.01	0.02
AIC	34.9	33.52	35.58	35.50
AUC (95% CI)	0.71 (0.62-0.78)	0.82 (0.45-1.00)	0.71 (0.47-0.85)	0.77 (0.54-0.92)
CD44				
LRT <i>P</i>	0.08	0.24	0.12	0.22
AIC	41.77	41.25	42.82	43.09
AUC (95% CI)	0.65 (0.50-0.79)	0.72 (0.49-0.90)	0.72 (0.48-0.87)	0.71 (0.48-0.87)
HIF1 α _{nucleus}				
LRT <i>P</i>	0.25	0.17	0.14	0.17
AIC	43.58	40.74	43.04	42.71
AUC (95% CI)	0.54 (0.49-0.57)	0.70 (0.48-0.95)	0.66 (0.45-0.87)	0.69 (0.49-0.91)
HIF1 α _{cytosol}				
LRT <i>P</i>	0.47	0.40	0.46	0.39
AIC	44.38	41.92	44.64	43.85
AUC (95% CI)	0.51 (0.36-0.62)	0.57 (0.36-0.93)	0.58 (0.37-0.79)	0.56 (0.37-0.92)
PTCH1				
LRT <i>P</i>	0.36	0.92	0.69	0.99
AIC	44.08	42.62	45.03	44.58
AUC (95% CI)	0.52 (0.37-0.63)	0.60 (0.36-0.87)	0.58 (0.37-0.86)	0.64 (0.40-0.88)
SHH				
LRT <i>P</i>	0.25	0.17	0.13	0.17
AIC	43.58	40.77	42.92	42.66
AUC (95% CI)	0.54 (0.49-0.57)	0.69 (0.47-0.96)	0.67 (0.48-0.89)	0.70 (0.49-0.94)

Abbreviations: LRT *P* = likelihood-ratio test *P* value; AIC = Akaike information criterion (AIC); AUC = area under the receiver operating characteristic; and boot = internally validated with bootstrapping.

none of the remainder biological tumor markers yielded a significant improved goodness-of-fit. Hence, all remainder biological tumor markers were discarded. An overview of the variables of the pre-defined prediction models and HER2 is provided in Figure 2.

DISCUSSION

This study evaluated various biological tumor markers to improve response prediction of nCRT in EC patients. We demonstrated that the biological marker HER2 added complementary information to our previous constructed clinico-radiomic prediction models. Low HER2 expression yielded a higher probability of achieving a pathologic complete response. Based on the reported performance measures, the preferred prediction model incorporated the

which yielded 75% sensitivity, 81% specificity, and 78.1% accuracy (26). Another study showed that enrichment of postoperative HER2 levels in surgical resection material compared to pre-treatment biopsies was associated with poor pathologic response (27).

According to the general guidelines, HER2 status should be verified from all patients with advanced esophageal adenocarcinoma, who are potential candidates for HER2-targeting, prior to the initiation of HER2-targeted therapy (22). This means that HER2 status assessment is not part of the pre-treatment staging of the disease. Indeed, HER2 status was verified at the initial staging of our study cohort in only 28% of the patients ($n = 12$). Our findings suggest a potential value of adding pre-treatment HER2 status in treatment stratification in locally advanced EC. Repeating the analysis without excluding the squamous cell carcinoma patient, yielded equivalent results.

Several potential limitations of the study should be noted. The study group ($n = 42$) is relatively small, allowing the substantial chance of model overfitting. To discourage overfitting, we used the AIC and bootstrapping for model discrimination. In the final model, overfitting effects were minimal, as indicated by low variance inflation factors and low AUC optimism. However, as even internal validation may not be indicative for the final model's performance in future settings, the model should be validated in an independent validation cohort. Moreover, the reproducibility of HER2 evaluation and radiomic feature extraction should be established before it can be safely implemented and generalized in a clinical setting. Immunohistochemistry protocols may differ among institutes and results depend on the level of experience of the researcher. In this study, inter-rater variation was minimized by the independent evaluation of two researchers. Discordant cases were validated by an experienced pathologist. Following the guidelines in the immunohistochemistry evaluation of HER2, tumors with 0 and 1+ are considered negative while tumors with 3+ are considered positive. Tumors with 2+ are considered ambiguous and are generally subjected to fluorescence in situ hybridization or silver in situ hybridization techniques to confirm the HER2 status (22, 25). Tumors with 2+ represent the majority of false-positive results when dichotomizing HER2 (24), while tumors with 1+ contribute to the occurrence of false-negative results (25). Hence, as mentioned earlier, we analyzed HER2 as a continuous variable. It should be noted that one ($n = 1$) patient had a 2+ score, however, fluorescence in situ hybridization showed no HER2 amplification. Moreover, only a few studies have reported on measurement errors (i.e. reliability, reproducibility, or repeatability) of radiomics (28-30). Although further research is warranted, initial results show that most radiomic features are sensitive to numerous confounding factors including image acquisition, reconstruction protocols, or delineation methods. Standardization in the radiomics extraction workflow therefore remains essential to achieve routine clinical adoption.

CONCLUSION

This study indicates that HER2 and CD44 in the initial diagnostic work-up could be useful biological tumor markers in predicting complete response to nCRT in EC and hence to improve treatment strategies. However, HER2 was the only marker which yielded significantly improved discriminative ability when being incorporated into the pre-defined clinico-radiomic prediction models. This study shows that integration of multimodel biomarkers may improve the performance of such prediction models.

REFERENCES

1. van Hagen P, Hulshof MC, van Lanschot JJ, et al. Preoperative chemoradiotherapy for esophageal or junctional cancer. *N Engl J Med*. 2012;366:2074-2084.
2. Kwee RM. Prediction of tumor response to neoadjuvant therapy in patients with esophageal cancer with use of ^{18}F FDG PET: a systematic review. *Radiology*. 2010;254:707-717.
3. Dexter DL, Leith JT. Tumor heterogeneity and drug resistance. *J Clin Oncol*. 1986;4:244-257.
4. O'Connor JP, Rose CJ, Waterton JC, Carano RA, Parker GJ, Jackson A. Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clin Cancer Res*. 2015;21:249-257.
5. Harris AL. Hypoxia--a key regulatory factor in tumour growth. *Nat Rev Cancer*. 2002;2:38-47.
6. Beukinga RJ, Hulshoff JB, van Dijk LV, et al. Predicting response to neoadjuvant chemoradiotherapy in esophageal cancer with textural features derived from pretreatment ^{18}F -FDG PET/CT imaging. *J Nucl Med*. 2017;58:723-729.
7. Bollschweiler E, Holscher AH, Schmidt M, Warnecke-Eberz U. Neoadjuvant treatment for advanced esophageal cancer: response assessment before surgery and how to predict response to chemoradiation before starting treatment. *Chin J Cancer Res*. 2015;27:221-230.
8. Tao CJ, Lin G, Xu YP, Mao WM. Predicting the response of neoadjuvant therapy for patients with esophageal carcinoma: an in-depth literature review. *J Cancer*. 2015;6:1179-1186.
9. Vallbohmer D, Brabender J, Grimminger P, Schroder W, Holscher AH. Predicting response to neoadjuvant therapy in esophageal cancer. *Expert Rev Anticancer Ther*. 2011;11:1449-1455.
10. Forshaw MJ, Gossage JA, Mason RC. Neoadjuvant chemotherapy for oesophageal cancer: the need for accurate response prediction and evaluation. *Surgeon*. 2005;3:373-382.
11. Wang D, Plukker JTM, Coppes RP. Cancer stem cells with increased metastatic potential as a therapeutic target for esophageal cancer. *Semin Cancer Biol*. 2017;44:60-66.
12. Yang Z, Cui Y, Ni W, Kim S, Xuan Y. Gli1, a potential regulator of esophageal cancer stem cell, is identified as an independent adverse prognostic factor in esophageal squamous cell carcinoma. *J Cancer Res Clin Oncol*. 2017;143:243-254.
13. Wang Y, Ding Q, Yen CJ, et al. The crosstalk of mTOR/S6K1 and Hedgehog pathways. *Cancer Cell*. 2012;21:374-387.
14. Ma X, Sheng T, Zhang Y, et al. Hedgehog signaling is activated in subsets of esophageal cancers. *Int J Cancer*. 2006;118:139-148.
15. Zoller M. CD44: can a cancer-initiating cell profit from an abundantly expressed molecule? *Nat Rev Cancer*. 2011;11:254-267.
16. Smit JK, Faber H, Niemantsverdriet M, et al. Prediction of response to radiotherapy in the treatment of esophageal cancer using stem cell markers. *Radiother Oncol*. 2013;107:434-441.
17. Gowryshankar A, Nagaraja V, Eslick GD. HER2 status in Barrett's esophagus & esophageal cancer: a meta analysis. *J Gastrointest Oncol*. 2014;5:25-35.
18. Miller S, Hung M. Regulation of her2/neu gene-expression (review). *Oncol Rep*. 1995;2:497-503.

19. Rice TW, Blackstone EH, Rusch VW. 7th edition of the AJCC Cancer Staging Manual: esophagus and esophagogastric junction. *Ann Surg Oncol*. 2010;17:1721-1724.
20. Mandard AM, Dalibard F, Mandard JC, et al. Pathologic assessment of tumor regression after preoperative chemoradiotherapy of esophageal carcinoma. Clinicopathologic correlations. *Cancer*. 1994;73:2680-2686.
21. Hofmann M, Stoss O, Shi D, et al. Assessment of a HER2 scoring system for gastric cancer: results from a validation study. *Histopathology*. 2008;52:797-805.
22. Bartley AN, Washington MK, Colasacco C, et al. HER2 testing and clinical decision making in gastroesophageal adenocarcinoma: guideline from the College of American Pathologists, American Society for Clinical Pathology, and the American Society of Clinical Oncology. *J Clin Oncol*. 2017;35:446-464.
23. Honing J, Pavlov KV, Mul VE, et al. CD44, SHH and SOX2 as novel biomarkers in esophageal cancer patients treated with neoadjuvant chemoradiotherapy. *Radiother Oncol*. 2015;117:152-158.
24. Tubbs RR, Pettay JD, Roche PC, Stoler MH, Jenkins RB, Grogan TM. Discrepancies in clinical laboratory testing of eligibility for trastuzumab therapy: apparent immunohistochemical false-positives do not get the message. *J Clin Oncol*. 2001;19:2714-2721.
25. Prins MJ, Ruurda JP, van Diest PJ, van Hillegersberg R, Ten Kate FJ. The significance of the HER-2 status in esophageal adenocarcinoma for survival: an immunohistochemical and an in situ hybridization study. *Ann Oncol*. 2013;24:1290-1297.
26. Warnecke-Eberz U, Metzger R, Bollschweiler E, et al. TaqMan low-density arrays and analysis by artificial neuronal networks predict response to neoadjuvant chemoradiation in esophageal cancer. *Pharmacogenomics*. 2010;11:55-64.
27. Vallbohmer D, Kuhn E, Warnecke-Eberz U, et al. Failure in downregulation of intratumoral survivin expression following neoadjuvant chemoradiation in esophageal cancer. *Pharmacogenomics*. 2008;9:681-690.
28. Yan J, Chu-Stern JL, Loi HY, et al. Impact of image reconstruction settings on texture features in ¹⁸F-FDG PET. *J Nucl Med*. 2015;56:1667-1673.
29. Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A. The impact of image reconstruction settings on ¹⁸F-FDG PET radiomic features: multi-scanner phantom and patient studies. *Eur Radiol*. 2017;27:4498-4509.
30. van Velden FH, 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;18:788-795.

Part III

Applications of ^{18}F -FDG PET Radiomics in Infectious Disorders

Chapter 7

Textural Features of ^{18}F -Fluorodeoxyglucose Positron Emission Tomography Scanning in Diagnosing Aortic Prosthetic Graft Infection

Ben R. Saleem*
Roelof J. Beukinga*
Ronald Boellaard
Andor W.J.M. Glaudemans
Michel M.P.J. Reijnen
Clark J. Zeebregts
Riemer H.J.A. Slart

*Both authors contributed equally.

Eur J Nucl Med Mol Imaging. 2017;44:886-894

ABBREVIATIONS

^{18}F -FDG PET	^{18}F -Fluorodeoxyglucose positron emission tomography
AGI	aorto-iliac graft infection
AUC	area under the receiver operating characteristics curve
CRP	C-reactive protein
CT	computed tomography
EANM	European Association of Nuclear Medicine
EVAR	endovascular aortic repair
GLCM	grey level co-occurrence matrix
GLRLM	grey level run-length matrix
GLSZM	grey level size-zone matrix
ICC	intraclass correlation coefficient
PTFE	polytetrafluoroethylene
SUV	standardized uptake value
TBR	tissue-to-background ratio
VGS	visual grading scale
VOI	volume of interest

ABSTRACT

Background: The clinical problem in suspected aortoiliac graft infection (AGI) is to obtain proof of infection. Although ^{18}F -fluorodeoxyglucose (^{18}F -FDG) positron emission tomography scanning (PET) has been suggested to play a pivotal role, an evidence-based interpretation is lacking. The objective of this retrospective study was to examine the feasibility and utility of ^{18}F -FDG uptake heterogeneity characterized by textural features to diagnose AGI.

Methods: Thirty patients with a history of aortic graft reconstruction who underwent ^{18}F -FDG PET/CT scanning were included. Sixteen patients were suspected to have an AGI (group I). AGI was considered proven only in the case of a positive bacterial culture. Positive cultures were found in 10 of the 16 patients (group Ia), and in the other six patients, cultures remained negative (group Ib). A control group was formed of 14 patients undergoing ^{18}F -FDG PET for other reasons (group II). PET images were assessed using conventional maximal standardized uptake value (SUV_{max}), tissue-to-background ratio (TBR), and visual grading scale (VGS). Additionally, 64 different ^{18}F -FDG PET based textural features were applied to characterize ^{18}F -FDG uptake heterogeneity. To select candidate predictors, univariable logistic regression analysis was performed ($\alpha=0.16$). The accuracy was satisfactory in case of an $\text{AUC}>0.8$.

Results: The feature selection process yielded the textural features named variance ($\text{AUC}=0.88$), high grey level zone emphasis ($\text{AUC}=0.87$), small zone low grey level emphasis ($\text{AUC}=0.80$), and small zone high grey level emphasis ($\text{AUC}=0.81$) most optimal for distinguishing between groups I and II. SUV_{max} , TBR, and VGS were also able to distinguish between these groups with AUCs of 0.87, 0.78, and 0.90, respectively. The textural feature named short run high grey level emphasis was able to distinguish group Ia from Ib ($\text{AUC}=0.83$), while for the same task the TBR and VGS were not found to be predictive. SUV_{max} was found predictive in distinguishing these groups, but showed an unsatisfactory accuracy ($\text{AUC}=0.75$).

Conclusion: Textural analysis to characterize ^{18}F -FDG uptake heterogeneity is feasible and shows promising results in diagnosing AGI, but requires additional external validation and refinement before it can be implemented in the clinical decision-making process.

Key Words: ^{18}F -FDG PET; Aortic prosthetic graft infection; Textural features

INTRODUCTION

Aortoiliac prosthetic graft infection (AGI) is a severe complication after prosthetic graft placement, which is associated with high morbidity and mortality (1-5). Positive bacterial cultures, either from percutaneous aspirated perigraft fluid or from surgically obtained material, are considered the gold standard for diagnosing AGI (1-5). However, a perigraft abscess or fluid collection is not always present, and even if present it may not always be suitable for puncture. The diagnosis of AGI, therefore, remains an important challenge.

Non-invasive ^{18}F -fluorodeoxyglucose positron emission tomography (^{18}F -FDG PET) imaging may play an important role in increasing the diagnostic accuracy of infectious diseases with inherent elevated intracellular glucose metabolism (6-11). The PET images can either be evaluated semi-quantitatively using the maximal standardized uptake value (SUV_{max}) and the tissue-to-background ratio (TBR), and visually using the visual grading scale (VGS). For diagnostic assessments which are based on a combination of these parameters, our group reported a high sensitivity (up to 91%), but a low specificity (up to 64%) (10), which was also confirmed by others (6, 7, 11). Additionally, there is no consensus with respect to the interpretation of these ^{18}F -FDG PET findings. Recently, $\text{SUV}_{\text{max}} > 8$ in the perigraft area was suggested as the cutoff value for proof of an infection of a thoracic prosthetic graft (12). However, this value must be interpreted with caution as this study was conducted in only nine patients and scans were not acquired according to European Association of Nuclear Medicine (EANM) recommendations (13). Moreover, our group found the above-mentioned quantitative measures when used as so-called cutoff values to determine infection to be of modest utility in the diagnosis of AGI (14). As a consequence, more accurate quantification methods are needed. Since a heterogeneous uptake is associated with infection, the distribution pattern of ^{18}F -FDG activity may help identify AGI with a higher diagnostic precision (6). A powerful tool for quantifying such distribution is textural analysis, which may provide valuable information regarding biological heterogeneity. The concept of textural analysis is generally based on the spatial arrangement of voxels in a predefined volume of interest (VOI). Spatial heterogeneity can be depicted from different spatial interrelationships on ^{18}F -FDG PET scans. Within the field of clinical oncology, textural analysis already has yielded promising results in predicting response by quantifying intra-tumoral heterogeneity (15-23).

In the current study, we introduced the principle of textural analysis into the field of infectious diseases and aimed to investigate feasibility and accuracy of textural features to diagnose AGI.

FIGURE 1 Flow chart of patient disposition

MATERIALS AND METHODS

Design of the Study

For this retrospective analysis we included all patients ($n = 30$) from a prospective database of patients with a history of aortoiliac prosthetic graft reconstruction who underwent ^{18}F -FDG PET/CT at our tertiary referral center between December 2009 and May 2015. Medical charts were analysed to identify those patients with a clinical suspicion of an AGI. An AGI was clinically suspected in case of undefined fever, a deep wound infection, persisting high laboratory infection parameters [e.g. erythrocyte sedimentation rate, white blood cell count, and/or C-reactive protein (CRP)], or a combination of these factors. AGI was considered proven only in cases where a positive bacterial culture of material was obtained from peri-prosthetic samples obtained during diagnostic work-up or surgery. Sixteen patients were clinically suspected to have an AGI (group I) of which 10 had a positive culture (group Ia) and six had a negative culture (group Ib). The remaining 14 patients underwent a ^{18}F -FDG PET/CT scan because of cancer staging and were used as control group. Fig. 1 shows a flow chart of the patient disposition.

Patient characteristics and information about the initial operation, type of graft material, clinical symptoms and laboratory parameters at the time of ^{18}F -FDG PET imaging, and definite type of treatment were collected from the medical records. Co-morbidities were defined as recommended by the Ad Hoc Committee on Reporting Standards (24). This study was approved by the institutional ethical review board (METc 2015/082). Patients' data were analysed anonymously.

^{18}F -FDG PET Imaging and Analysis

Non-gated PET/CT imaging was performed with a dedicated integrated PET/CT system (Biograph mCT PET/CT, Siemens, Knoxville, TN, USA). All patients fasted overnight with

no restrictions on drinking water and with a minimum fasting time of 6 h prior to PET/CT. ^{18}F -FDG was administered intravenously with a weight-based activity of 3 MBq/kg. Sixty minutes after tracer injection, patients were positioned on the camera table with the arms in upright position. PET images were acquired with 3 min per bed position. An initial low dose CT scan was performed to ensure that the region of interest was included in the field of view, where after an inspiration breath-hold low-dose CT for attenuation correction was performed with 100 kVp and 30 mAs. Image data were reconstructed using standard methods and images were standardized according to EANM guidelines (13).

An experienced nuclear medicine physician assessed the ^{18}F -FDG PET images, including VGS, SUV_{max} and TBR. The five-point VGS was graded as follows: grade 0, ^{18}F -FDG uptake similar to that in the background; grade I, low ^{18}F -FDG uptake, comparable with inactive muscles and fat; grade II, moderate ^{18}F -FDG uptake, clearly visible and higher than uptake by inactive muscles and fat; grade III, strong ^{18}F -FDG uptake, but distinctly less than the physiologic urine bladder activity; and grade IV, very strong ^{18}F -FDG uptake, comparable with the physiologic urinary activity of the bladder (10, 25). A VOI was drawn around the area of the vascular prosthesis to calculate the SUV_{max} . SUV_{max} corresponded to the voxel with the highest ^{18}F -FDG uptake. The TBR was defined as the SUV_{max} divided by the mean SUV of the caval vein (blood pool).

Volume of Interest

^{18}F -FDG PET based textural features to characterize heterogeneity of ^{18}F -FDG uptake in the aortic prosthetic graft were measured for each patient. Fig. 2 displays the image processing and extraction of all textural features ($n=64$) from a predefined VOI. VOIs were manually delineated (Fig. 2a) by two independent experienced nuclear medicine physicians on axial planes of the low-dose CT to enclose three-dimensional coverage of the entire suspected prosthetic graft, using PMOD 3.6 software (PMOD Inc., Zurich, Switzerland). However, as the partial volume effect may cause the activity to be smeared out over a larger area than the actual structure, and the total number of counts is preserved, we also included high uptake regions around the graft material with clear contamination of activity from the graft (spill over). Moreover, we assessed whether ^{18}F -FDG uptake was physiological or non-physiological based on the CT-based anatomical location and excluded high uptake regions with clear spill-in effects from physiological neighbouring tissues such as the kidneys, ureter, and urine bladder. Although this VOI definition leads to analysis on the whole prosthetic graft volume, we chose for this definition since analysis of the most-diseased segment of the aortic graft would require a semi-objective identification of the ^{18}F -FDG-avid area (13). Additionally, prosthetic grafts of group II do often not contain a ^{18}F -FDG-avid area, which, therefore, complicates the comparison with Group I. After VOI delineation, the PET/CT

imaging data and VOI delineations were loaded into Matlab 2014b (Mathworks, Natick, MA, USA) for processing and analysis (Fig. 2b). The VOI delineations were then registered to the PET images. PET voxels, which were enclosed for $\geq 50\%$ coverage, were considered to be part of the suspected prosthetic graft (Fig. 2c). For noise reduction, SUV was discretized with

FIGURE 2 A) Manual delineation on LD-CT; B) Overlay of colour-mapped PET image onto LD-CT; C) Cropping of PET VOI; D) Feature extraction; I) Global assessment of tonal distribution by means of first order textural features; II) Assessment of pairwise arrangement of voxels by means of construction of grey level co-occurrence matrix (GLCM); III) Assessment of alignment of voxels with the same intensity by means of construction of grey level run-length matrix (GLRLM); IV) Assessment of characteristics of homogenous zones by means of construction of grey level size-zone matrix (GLSZM)

fixed 0.5 g/mL increments according to Doane's optimal bin width (26). Textural analysis was performed on the cropped PET VOI. To find the influence of the volume of the VOI on textural features the Pearson correlation coefficient was calculated.

Textural Features Extraction

Fig. 2d and supplemental table 1 provide a full overview of all 64 analysed textural features. We extracted 19 first order textural features (based on the grey level distribution, but without spatial information of voxels). Texture can be characterized by replications of (small) texture elements. These texture elements consist of contiguous voxels with certain spatial and intensity properties. We obtained the distributions of three different texture elements, i.e. the grey level co-occurrence (or spatial dependence) matrix (GLCM) for pairwise arrangement of voxels (27), the grey level run-length matrix (GLRLM) for alignment of voxels with an identical intensity (28), and the grey level size-zone matrix (GLSZM) for characteristics of homogeneous zones (29). From these matrices, we extracted 46 second order textural features (which are thus based on spatial information of the grey levels). These textural features were extracted with a voxel-to-voxel distance offset of $d = 1$ and directional analysis was performed with a connectivity of 26 voxels (analysis in 13 angular directions). All extracted textural features were normalized to the range $[0,1]$. To determine the influence of noise, it was tested whether noise was equally distributed among the groups and the correlation between noise and each textural feature was computed. Therefore, a sphere of 3 cm in diameter was drawn in the liver; the coefficient of variation was determined as noise parameter.

Statistical Analysis

Baseline characteristics are presented as mean $\pm$ standard deviation or percentages. To select candidate predictors to identify infection, univariable logistic regression analysis was performed. All potential predictors that met the Akaike Information Criterion (AIC) were considered significant (30). To discourage overfitting, the AIC is based on rewarding goodness of fit and penalizing complexity in the model. The AIC requires $\chi^2 > 2$ df, i.e. when considering a predictor with one degree of freedom df; this implies a significance level $\alpha = P(\chi^2 \geq 2) = 0.16$ (31). The accuracy of all candidate predictors was measured by the area under the receiver operating characteristic curve (AUC). Textural features were considered to have a good accuracy in case of an $AUC > 0.8$. Moreover, textural feature values may be subject to inter-observer variability in delineation of the prosthesis. Textural features were considered stable in case of a minimum acceptable excellent agreement indicated by an intra-class correlation coefficient (ICC) level of 0.75 (32). To obviate multicollinearity among all significant, accurate, and stable considered textural features, the pairwise Pearson correlation coefficient was evaluated. When the correlation of a pair of variables was > 0.8 , the variable with the lowest AUC was excluded from the set of features chosen for AGI characterization.

TABLE 1
Patient characteristics

Total	Group Ia	Group Ib	Group II
<i>n</i>	10	6	14
Sex			
Male	7	4	13
Female	3	2	1
Age; years, mean (range)	66 (42–76)	68 (58–77)	70 (52–77)
Co-morbidity*			
Diabetes mellitus	6	2	5
Tobacco use	7	5	12
Hypertension	8	4	3
Hyperlipemia	6	4	7
Cardiac disease	4	5	12
Renal disease	7	2	11
Pulmonary disease	9	3	8

* Defined according to the Ad Hoc Committee on Reporting Standards (24)

Data were collected and analysed using IBM SPSS 20.0 software (IBM Corp, Armonk, NY, USA).

RESULTS

Baseline Patient Characteristics

Twenty-four (80%) of the 30 patients were male. The mean age at the time of PET/CT scanning was 68 years, ranging from 42 to 77 years. Patient characteristics are described in Table 1. There were no significant differences between the groups with respect to demographic data (Pearson Chi-Square test). Table 2 shows the organisms obtained from culture in group I. Twenty-one (70%) patients were initially treated because of aneurysmal disease and the remaining for occlusive disease. Operative details of the primary operation are listed in Table 3. The time intervals between the initial operation and the ^{18}F -FDG PET scan for diagnosing AGI were 43.6 ± 42.0 months and 75.5 ± 55.7 months for group Ia and Ib, respectively. In group Ia and Ib, antibiotic treatment was initialized before the ^{18}F -FDG PET scan was performed in eight (80%) and five (83%) patients with 2.2 ± 31.5 days and 25.8 ± 30.0 days since clinical suspicion of AGI, respectively. Pre-scan serum glucose levels ranged from 4.0 to 7.7 mmol/L (median 5.6 mmol/L, mean 5.7 mmol/L), including the diabetic patients. Two patients in group Ib died within 30 days after surgical removal of the suspected graft due to postoperative complications. No autopsy was performed on these two patients. The other four patients had no signs of infection during further follow-up (mean 7 months), suggesting that these grafts were correctly classified as non-infected. Nine patients in group

TABLE 2
Bacteriology of infected prosthetic graft material in group I

	Culture obtained from:		n (%)
	Perigraft fluid (n = 10)	Surgery (n = 6)	
Group Ia	5	5	10 (63)
Group Ib	5	1	6 (37)
Organism			
Candida albicans	0	1	3
Coagulase-negative staphylococci	2	1	1
Enterococcus faecalis/faecium	0	2	3
Escherichia coli	1	1	1
Granulicatella adiacens	0	0	1
Nocardia farcinica	1	1	1
Proteus mirabilis	0	1	1
Proteus vulgaris	0	1	1
Staphylococcus aureus	2	0	2

II died during follow up (67.6 ± 37.5 months) mainly because of malignancy, without any sign of infection.

Visual Grading Scale

VGS for group I was 3.25 ± 1.06 . VGS was 3.50 ± 0.71 for group Ia and 2.83 ± 1.47 for group Ib. The VGS for group II was 1.4 ± 0.94 . The VGS was found to differ significantly between groups I and II ($P < 0.01$), with an AUC of 0.90. However, the VGS appeared not to differ significantly between groups Ia and Ib ($P = 0.26$), with an AUC of 0.64.

TABLE 3
Graft location and material at initial operation

	Group Ia n (%)	Group Ib n (%)	Group II n (%)
Total			
Underlying disease			
Aneurysmatic	5 (50)	4 (57)	12 (86)
Occluding	5 (50)	2 (43)	2 (14)
Graft location			
Aortoiliac	6 (60)	5 (83)	13 (93)
Iliofemoral	4 (40)	1 (17)	1 (7)
Type of reconstruction			
Open	9 (90)	6 (100)	8 (57)
Endovascular	1 (10)	0 (0)	6 (43)
Graft material			
Dacron*	7 (70)	6 (100)	7 (50)
PTFE	2 (20)	0 (0)	0 (0)
EVAR	1 (10)	0 (0)	7 (50)
Medtronic*	1 (10)	0 (0)	4 (29)
Cook-Zenith*	0 (0)	0 (0)	3 (21)

Abbreviations: PTFE = polytetrafluoroethylene; and EVAR = endovascular aortic repair

TABLE 4

Regression analysis results, accuracy, and robustness of the selected variables

Variable	I vs. II		Ia vs. Ib		ICC
	P	AUC	P	AUC	
Conventional measures					
Maximal standardized uptake value	0.01	0.87	0.13	0.75	
Tissue to background ratio	0.06	0.45	0.35	0.70	
Visual grading scale	<0.01	0.90	0.26	0.64	
First order textural features					
Variance	0.01	0.88	0.17	0.70	0.85
GLRLM-based textural features					
Short run high grey level emphasis [*]	0.02	0.79	0.07	0.83	0.75
GLSZM-based textural features					
High grey level zone emphasis [†]	0.01	0.87	0.12	0.78	0.83
Small zone low grey level emphasis	0.01	0.80	0.16	0.73	0.86
Small zone high grey level emphasis	0.04	0.81	0.15	0.75	0.79

Variables satisfying $P < 0.16$, $AUC > 0.8$, and $ICC > 0.75$ are given in bold

Abbreviations: GLRLM = grey level run-length matrix; and GLSZM = grey level size-zone matrix

* Textural feature selected for separating group Ia from Ib

† Textural feature selected for separating group I from II

Maximum Standardized Uptake Value

The SUV_{max} for group I was 7.01 ± 2.31 compared to 4.17 ± 1.86 for group II. The SUV_{max} for group Ia was 7.72 ± 2.22 compared to 5.83 ± 2.09 for group Ib. SUV_{max} was found to be predictive in distinguishing group I from II ($P = 0.01$) with an accuracy of $AUC = 0.87$. SUV_{max} was also able to distinguish group Ia from Ib ($P = 0.13$), but with an unsatisfactory accuracy of $AUC = 0.75$.

Tissue-to-Background Ratio

TBR for group I was 4.57 ± 2.14 . TBR was 4.86 ± 2.15 for group Ia and 3.82 ± 2.14 for group Ib. The TBR for group II was 2.94 ± 1.54 . The TBR was found to differ significantly between groups I and II ($P = 0.06$) with an AUC of 0.78, but appeared not to differ significantly between groups Ia and Ib ($P = 0.35$) with an AUC of 0.70.

Textural Features

Textural analysis was completed for 30 patients. No significant correlations were found between volume and texture features. Fifteen (22%) of the studied textural features, were found to be robust for inter-observer variability in delineation of the prosthesis and were suitable for AGI prediction. Four textural features fulfilled all selection criteria (P -value < 0.16 in the univariable analysis, $AUC > 0.80$, and $ICC > 0.75$) in distinguishing suspected from non-suspected graft infection (Table 4). The high-grey-level-zone-emphasis and small-zone-high-grey-level-emphasis were found to correlate, hence the small-zone-high-grey-level-

FIGURE 3 A coronal view of ^{18}F -FDG PET images of proven and non-proven infected prosthetic aortic grafts, and the corresponding values of conventional measures and the set of textural features chosen for AGI characterization. Abbreviations: SRHGLe = short-run-high-grey-level-emphasis and HGLZe = high-grey-level-zone-emphasis

emphasis was excluded from further analysis since this variable had the lowest AUC of the two. Variance, high-grey-level-zone-emphasis, and small-zone-low-grey-level-emphasis remained predictive for AGI characterization with AUCs of 0.88, 0.87, and 0.81, respectively. Short-run-high-grey-level-emphasis was the only textural feature to fulfil all selection criteria in distinguishing proven (group Ia) from non-proven (group Ib) infection (Table 4). The AUC of the short-run-high-grey-level-emphasis was 0.83, allowing a sensitivity of 80% and a specificity of 100% using an optimal threshold of 0.70. The short-run-high-grey-level-emphasis value for group Ia was 20.45 ± 11.02 compared to 10.05 ± 3.92 for group Ib, 16.55 ± 10.25 for group I, and 7.65 ± 4.38 for group II. Fig. 3 shows a coronal view of ^{18}F -FDG PET images of proven and non-proven infected prosthetic grafts and the corresponding values of the conventional measures, and the selected textural features. No significant differences were found in the amount of noise between group I and II ($P=0.34$) neither between group Ia and Ib ($P=0.10$).

DISCUSSION

This study investigates the relationship between ^{18}F -FDG uptake heterogeneity in the aortic prosthetic graft, as characterized by textural features, and AGI. This study shows that textural analysis of AGI is feasible and may increase the accuracy to diagnose AGI compared to conventional assessment.

In this study, several textural features were found to be robust for inter-observer variability in delineation of the prosthesis and seem to be suitable for AGI prediction. Short-run-high-grey-level-emphasis, which is highly dependent on the occurrence of short runs (and thus a heterogeneous ^{18}F -FDG uptake) with high grey levels, was the only textural feature to distinguish proven (group Ia) from non-proven (group Ib) infection. The short-run-high-grey-level-emphasis demonstrated higher values for the studied infected prosthetic grafts compared to the uninfected prosthetic grafts. This finding, therefore, supports the hypothesis that a high and heterogeneous ^{18}F -FDG uptake is associated with infected prosthetic grafts. The short-run-high-grey-level-emphasis was most efficient in identifying AGI within the suspected group, whereas for the same task the performances of SUV_{max} , TBR, and VGS measurements were all limited.

Diagnoses based on the standard parameters could significantly distinguish between patients being suspected (group I) and non-suspected (group II) of having AGI. VGS showed the highest accuracy of all studied parameters, indicating that these groups can sufficiently be distinguished without textural analysis. However, among these standard parameters, only the SUV_{max} was able to distinguish group Ia from Ib. Spacek et al. visually interpreted ^{18}F -FDG uptake as “intense”, “inhomogeneous”, or “none” and found intense focal ^{18}F -FDG uptake to be a significant predictor for AGI (9), which confirms that SUV_{max} results were found significant in the current study. However, as was also supported by our previous study (14), the accuracy of SUV_{max} was moderate and insufficient for changing clinical decision-making. Of interest, Keidar et al. visually assessed patterns of uptake for non-infected vascular grafts in patients undergoing a ^{18}F -FDG PET for other reasons than suspected AGI (33). A diffuse homogeneous uptake was observed in 67 grafts (63%) and heterogeneous uptake was observed in 31 grafts (29%). Nine grafts (8%) demonstrated no ^{18}F -FDG uptake and none of the grafts displayed focal ^{18}F -FDG uptake. Keidar et al. hypothesized that diffuse ^{18}F -FDG PET uptake in non-infected grafts is a result of a local sterile inflammatory process around the prosthesis due to a foreign body-related reaction, and also related to the type of implanted material. Moreover, Berger et al. found the mentioned standard parameters largely to overlap in infected and uninfected central vascular grafts (34), which confirms our findings that conventional parameters are not sufficient in distinguishing these groups.

The current study demonstrated the association of high and heterogeneous ^{18}F -FDG uptake with AGI; however, due to the relatively small patient cohort it seems not applicable for clinical decision making yet. Nevertheless, this finding is of utmost importance, since it warrants studies with larger, prospective patient cohorts with the construction of multivariable prediction models. Such studies should focus on the textural features, which were found robust for inter-observer variability in the current study. One of the most important limitations in testing the utility of textural analysis in the diagnosis of AGI is the fact that 80% of the patients received antibiotic therapy at the time of scanning, which may have yielded false-negative cultures. However, this bias is the result of the clinical reality, since the omission of antibiotic therapy could adversely affect patients and this issue is related to each study predicting AGI. Moreover, routine microbiologic techniques can sometimes fail to isolate the microorganism from perigraft material (35, 36). Sonication techniques have been described to identify indolent gram-positive microorganisms by using ultrasound energy to agitate particles from the graft sample for microbiology (35, 36). However, we did not apply these sonication techniques in this retrospective study because they were not available in our hospital. In addition, the used definition for VOI leads to analysis on the whole prosthetic graft volume, which possibly results in an underestimation of the predictive value of the textural features. However, as we mentioned earlier, we chose for this definition since analysis of the most-diseased segment of the aortic graft would require a semi-objective identification of the ^{18}F -FDG-avid area (13) and since group II does not contain such an area.

CONCLUSION

Textural analysis to characterize ^{18}F -FDG uptake heterogeneity is feasible and shows promising results in diagnosing AGI and can encourage further research to facilitate implementation of automated textural analysis algorithms into clinical practice. Further research regarding the construction, refinement, and validation of prediction models in larger prospective cohorts is required before it can be implemented in the clinical decision-making process.

REFERENCES

1. Valentine RJ. Diagnosis and management of aortic graft infection. *Semin Vasc Surg.* 2001;14:292-301.
2. O'Connor S, Andrew P, Batt M, Becquemin JP. A systematic review and meta-analysis of treatments for aortic graft infection. *J Vasc Surg.* 2006;44:38-45.
3. Perera GB, Fujitani RM, Kubaska SM. Aortic graft infection: update on management and treatment options. *Vasc Endovascular Surg.* 2006;40:1-10.
4. Saleem BR, Meerwaldt R, Tielliu IF, Verhoeven EL, van den Dungen, J J, Zeebregts CJ. Conservative treatment of vascular prosthetic graft infection is associated with high mortality. *Am J Surg.* 2010;200:47-52.

5. Legout L, D'Elia PV, Sarraz-Bournet B, et al. Diagnosis and management of prosthetic vascular graft infections. *Med Mal Infect.* 2012;42:102-109.
6. Fukuchi K, Ishida Y, Higashi M, et al. Detection of aortic graft infection by fluorodeoxyglucose positron emission tomography: comparison with computed tomographic findings. *J Vasc Surg.* 2005;42:919-925.
7. Keidar Z, Pirmisashvili N, Leiderman M, Nitecki S, Israel O. ^{18}F -FDG uptake in noninfected prosthetic vascular grafts: incidence, patterns, and changes over time. *J Nucl Med.* 2014;55:392-395.
8. Keidar Z, Engel A, Hoffman A, Israel O, Nitecki S. Prosthetic vascular graft infection: the role of ^{18}F -FDG PET/CT. *J Nucl Med.* 2007;48:1230-1236.
9. Lauwers P, Van den Broeck S, Carp L, Hendriks J, Van Schil P, Blockx P. The use of positron emission tomography with (18)F-fluorodeoxyglucose for the diagnosis of vascular graft infection. *Angiology.* ;58:717-724.
10. Spacek M, Belohlavek O, Votrubova J, Sebesta P, Stadler P. Diagnostics of "non-acute" vascular prosthesis infection using ^{18}F -FDG PET/CT: our experience with 96 prostheses. *Eur J Nucl Med Mol Imaging.* 2009;36:850-858.
11. Bruggink JL, Glaudemans AW, Saleem BR, et al. Accuracy of FDG-PET-CT in the diagnostic work-up of vascular prosthetic graft infection. *Eur J Vasc Endovasc Surg.* 2010;40:348-354.
12. Tokuda Y, Oshima H, Araki Y, et al. Detection of thoracic aortic prosthetic graft infection with ^{18}F -fluorodeoxyglucose positron emission tomography/computed tomography. *Eur J Cardiothorac Surg.* 2013;43:1183-1187.
13. Boellaard R, O'Doherty MJ, Weber WA, et al. FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0. *Eur J Nucl Med Mol Imaging.* 2010;37:181-200.
14. Saleem BR, Berger P, Vaartjes I, et al. Modest utility of quantitative measures in ^{18}F -fluorodeoxyglucose positron emission tomography scanning for the diagnosis of aortic prosthetic graft infection. *J Vasc Surg.* 2015;61:965-971.
15. Eary JF, O'Sullivan F, O'Sullivan J, Conrad EU. Spatial heterogeneity in sarcoma ^{18}F -FDG uptake as a predictor of patient outcome. *J Nucl Med.* 2008;49:1973-1979.
16. Yu H, Caldwell C, Mah K, et al. Automated radiation targeting in head-and-neck cancer using region-based texture analysis of PET and CT images. *Int J Radiat Oncol Biol Phys.* 2009;75:618-625.
17. Yu H, Caldwell C, Mah K, Mozeg D. Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning. *IEEE Trans Med Imaging.* 2009;28:374-383.
18. El Naqa I, Grigsby P, Apte A, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.* 2009;42:1162-1171.
19. van Velden FH, Cheebsumon P, Yaqub M, et al. Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies. *Eur J Nucl Med Mol Imaging.* 2011;38:1636-1647.
20. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline ^{18}F -FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med.* 2011;52:369-378.
21. Vaidya M, Creach KM, Frye J, Dehdashti F, Bradley JD, El Naqa I. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. *Radiother Oncol.* 2012;102:239-245.
22. Tan S, Kligerman S, Chen W, et al. Spatial-temporal [^{18}F]FDG-PET features for predicting pathologic response of esophageal cancer to neoadjuvant chemoradiation therapy. *Int J Radiat Oncol Biol Phys.* 2013;85:1375-1382.
23. Zhang H, Tan S, Chen W, et al. Modeling pathologic response of esophageal cancer to chemoradiation therapy using spatial-temporal ^{18}F -FDG PET features, clinical parameters, and demographics. *Int J Radiat Oncol Biol Phys.* 2014;88:195-203.
24. Rutherford RB, Baker JD, Ernst C, et al. Recommended standards for reports dealing with lower extremity ischemia: revised version. *J Vasc Surg.* 1997;26:517-538.
25. Saleem BR, Pol RA, Slart RH, Reijnen MM, Zeebregts CJ. ^{18}F -Fluorodeoxyglucose positron emission tomography/CT scanning in diagnosing vascular prosthetic graft infection. *Biomed Res Int.* 2014;2014:471971.
26. Doane DP. Aesthetic Frequency Classifications. *The American Statistician.* 1976;30:181-183.
27. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern.* 1973;3:610-621.

28. Galloway MM. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*. 1975;4:172-179.
29. Thibault G, Angulo J, Meyer F. Advanced statistical matrices for texture characterization: application to cell classification. *IEEE Trans Biomed Eng*. 2014;61:630-637.
30. Akaike H. A new look at the statistical model identification. *TAC*. 1974;19:716-723.
31. Steyerberg EW. *Clinical Prediction Models*. New York, NY: Springer; 2009.
32. Cicchetti DV. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychol Assess*. 1994;6:284-90.
33. Keidar Z, Nitecki S. FDG-PET in prosthetic graft infections. *Semin Nucl Med*. 2013;43:396-402.
34. Berger P, Vaartjes I, Scholtens A, et al. Differential FDG-PET Uptake Patterns in Uninfected and Infected Central Prosthetic Vascular Grafts. *Eur J Vasc Endovasc Surg*. 2015;50:376-383.
35. Bergamini TM, Bandyk DF, Govostis D, Vetsch R, Towne JB. Identification of *Staphylococcus epidermidis* vascular graft infections: a comparison of culture techniques. *J Vasc Surg*. 1989;9:665-670.
36. Tollefson DF, Bandyk DF, Kaebnick HW, Seabrook GR, Towne JB. Surface biofilm disruption. Enhanced recovery of microorganisms from vascular prostheses. *Arch Surg*. 1987;122:38-43.

Chapter 8

**Summary, General Discussion, and
Future Perspectives**

ABBREVIATIONS

¹⁸ F-FAZA	¹⁸ F-Fluoroazomycin-arabinoside
¹⁸ F-FDG PET	¹⁸ F-fluorodeoxyglucose positron emission tomography
¹⁸ F-FMISO	¹⁸ F-Fluoromisonidazole
⁶⁴ Cu-ATSM	⁶⁴ Cu-diacetyl-bis(N4-methylthiosemicarbazone)
AGI	aorta-iliac graft infections
CD44	cluster of differentiation 44
CROSS	ChemoRadiotherapy for Oesophageal cancer followed by Surgery Study
CT	computed tomography
EC	esophageal cancer
EMT	epithelial-mesenchymal transition
HER2	human epidermal growth factor receptor 2
HIF1α	hypoxia-inducible factor alpha
IBSI	image biomarker standardisation initiative
MRI	magnetic resonance imaging
nCRT	neoadjuvant chemoradiotherapy
pCR	pathologic complete response
PTCH1	patched homolog 1
SHH	Sonic Hedgehog
SPECT	single-photon emission computed tomography
SUV	standardized uptake value

SUMMARY

Part I: Validation of ^{18}F -FDG PET Radiomic Features

In **Part I**, we investigated the suitability of the most commonly used radiomic features derived from ^{18}F -fluorodeoxyglucose positron emission tomography (^{18}F -FDG PET) as new potential biomarkers. In the last decade, numerous radiomic studies have been published regarding the role of radiomics in the clinical decision-making. Although the majority of the published studies have reported promising clinical results, the lack of standardization in the feature extraction process complicates the validation of those results and the implementation of these features in common practice. In addition, other important challenges facing radiomics are the establishment of feature measurement errors and the reduction of the multitude of features that are generated.

In **Chapter 2**, the image biomarker standardisation initiative (IBSI) was launched to work towards standardization in quantitative image analysis and radiomics. Software implementations of 21 participating teams were iteratively compared over time against digital phantom and computed tomography (CT) data. The large initial discrepancy across the different radiomic feature definitions and implementations were resolved by the IBSI. Based on these investigations, the IBSI provided a reference manual, containing nomenclature and definitions for the most commonly used radiomic features, an image processing scheme, benchmark data sets and associated values, and a set of reporting guidelines. This manual constitutes an important tool available to the community to increase reproducibility and generalizability.

In **Chapter 3**, we explored how dimensionality reduction and reliability of ^{18}F -FDG PET radiomic features are affected by the underlying data, image reconstruction methods and settings, noise, discretization method, and delineation method. For every combination of these settings, ten statistically equal replicates were generated of the NEMA NU 2-2012 image quality phantom. Feature space reduction was found to be repeatable, but was sensitive to discretization, sphere size, and activity uptake, and is therefore only generalizable among studies using the same settings. Reliability is an important measurement of error and was obtained using the replicates of each combination of settings. Based on these experiments, we demonstrated that clinical PET studies and examinations need to be standardized in order to use ^{18}F -FDG PET radiomics as quantitative imaging biomarkers. In particular, efforts should focus on noise reduction even at the cost of spatial resolution and optimizing the choice for image reconstruction method, discretization method, and segmentation method. A separate validation is recommended for every type of PET study.

Part II: Applications of ^{18}F -FDG PET/CT Radiomics in Oncology

In **Part II**, we evaluated potential applications of quantitative ^{18}F -FDG PET/CT imaging in esophageal cancer (EC). In the Netherlands, the current standard treatment for patients with locally advanced ($T_1/N_{1-3}/M_0$; $T_{2-4a}/N_{0-3}/M_0$) EC consists of esophagectomy with curative intent 6-8 weeks after neoadjuvant chemoradiotherapy (nCRT; carboplatin/paclitaxel/41.4 Gy) according to the ChemoRadiotherapy for Oesophageal cancer followed by Surgery Study (CROSS) regimen. Prediction of response to nCRT is an important focus of research as a challenging and risky esophagectomy may be refrained for complete responders to nCRT. In this part, we have constructed candidate prediction models for nCRT response based on clinical parameters and quantitative ^{18}F -FDG PET/CT radiomic features. We have demonstrated that radiomic features can effectively be used to characterize underlying tumoral pathophysiological causes and may play an important future role in the prediction of pathologic complete response (pCR).

In **Chapter 4**, we reviewed 97 patients with locally advanced EC to construct 6 candidate models to predict pCR to nCRT based on pre-nCRT clinical parameters and ^{18}F -FDG PET/CT-derived radiomic features. A feature selection procedure selected the clinical parameters histologic type and clinical T stage, the ^{18}F -FDG PET-derived textural feature long run low gray level emphasis, and the CT-derived textural feature run percentage. The candidate prediction models incorporated different combinations of these variables. The prediction models had a higher discriminatory accuracy than the current standard method, based on maximum standardized uptake value ($\text{SUV}_{\max}$) measurements. These results demonstrate a high incremental value of ^{18}F -FDG PET/CT quantification and can be considered as an initial step in predicting pCR to nCRT in patients with EC.

In **Chapter 5**, the quantitative PET-based prediction strategy of Chapter 4 was expanded by analyzing radiomic features acquired from pre-nCRT and post-nCRT ^{18}F -FDG PET scans of 73 patients. Having pre- and post-nCRT scans available enables close follow-up of tumor response during treatment and extraction of radiomic feature changes and histogram distances to assess nCRT-induced changes in ^{18}F -FDG uptake distribution. We identified a logistic regression prediction model consisting of traditional clinical T-staging and post-nCRT joint maximum. Joint maximum is a measure of orderliness, which quantifies the systematic arrangement of voxel intensity differences. We found that patients with high post-nCRT ^{18}F -FDG PET orderliness showed an increased probability for achieving pathologic complete response. This model yielded substantially higher discriminatory accuracies in phenotyping pCR in EC patients than the pre-defined reference prediction models based on conventional $\text{SUV}_{\max}$ measurements, clinical parameters, and clinical parameters along with pre-nCRT radiomic features. Moreover, we found that post-nCRT joint maximum was

redundant with the post-nCRT radiomic features median absolute deviation, joint entropy, sum entropy, angular second moment, and inverse variance. In the multivariable model, post-nCRT joint maximum could be exchanged by each of these radiomic features, while maintaining the performance of the prediction model.

In **Chapter 6**, we integrated the quantitative PET/CT-based prediction strategy into a multimodal prediction strategy. Pre-nCRT tumor biopsies of 43 patients with locally advanced EC were evaluated by immunohistochemistry for the potential biological tumor markers human epidermal growth factor receptor 2 (HER2), cluster of differentiation 44 (CD44), hypoxia-inducible factor alpha (HIF1 α), patched homolog 1 (PTCH1), and Sonic Hedgehog (SHH). Of all studied biological tumor markers, we only found a significant association between pCR and HER2. Conformable to these results, the goodness-of-fit and discriminatory accuracy of the clinico-radiomic prediction models described in Chapter 4 were substantially improved by the incorporation of HER2. The remaining studied biological tumor markers did not result in any model improvement. These results demonstrate that integration of multimodal biomarkers can enhance the discriminatory accuracy of predicting response to nCRT in EC patients.

Part III: Applications of ^{18}F -FDG PET Radiomics in Infectious Disorders

In **Part III**, we have evaluated the potential applications of quantitative ^{18}F -FDG PET imaging as a surrogate for invasive diagnostic procedures in infected graft reconstructions in occlusive vascular diseases. Obtaining definite proof of infection is crucial after vascular graft reconstruction and remains a major challenge due to the lack of evidence-based PET interpretation. Although most research on radiomics has focused on clinical oncological applications, we demonstrated that quantifying heterogeneous ^{18}F -FDG uptake using textural analysis also has a potential application in detecting vascular graft infections. In this thesis, we focused on quantifying aorta-iliac graft infections (AGI).

In **Chapter 7**, we evaluated 30 patients with a history of aorta-iliac graft reconstruction after aneurysmal or occlusive disease. Patients were screened to identify those with a clinical suspicion of an AGI. We found that the conventional measures SUV_{max} , tissue-to-background ratio, and the visual grading scale were able to distinguish between clinically suspected and non-suspected (control group) graft infections. The selected textural features did not have complementary value beyond these conventional measures in distinguishing these groups. However, within the patient group suspected for AGI, we found a limited discriminatory value for the conventional measures and identified that the textural feature short run high gray level emphasis was able to distinguish between patients with positive and negative bacterial cultures. Quantitative PET may therefore be a potential tool in improving PET interpretation of patients suspected of AGI.

GENERAL DISCUSSION

Medical imaging has been one of the most important advances in modern medicine and has become essential for diagnosis, treatment, and prognosis for almost all major medical conditions and diseases (1). Traditionally, medical imaging has been evaluated visually, however, this approach is subjective and fails to identify beneficial information such as heterogeneity, which cannot readily be detected by visual assessment. Quantifying heterogeneity has gained clinical interest as it may be associated with patient outcome and subsequently facilitate the clinical decision-making process. Therefore, additional techniques are required which allow reproducible and objective assessment and overcome the physician's subjective assessment. Heterogeneity of voxel intensities can be quantified by numerous techniques including textural analysis quantifying the spatial arrangement of voxel intensities (2-5), fractal techniques and Minkowski functionals measuring repetitive geometrical patterns (6), shape models measuring the deviation of the ^{18}F -FDG distribution from a homogeneous ellipsoid (7), intensity histogram analysis plotting the volume as a function of intensity that is equal to or higher than a certain value (8), the heterogeneity factor defined as the derivative (dV/dT) of a volume-threshold function (9), and parcellation techniques using multiple imaging parameters to identify subregions with common underlying biological properties (6).

The aim of this thesis was to apply quantitative radiomic features to objectively evaluate and adjust the clinical decision-making in the treatment of patients with esophageal cancer and vascular graft infections. The emerging field of radiomics extracts a large number of quantitative (morphological, intensity, and textural) image features to quantify the spatial arrangement and heterogeneity of voxel intensities. These features may have a prognostic and/or predictive impact on the outcomes. While prognostic features can stratify the likelihood of a particular outcome, predictive features can predict response to a particular treatment.

Despite these promising applications, the initial lack of standardization for feature extraction and image processing complicated the comparison and validation of different study results. The complex radiomics workflow includes a sequence of operations and related methodological choices concerning the pre-processing of imaging data, such as functional volume segmentation methods, interpolation algorithms to isotropic voxels, and approaches to discretization of voxel intensities. In this thesis, these issues were resolved by reaching consensus-based definitions and recommendations for calculating image biomarkers in an international collaboration. Moreover, we exposed the mechanisms of confounding factors which drive the reduction and reliability of ^{18}F -FDG PET radiomic features. The described mechanisms need to be validated in clinical data as phantom data may not be representative for patient data. It was concluded that standardization of clinical PET/CT studies and

examinations is required in order to use ^{18}F -FDG PET radiomics as quantitative imaging biomarkers. This PET imaging standard for radiomic studies needs to be established based on the mechanisms which drive the reliability of ^{18}F -FDG PET radiomic features. Using this imaging standard, a set of reliable radiomic features should be identified, which can be tested for its clinical relevance. To facilitate a more intuitive clinical interpretation, this set needs to be translated to qualitative imaging properties which are familiar to radiologists.

Subsequently, we demonstrated potential clinical applications for radiomics with state-of-the-art imaging. Several prediction models were constructed based on ^{18}F -FDG PET/CT radiomics along with clinical markers and biological tumor markers to predict the response of nCRT in EC. Some of these prediction models were able to provide up to a 30% absolute improvement in discriminatory accuracy beyond the current standard prediction methods. These prediction models are easy to apply for clinicians by completing a single formula. The formula provides the probability of achieving a pathologic complete response on a per-patient level, and thereby adds potentially useful quantitative information to the visual image assessment. From all prediction models constructed in this thesis, the restaging model of Chapter 5 yielded the highest internally validated discrimination. Quantitative restaging PET/CT therefore appears to be more valuable than quantitative baseline PET/CT for predicting pCR after nCRT. Moreover, it is of interest how incorporating HER2 into this model will affect the model's performance. Despite these promising results, this benefit did not yet translate into a change in the clinical decision-making process. Although all prediction models were internally validated using resampling techniques like cross-validation and bootstrapping, the constructed prediction models lack validation in external patient cohorts. Prediction models can only be constructed and validated in a reliable fashion with the availability of large patient cohorts or big data, but for specific patient groups imaging data are limited available. As a result, (international) initiatives of imaging data-sharing and pooling of relevant patient data are required to move the field forward. Yet it is worth mentioning that the clinical benefit of radiomics using big data is only fully appreciated in uniform data. The generation of used imaging methods over the course of time is therefore also of interest. Big data can facilitate robust statistical analyses such as machine learning, but unfortunately this approach has rarely been used (10). Besides, to further improve the discriminatory accuracy, other relevant predictors should be incorporated into the prediction models such as quantitative image features extracted from magnetic resonance imaging (MRI). So far, most research on radiomics has focused on the prediction of complete response to nCRT. The role of radiomics in prognosis and prediction of non-response to nCRT should therefore be further investigated in order to provide further support in the clinical decision making. A reliable prediction of non-response may result in omission of treatment and the avoidance of potentially harmful treatment-induced damage. Evaluation of the change in radiomic

features between baseline and interim scanning may be a promising tool to assess treatment response early in the course of treatment. Besides, the use of these so-called delta-radiomics may enable the assessment of the effect of different types of radiotherapy (photon and proton irradiation).

Furthermore, ^{18}F -FDG PET radiomic features were applied to improve the non-invasive diagnosis of AGI after vascular graft reconstructions. In a pilot study, we have found a limited discriminatory value for the conventional measures and identified a radiomic feature which was able to distinguish between patients suspected for AGI with positive and negative bacterial cultures. Quantitative PET may therefore be a potential tool in improving PET interpretation of patients suspected of AGI. These conclusions were based on a small patient cohort, which complicated the construction of multivariable prediction models. As a result, there was insufficient evidence to support the decision-making. Larger external datasets should reveal whether these results can be verified. This study should mainly be interpreted as a proof-of-principle and introduction of radiomics in the field of diagnosing inflammation and infection.

FUTURE PERSPECTIVES

Radiomics provides information on underlying biology, tumor phenotype, and its microenvironment, which may complement traditional patient characteristics, clinical evaluation, visual medical image evaluation, laboratory tests, and genomic assays. We expect that relevant quantitative image features eventually will be extracted along with standard-of-care images and will be integrated with other available patient information to improve diagnostic, predictive, and prognostic accuracy. The availability of quantitative data of thousands of patients will lead to more robust prediction models, which can support future clinical decision-making. In a subsequent section, we will highlight some **potential future applications of radiomics**.

Other Radiotracers and Modalities

The vast majority of the current available studies regarding radiomics have focused on ^{18}F -FDG PET and CT images. However, ^{18}F -FDG PET is characterized by low specificity because it accumulates in malignant, infectious, and inflammatory tissues. Characterizing different functional tissue properties with more specific PET radiotracers may therefore lead to more specific and more meaningful clinical data to construct predictive models. For instance, evaluation of hypoxia in the tumor and its microenvironment is of clinical interest because the concentration of free oxygen radicals is considered one of multiple factors related to the effectiveness of radiation treatment. Hypoxia in the tumor microenvironment

is predominantly caused by molecular mechanisms leading to upregulation of aerobic glycolysis. A number of tracers have been developed to evaluate tumoral hypoxia, including ^{18}F -Fluoromisonidazole (^{18}F -FMISO), ^{18}F -Fluoroazomycin-arabinoide (^{18}F -FAZA), and ^{64}Cu -diacetyl-bis(N4-methylthiosemicarbazone) (^{64}Cu -ATSM), however each of these PET tracers has still its own advantages and pitfalls. Tracing specific functional tissue properties may also be promising in imaging inflammation and infection. For instance, $^{99\text{m}}\text{Tc}$ -labeled white blood cell single-photon emission computed tomography (SPECT) imaging has a higher specificity than ^{18}F -FDG PET imaging. However, this time-consuming study is accompanied with a lower resolution and relative high costs for tagging white blood cells. Other promising alternatives include infection tracers such as ^{18}F -fluorodeoxyisorbital (^{18}F -FDS) and specific bacteria-targeted SPECT tracers such as $^{99\text{m}}\text{Tc}$ -ciprofloxacin (Infecton®) (11, 12), however these tracers are still in an experimental phase. Moreover, quantitative measurements based on other modalities such as multi-sequence MRI (e.g. T1, T2, diffusion weighted imaging, and dynamic contrast-enhanced imaging) may possibly lead to an improved tissue characterization. Using these sequences, the perfusion and diffusion of tissues can be imaged, which provides valuable information about the functional architecture of the tissue. Besides, recent advances have now led to the introduction of fully integrated single gantry PET-MRI systems with increased spatial resolution. Hybrid PET/MRI systems result in less patient motion artifacts compared to sequential PET-MRI. This improves spatial registration of PET/MRI and consequently improves volume of interest definition.

Radiogenomics-based Diagnosis

Radiogenomics is a field of study, which hypothesizes that quantitative image features have an underlying association with genomic data or clinical factors. It attempts to reveal radiogenomic atlases to define the biological basis of radiomic features. Equivalent to cancer genomics projects such as The Cancer Genome Atlas (13), multi-dimensional maps of radiomic features can be generated quantifying tumor phenotypes which are captured in medical images. Similar to Cancer Radiomics Atlases, Infection Radiomics Atlases may be generated to discriminate between infections and inflammation. Radiogenomic atlases may guide the diagnosis of infections and cancer by guiding the selection to identify certain gene expression alterations in biopsy material and by confirming histopathologic findings to reduce substantial sampling errors and observer variability (14). Unfortunately, these relationships are largely undiscovered and only a limited amount of studies demonstrated a correlation between CT, PET, or MRI radiomic features and underlying biological rationale (15-20).

Patch-wise Radiomics

The heterogeneous nature of tumors often leads to incomplete characterization of the total tumor burden and inadequate identification of local tumor extension. Intratumoral heterogeneity causes the molecular tumor profile of biopsies to be depended on the sample, which may therefore not represent the entire tumor. In addition, inadequate identification of the primary tumor may yield microscopic tumor outside the radiotherapeutic clinical target volume and irradiated resections. These issues may be addressed by computing patch-wise radiomic features allowing non-invasive analysis of the total tumor burden on a per-voxel basis. For this technique, the radiomic feature is calculated over a moving window with a predefined size. This provides radiomic feature based images which may allow improved identification of the degree of local tumor extension and intratumoral heterogeneous areas. These images can therefore be useful in the guidance of biopsies or resections and have already led to first attempts to facilitate a radiomics-based decision support framework for targeted radiation treatment planning (21).

Quantifying the Environment

Given that tumors and infections constantly interact with the surrounding microenvironment, the microenvironment is substantially related to the prognosis of the patient and would therefore be an interesting radiomics target. Modulation of immune recognition and immune response creates a distinct tumor microenvironment, including angiogenesis, increased vascular permeability, cytokine release, and lymphocyte infiltration (22). The tumor microenvironment promotes progression and metastasis due to tumor cell migration and may induce the occurrence of epithelial-mesenchymal transition (EMT) in tumor cells. During this EMT process, cancer cells lose their polarity and acquire the invasive mesenchymal phenotype of the tumor. We hypothesize that this process can be captured by radiologic characteristics and quantitative image features such as radiomics. Another radiomic application to quantify the environment is the characterization of structure, shape, and composition of organs at risk. Radiomics have shown to improve prediction of radiation-induced morbidity and may therefore affect radiation treatment plans and corresponding organ dose distributions (23).

REFERENCES

1. Editorial. Looking Back on the Millennium in Medicine. *The New England Journal of Medicine*. 2000;342:42-49.
2. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. *IEEE Trans Syst Man Cybern*. 1973;3:610-621.
3. Galloway MM. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*. 1975;4:172-179.

4. Thibault G, Angulo J, Meyer F. Advanced statistical matrices for texture characterization: application to cell classification. *IEEE Trans Biomed Eng.* 2014;61:630-637.
5. Amadasun M, King R. Textural features corresponding to textural properties. *IEEE Trans Syst Man Cybern.* 1989;19:1264-1274.
6. Alic L, Niessen WJ, Veenland JF. Quantification of heterogeneity as a biomarker in tumor imaging: a systematic review. *PLoS ONE.* 2014;9:e110300.
7. O'Sullivan F, Roy S, O'Sullivan J, Vernon C, Eary J. Incorporation of tumor shape into an assessment of spatial heterogeneity for human sarcomas imaged with FDG-PET. *Biostatistics.* 2005;6:293-301.
8. El Naqa I, Grigsby P, Apte A, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.* 2009;42:1162-1171.
9. Kidd EA, Grigsby PW. Intratumoral metabolic heterogeneity of cervical cancer. *Clin Cancer Res.* 2008;14:5236-5241.
10. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep.* 2015;5:13087.
11. Weinstein EA, Ordonez AA, DeMarco VP, et al. Imaging Enterobacteriaceae infection in vivo with 18F-fluorodeoxyisobutyl positron emission tomography. *Sci Transl Med.* 2014;6:259ra146.
12. Britton KE, Wareham DW, Das SS, et al. Imaging bacterial infection with (99m)Tc-ciprofloxacin (Infecton). *J Clin Pathol.* 2002;55:817-823.
13. Anonymous The Cancer Genome Atlas homepage. .
14. Manion E, Cohen MB, Weydert J. Mandatory second opinion in surgical pathology referral material: clinical consequences of major disagreements. *Am J Surg Pathol.* 2008;32:732-737.
15. Gevaert O, Mitchell LA, Achrol AS, et al. Glioblastoma Multiforme: Exploratory Radiogenomic Analysis by Using Quantitative Image Features. *Radiology.* 2015;276:313.
16. Segal E, Sirlin CB, Ooi C, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat Biotechnol.* 2007;25:675-680.
17. Wan T, Bloch BN, Plecha D, et al. A Radio-genomics Approach for Identifying High Risk Estrogen Receptor-positive Breast Cancers on DCE-MRI: Preliminary Results in Predicting OncotypeDX Risk Scores. *Sci Rep.* 2016;6:21394.
18. Itakura H, Achrol AS, Mitchell LA, et al. Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. *Sci Transl Med.* 2015;7:303ra138.
19. Grossmann P, Gutman DA, Dunn WD, Holder CA, Aerts HJW. Imaging-genomics reveals driving pathways of MRI derived volumetric tumor phenotype features in Glioblastoma. *BMC Cancer.* 2016;16.
20. Yip SS, Kim J, Coroller T, et al. Associations between somatic mutations and metabolic imaging phenotypes in non-small cell lung cancer. *Journal of Nuclear Medicine.* 2016;58:569.
21. Shiradkar R, Podder TK, Algohary A, Viswanath S, Ellis RJ, Madabhushi A. Radiomics based targeted radiotherapy planning (Rad-TRaP): a computational framework for prostate cancer treatment planning with MRI. *Radiat Oncol.* 2016;11:148.
22. Chad Tang, Brian Hobbs, Ahmed Amer, et al. Development of an Immune-Pathology Informed Radiomics Model for Non-Small Cell Lung Cancer. *Scientific Reports (Nature Publisher Group).* 2018;8:1-9.
23. van Dijk LV, Brouwer CL, van der Schaaf A, et al. CT image biomarkers to improve patient-specific prediction of radiation-induced xerostomia and sticky saliva. *Radiother Oncol.* 2017;122:185-191.

Chapter 9

**Nederlandse Samenvatting/Dutch
Summary**

AFKORTINGEN

^{18}F -FDG PET	^{18}F -fluorodeoxyglucose positron emission tomography
CROSS	ChemoRadiotherapy for Oesophageal cancer followed by Surgery Study
CT	computed tomography
HER2	human epidermal growth factor receptor 2
IBSI	image biomarker standardisation initiative
nCRT	neoadjuvante chemoradiotherapie
SUV_{max}	maximale standardized uptake value

In de laatste decennia is het duidelijk geworden dat medische behandelingen steeds meer op het individu toegesneden moeten worden. Vooral in de oncologische zorg is het individualiseren van specifieke kankerbehandelingen één van de grote uitdagingen. Deze aanpak is mogelijk door betere fenotypering en analyse van genetische mutaties die verantwoordelijk zijn voor de resistentie van bepaalde behandelingen en de daaraan gerelateerde nieuwe behandelingsvormen. Beeldvormende technieken zoals de computed tomography (CT), ¹⁸F-fluorodeoxyglucose positron emission tomography (¹⁸F-FDG PET) of hybride PET/CT spelen hierbij mogelijk een belangrijke rol. De klinische beoordeling van de diagnostische beeldvorming is in veel gevallen toereikend, maar voor de beoordeling en voorspelling van respons op de gegeven behandeling en het verbeteren van diagnose en prognose is een vorm van kwantificatie (metingen aan PET/CT) nodig. De klinische beoordeling gaat vaak gepaard met variatie in de beoordeling tussen waarnemers. Kwantificatie van PET/CT biedt hierin een mogelijke oplossing door het leveren van reproduceerbare en objectieve informatie. Bovendien kunnen deze metingen informatie opleveren die buiten de menselijke waarneming valt. Radiomics is een opkomende kwantificatie techniek die door middel van grote hoeveelheden beeldvormende karakteristieken (*radiomic features*) onderliggende pathofysiologische eigenschappen beschrijven. Radiomic features kunnen worden onderverdeeld in een aantal categorieën waarvan de statistische, de lokale intensiteit, de morfologische en de textural features de meest voorkomende zijn. De wetenschappelijke literatuur suggereert dat met name het kwantificeren van biologische heterogeniteit op beeldvorming met textural features bruikbaar is voor verschillende klinische doeleinden. Zo tonen tumoren een uitgebreide variatie aan genotypes en fenotypes (tumor heterogeniteit), dat uiteindelijk leidt tot het mogelijk falen van kankerbehandelingen en een slechte prognose. Dit proefschrift gaat in op de validatie en verschillende toepassingen van ¹⁸F-FDG PET/CT radiomic features.

DEEL I: VALIDATIE VAN ¹⁸F-FDG PET RADIOMIC FEATURES

In **Deel I** van dit proefschrift werd onderzocht of de meest gebruikte ¹⁸F-FDG PET radiomic features geschikt zijn om te fungeren als nieuwe potentiële biomarkers (radiomic gebaseerde biomarkers). De laatste jaren zijn er tal van radiomic studies gepubliceerd naar een mogelijke rol in de ondersteuning van de huidige klinische besluitvorming. Hoewel de meerderheid van de gepubliceerde studies veelbelovende resultaten vermeldt, wordt de validatie van die resultaten bemoeilijkt door het gebrek aan standaardisatie in het feature extractie proces. Voordat de klinische relevantie vastgesteld kan worden, is het daarnaast van belang om de betrouwbaarheid van deze radiomic features vast te stellen en het grote aantal radiomic features naar een kleiner aantal niet-correlerende, relevante features te reduceren.

Hoofdstuk 2 betreft de standaardisatie van radiomic features door het *image biomarker standardisation initiative* (IBSI), een internationaal samenwerkingsverband. Software implementaties van 21 deelnemende teams werden vergeleken op basis van digitale fantoom en CT-data. Het IBSI verminderde de initiële grote discrepantie in feature definities en implementaties. Gebaseerd op deze resultaten bracht het IBSI een handleiding uit met naamgeving en definities van de meest voorkomende radiomic features, een schema voor beeldbewerking, benchmark data sets met alle corresponderende feature waarden en rapportage richtlijnen. Deze handleiding vormt een belangrijk instrument om de reproduceerbaarheid van op radiomic features gebaseerde modellen te verbeteren.

In **Hoofdstuk 3** werd de invloed onderzocht van het te scannen object, beeldreconstructie instellingen, ruis, discretisatie methode en segmentatie methode op de reductie en betrouwbaarheid van ^{18}F -FDG PET radiomic features. De experimenten werden uitgevoerd met een fantoom met zes bolvormige elementen van variabele grootte die gevuld werden met verschillende concentraties ^{18}F -FDG. Het reduceren van het aantal features was herhaalbaar, maar bleek wel gevoelig te zijn voor discretisatie, bolgrootte en concentratie ^{18}F -FDG, waardoor feature reductie alleen reproduceerbaar lijkt te zijn op voorwaarde van gelijke scan omstandigheden. Op basis van deze experimenten blijkt dat PET beelden moeten worden gestandaardiseerd om ^{18}F -FDG PET-radiomics te gebruiken als biomarkers voor kwantitatieve beeldvorming. De betrouwbaarheid van ^{18}F -FDG PET radiomic features kon worden vergroot door men zich te richten op ruisonderdrukking, ook al gaat dat ten koste van de daaraan gekoppelde spatiële resolutie. Bovendien zou de keuze voor beeldreconstructie, discretisatie en segmentatie voor elk type PET-onderzoek geoptimaliseerd moeten worden. Voor elk type PET-onderzoek wordt daarnaast een afzonderlijke validatie aanbevolen, aangezien de resultaten afhangen van het te scannen object.

DEEL II: ONCOLOGISCHE TOEPASSINGEN VAN ^{18}F -FDG PET RADIOMIC FEATURES

In **Deel II** van dit proefschrift werden toepassingen onderzocht van ^{18}F -FDG PET radiomics bij de behandeling van patiënten met een slokdarmcarcinoom. In Nederland bestaat de standaardbehandeling voor patiënten met een lokaal uitgebreid slokdarmcarcinoom uit een voorbehandeling gevolgd door een slokdarmresectie 6-8 weken na beëindiging van de voorbehandeling. Deze voorbehandeling, de neoadjuvante chemoradiotherapie (nCRT), bestaat uit een combinatie van chemotherapie (carboplatin en paclitaxel) en externe radiotherapie (41.4 Gy) volgens het schema van de *ChemoRadiotherapy for Oesophageal cancer followed by Surgery Study* (CROSS) studie. Ondanks dat deze behandeling gepaard gaat met een significante verbetering in overleving, hebben patiënten die geen vitaal tumorweefsel hebben na nCRT geen baat bij een aan verschillende complicaties onderhevige slokdarmresectie.

Helaas kunnen deze “complete responders” pas definitief worden geïdentificeerd ná chirurgie en pathologische beoordeling van het resectiepreparaat (pathologisch complete responders). Dit deel van het proefschrift richt zich op de voorspelling van pathologisch complete respons na nCRT om zo een onnodige risicovolle slokdarmresectie te voorkomen.

In **Hoofdstuk 4** ontwikkelden we 6 modellen voor de voorspelling van pathologisch complete respons op basis van 97 patiënten met een lokaal uitgebreid slokdarmcarcinoom. De predictiemodellen waren gebaseerd op klinische parameters en baseline (pre-nCRT) radiomic features. Het histologisch subtype en het klinische T-stadium (van de tumor-node-metastasis classificatie) werden samen met de ^{18}F -FDG PET radiomic feature *long run low gray level emphasis* en de CT radiomic feature *run percentage* gebruikt voor de ontwikkeling van de predictiemodellen. De modellen toonden een absolute verbetering in discriminatoire nauwkeurigheid tot wel 22% ten opzichte van de huidige klinische standaard die gebaseerd is op de maximale standardized uptake value (SUV_{max}). Deze veelbelovende resultaten kunnen worden beschouwd als een eerste stap naar adequate voorspelling van pathologisch complete responders na nCRT.

In **Hoofdstuk 5** werd de ingezette strategie uit Hoofdstuk 4 verder uitgebreid door het analyseren van radiomic features van zowel pre-nCRT als post-nCRT ^{18}F -FDG PET scans van 73 slokdarmcarcinoom patiënten. Door het toevoegen van post-nCRT scans vlak voor de operatie kan de tumor respons gedurende de behandeling nauwlettend gevolgd worden. Naast het bepalen van radiomic features op de twee afzonderlijke scans, werden ook veranderingen van radiomic features en histogram afstanden (voor het meten van veranderingen in ^{18}F -FDG uptake distributie) berekend. We ontwikkelden een predictiemodel bestaande uit het klinische T-stadium en het post-nCRT *joint maximum*. Het joint maximum is een maat voor ordelijkheid of systematiek van voxel intensiteiten afkomstig van de PET-afbeelding. Een hoge mate van post-nCRT ^{18}F -FDG PET ordelijkheid verhoogde de kans op een pathologisch complete respons. Dit model gaf een absolute verbetering tot wel 30% in de discriminatoire nauwkeurigheid ten opzichte van de referentie predictiemodellen. De referentie modellen waren allen gebaseerd op conventionele SUV_{max} metingen, klinische parameters en klinische parameters samen met pre-nCRT radiomic features. Er werd een hoge onderlinge correlatie gevonden tussen de post-nCRT radiomic features *joint maximum*, *median absolute deviation*, *joint entropy*, *sum entropy*, *angular second moment* en *inverse variance*. In het predictiemodel kon de joint maximum vervangen worden door elk van deze features onder behoud van de prestaties van het predictiemodel.

In **Hoofdstuk 6** werd een multimodaliteit predictie strategie geïntegreerd. Er werd onderzocht of de predictiemodellen uit Hoofdstuk 4 verbeterd konden worden door het toevoegen van potentiële biologische tumor markers. Er werd een immunohistochemische

beoordeling verkregen van de biologische tumor markers human epidermal growth factor receptor 2 (HER2), cluster of differentiation 44, hypoxia-inducible factor alpha, patched homolog 1 en sonic hedgehog. De bepalingen werden gedaan op tumor biopten voor de behandeling met nCRT (“pre-nCRT-biopten”) van 43 patiënten met een lokaal uitgebreid slokdarmcarcinoom. Van alle biologische markers, werd er enkel een significante associatie gevonden tussen pathologisch complete respons en HER2. Daarnaast zorgde inclusie van HER2 voor een substantiële verbetering van het predictiemodel uit Hoofdstuk 4. Geen van de andere biologische markers leidden tot een verbetering van het predictiemodel. Deze resultaten laten zien dat een multimodaliteit predictie strategie, gebaseerd op een combinatie van klinische, genetische en radiomic features, kan leiden tot een verbeterde voorspelling van pathologisch complete response op nCRT in patiënten met een slokdarmcarcinoom.

DEEL III: TOEPASSINGEN VAN ^{18}F -FDG PET RADIOMIC FEATURES IN INFECTIES

In **Deel III** werden de mogelijke toepassingen van ^{18}F -FDG PET radiomic features onderzocht bij aorta-iliacale prothese infecties na chirurgische reconstructies. Voor de diagnose van prothese infecties zijn vaak beeldvormende technieken nodig. Echter, de beoordeling van PET beelden blijft een uitdaging door het gebrek aan evidence-based interpretatie criteria. Alhoewel de meeste research naar radiomics zich richt op oncologische toepassingen toonden we in dit deel van het proefschrift aan dat radiomics ook een mogelijke rol kan vervullen in de diagnose van vasculaire graft infecties.

In **Hoofdstuk 7** werden er 30 patiënten met aorta-iliacale prothesen ingedeeld naar klinische verdenking op een prothese infectie. De diagnose van een infectie werd gesteld op basis van een positieve kweek van peri-vasculair debris rond de vaatprothese. Zestien patiënten werden klinisch verdacht voor een prothese infectie (groep I). Van deze groep hadden 10 patiënten een positieve (groep Ia) en 6 patiënten een negatieve kweek (groep Ib) van het vocht rondom de prothese. De overige 14 patiënten werden klinisch niet verdacht voor infectie en werden gebruikt als controle (groep II). Deze patiënten ondergingen een ^{18}F -FDG PET/CT scan voor andere klinische doeleinden zoals kanker stadiëring. Conventionele metingen gerelateerd aan de PET beelden zoals de SUV_{max} , tissue-to-background en visual grading scale waren in staat om groep I van groep II te onderscheiden. Hierbij hadden textural features geen aanvullende waarde op de conventionele metingen. Echter, de conventionele aan de PET gerelateerde metingen waren niet in staat om groep Ia van groep Ib te onderscheiden. Dit onderscheid kon wel worden aangetoond door de PET textural feature *short run high grey level emphasis*. PET radiomics is derhalve een potentiële techniek om de diagnostiek van prothese infecties te verbeteren zodat invasieve diagnostische procedures op termijn vervangen kunnen worden.

Chapter 10

Author Affiliations

Dankwoord/Acknowledgements

Over de Auteur

List of Publications

AUTHOR AFFILIATIONS

Author affiliations for Chapter 1:

Alex Zwanenburg^{1,2,3}, Mahmoud A. Abdalah⁴, Hugo J.W.L. Aerts⁵, Vincent Andrearczyk⁶, Aditya Apte⁷, Saeed Ashrafinia^{8,9}, Roelof J. Beukinga¹⁰, Ronald Boellaard^{10,11}, Marta Bogowicz¹², Luca Boldrini¹³, Adrien Depeursinge^{6,14}, Marie-Charlotte Desseroit¹⁵, Nicola Dinapoli¹⁶, Cuong Viet Dinh¹⁷, Sebastian Echegaray¹⁸, Issam El Naqa^{19,20}, Andriy Y. Fedorov²¹, Roberto Gatta¹⁶, Robert J. Gillies⁴, Matthias Guckenberger¹², Nils Gähler²², Michael Götz²², Mathieu Hatt¹⁵, Fabian Isensee²², Philippe Lambin²³, Stefan Leger^{1,3}, Ralph T.H. Leijenaar²³, Jacopo Lenkowicz¹³, Fiona Lippert²⁴, Are Losnegård²⁵, Klaus H. Maier-Hein²², Olivier Morin²⁶, Henning Müller^{6,27}, Sandy Napel^{18,28,29}, Arman Rahmim³⁰, Arvind U.K. Rao³¹, Nanna M. Sijtsma³², Jairo Socarras Fernandez²⁴, Emiliano Spezi^{33,34}, Roel J.H.M. Steenbakkers³², Stephanie Tanadini-Lang¹², Daniela Thorwarth²⁴, Esther G.C. Troost^{1,2,3,35,36}, Taman Upadhaya^{15,37}, Vincenzo Valentini¹³, Martin Vallières^{15,20}, Uulke A. van der Heide¹⁷, Lisanne V. van Dijk³², Joost van Griethuysen^{38,39,40}, Floris H.P. van Velden⁴¹, Philip Whybra³³, Christian Richter^{1,3,36*}, and Steffen Löck^{1,3,35*}

*These authors share senior authorship.

1. OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Helmholtz-Zentrum Dresden - Rossendorf, Dresden, Germany
2. National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany, and; Helmholtz Association / Helmholtz-Zentrum Dresden - Rossendorf (HZDR), Dresden, Germany
3. German Cancer Consortium (DKTK), Partner Site Dresden, and German Cancer Research Center (DKFZ), Heidelberg, Germany
4. Department of Cancer Imaging and Metabolism, Moffitt Cancer Center, Tampa (FL), USA
5. Dana-Farber Cancer Institute, Brigham and Women's Hospital, and Harvard Medical School, Harvard University, Boston (MA), USA
6. Institute of Information Systems, University of Applied Sciences Western Switzerland (HES-SO), Switzerland
7. Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York (NY), USA
8. Department of electrical and computer engineering, Whiting School of Engineering, Johns Hopkins University, Baltimore (MD), USA
9. Russell H. Morgan department of radiology and radiological science, School of medicine, Johns Hopkins University, Baltimore (MD), USA
10. Department of Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen (UMCG), Groningen, the Netherlands
11. Radiology and Nuclear Medicine, VU University Medical Centre (VUMC), Amsterdam, The Netherlands
12. Department of Radiation Oncology, University Hospital Zurich, University of Zurich, Switzerland
13. Polo Scienze Oncologiche ed Ematologiche, Istituto di Radiologia, Università Cattolica del Sacro Cuore, Fondazione Policlinico Universitario Agostino Gemelli, Roma, Italia
14. Biomedical Imaging Group, Ecole polytechnique federale de Lausanne (EPFL), Switzerland

15. Laboratory of Medical Information Processing (LaTIM) - team ACTION (image-guided therapeutic action in oncology), INSERM, UMR 1101, IBSAM, UBO, UBL, Brest, France
16. Polo Scienze Oncologiche ed Ematologiche, Università Cattolica del Sacro Cuore, Fondazione Policlinico Universitario Agostino Gemelli, Roma, Italia
17. Department of Radiation Oncology, the Netherlands Cancer Institute (NKI), Amsterdam, the Netherlands
18. Department of Radiology, Stanford University School of Medicine, Stanford (CA), USA
19. Department of Radiation Oncology, Physics Division, University of Michigan
20. Medical Physics Unit, McGill University, Montréal, Québec, Canada
21. Surgical Planning Laboratory, Brigham and Women's Hospital and Harvard Medical School, Harvard University, Boston (MA), USA
22. Department of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany
23. Department of Radiation Oncology (The D-Lab), GROW-School for Oncology and Developmental Biology, Maastricht University Medical Centre+, Maastricht, The Netherlands
24. Section for Biomedical Physics, Department of Radiation Oncology, University of Tübingen, Germany
25. Department of Clinical Medicine, University of Bergen, Bergen, Norway
26. Department of Radiation Oncology, University of California, San Francisco (CA), USA
27. University of Geneva, Switzerland
28. Department of Electrical Engineering, Stanford University, Stanford (CA), USA
29. Department of Medicine (Biomedical Informatics Research), Stanford University School of Medicine, Stanford (CA), USA
30. Russell H. Morgan Department of Radiology and Radiological Science, School of medicine, Johns Hopkins University, Baltimore (MD), USA
31. Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston (TX), USA
32. Department of Radiation Oncology, University of Groningen, University Medical Center Groningen (UMCG), Groningen, The Netherlands
33. School of Engineering, Cardiff University, Cardiff, United Kingdom
34. Department of Medical Physics, Velindre Cancer Centre, Cardiff, UK
35. Department of Radiotherapy and Radiation Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany
36. Helmholtz-Zentrum Dresden - Rossendorf, Institute of Radiooncology – OncoRay, Dresden, Germany
37. Department of Nuclear Medicine, CHU Milétrie, Poitiers, France
38. Department of Radiology, the Netherlands Cancer Institute (NKI), Amsterdam, the Netherlands
39. GROW-School for Oncology and Developmental Biology, Maastricht University Medical Center, Maastricht, The Netherlands
40. Department of Radiation Oncology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, Harvard Medical School, Boston, MA
41. Department of Radiology, Leiden University Medical Center (LUMC), Leiden, the Netherlands

All authors, except for the first and senior authors, are ordered alphabetically as the proportional relevance of each contribution could not be determined with sufficient precision.

Author affiliations for Chapters 2-7:

Roelof J. Beukinga^{1,2,3}, Ronald Boellaard^{2,4}, Johannes G.M. Burgerhof⁵, Robert P. Coppes^{6,7}, Rudi A.J.O. Dierckx², Lisanne V. van Dijk⁷, Willemieke P.M. Dijksterhuis³, Hette Faber⁶, Andor W.J.M. Glaudemans², Jan B. Hulshoff^{3,8}, Johan R. de Jong², Arend Karrenbeld⁹, Gursah Kats-Ugurlu⁹, Christina T. Muijs⁷, Véronique E.M. Mul⁷, Walter Noordzij², Elisabeth Pfaehler², John Th.M. Plukker³, Michel M.P.J. Reijnen¹⁰, Ben R. Saleem¹¹, Riemer H.J.A. Slart^{1,2}, Cornelis H. Slump¹², Da Wang^{3,6}, and Clark J. Zeebregts¹¹

1. Biomedical Photonic Imaging (BMPI), University of Twente, Enschede, The Netherlands.
2. Nuclear Medicine and Molecular Imaging, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
3. Surgery (Division of Surgical Oncology), University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
4. Department of Radiology & Nuclear Medicine, VU University Medical Center, Amsterdam, The Netherlands.
5. Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
6. Cell Biology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
7. Radiation Oncology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
8. Radiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
9. Pathology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
10. Surgery, Rijnstate Hospital, Arnhem, The Netherlands.
11. Surgery (Division of Vascular Surgery), University of Groningen, University Medical Center Groningen, Groningen, The Netherlands.
12. MIRA Institute for Biomedical Technology and Technical Medicine, University of Twente, Enschede, The Netherlands.

All authors are ordered alphabetically.

DANKWOORD

En dan is het helemaal klaar! Wat ben ik blij met het resultaat, maar dit proefschrift had er niet gelegen zonder de hulp van een aantal mensen. Hierbij wil ik iedereen bedanken die op welke manier dan ook heeft bijgedragen aan de totstandkoming van dit proefschrift. Ik noem graag een aantal personen in het bijzonder.

Promotor **prof. dr. R.H.J.A. Slart**, beste Riemer, vanaf het moment dat je plaatsnam in de commissie voor mijn afstudeerstage van de opleiding Technische Geneeskunde, hebben we gezamenlijk gezocht naar mogelijkheden om mijn afstudeerwerk in een promotietraject voort te kunnen zetten. Zo kwam ik uiteindelijk terecht op de afdeling Biomedical Photonic Imaging op de Universiteit Twente. Ik wil je bedanken voor die geschonken mogelijkheid en het in mij gestelde vertrouwen. We hebben altijd heel prettig en vooral efficiënt kunnen samenwerken. Naast al het onderzoek kwamen soms ook andere dingen in het leven aan bod, zoals onze passie voor zeilen. Zo kwamen we er op een bepaald moment achter dat we allebei vaak in de jachthaven van Makkum te vinden zijn, maar we elkaar daar nog nooit waren tegen gekomen. Ik hoop dat onze samenwerking in de toekomst nog tot vele mooie dingen gaat leiden.

Promotor **prof. dr. J.Th.M. Plukker**, beste John, in 2014 begon ik op de chirurgische oncologie met mijn afstudeerstage. Jouw voorstel om onderzoek te doen naar de rol van PET, MRI en kwantitatieve beeldvorming in de voorspelling van therapie bij patiënten met slokdarmkanker sprak mij erg aan. Het verbaasde me dat jij als chirurg een buitengewoon grote interesse had voor beeldvorming en allerlei technische details. Dit leidde met mijn achtergrond als Technisch Geneeskundige vaak tot unieke en frisse ideeën. Maar misschien fungeerde onze prettige omgang nog wel het meest als basis voor onze goede samenwerking. Dit leidde onder andere tot één van de hoogtepunten van mijn PhD, de ASCO GI 2016 in San Francisco. Naast de leerzame momenten en de nodige sightseeing, leverde dit mij een bananen-incident op met de Amerikaanse douane. Bedankt voor de afgelopen jaren, ik ben je veel dank verschuldigd voor alle tijd die je in mij hebt geïnvesteerd.

Co-promotor **prof. dr. ir. C.H. Slump**, beste Kees, jouw technische begeleiding zorgde voor een rode draad door mijn mastertraject van de opleiding Technische Geneeskunde en mijn promotie. Ik heb je leren kennen als iemand van de sterke oneliners en de gevatte grapjes. Tijdens mijn promotietraject waren er tijden dat we elkaar maar weinig spraken, maar juist op de belangrijke momenten kon ik op je rekenen. Je kon me dan vaak even uit mijn tunnelvisie trekken en ik ging dan alles weer in het juiste perspectief zien. Naast de inhoudelijke begeleiding was je vooral ook een goede mentor voor mij. Je hebt me vaak van advies voorzien bij moeilijke beslissingen ten aanzien van mijn carrière. Je zorgde daarbij

na mijn afstuderen zelfs voor een tijdelijke aanstelling in jouw vakgroep Robotics and Mechatronics aan de Universiteit Twente. Ik wil je graag bedanken voor alle begeleiding die ik van je heb mogen ontvangen.

Prof. dr. ir. R. Boellaard, beste Ronald, ondanks dat je officieel niet betrokken was bij mijn promotie, hebben we veel gesprekken gevoerd over gezamenlijk wetenschappelijk onderzoek. Ik heb grote waardering voor de wijze waarop jij onderzoek doet. De manier waarop jij resultaten interpreteert en jouw ervaring in het doen van wetenschappelijk onderzoek waren een grote inspiratiebron en zorgden ervoor dat we ons onderzoek telkens weer konden verbeteren. Ik hoop dat we in de toekomst kunnen blijven samenwerken.

Prof. dr. C.J.A.M. Zeebregts, beste Clark, we kennen elkaar sinds ik een klinische masterstage stage kwam lopen op de vaatchirurgie van het UMCG. Wat toevallig dat we elkaar een paar jaar later opnieuw tegen kwamen tijdens mijn promotietraject met een project naar het verbeteren van de diagnose van aorta-iliacale infecties. Bedankt voor de behulpzaamheid en vruchtbare samenwerking.

Jan Binne Hulshoff, met de medische opleiding die jij hebt genoten kon je mij altijd helpen als het mij net even té medisch werd. Omdat we alles met elkaar konden overleggen, was het voor mij wel even wennen toen jij het kantoor verliet omdat je je promotie had afgerond. Eén van de hoogtepunten was onze reis naar San Francisco voor de ASCO GI 2016. We hebben het daar ontzettend naar onze zin gehad en buiten de officiële gedeeltes om gaf je mij een persoonlijke rondleiding door San Francisco. Op naar nog veel meer mooi onderzoek en gezelligheid!

Sanne van Dijk, vooral tijdens mijn afstuderen en het begin van mijn promotietraject hebben we onder het genot van menig *bakkie* elkaar verder geholpen met het programmeren van de radiomic features. Daarbij was je nooit te beroerd om je mening te ventileren. Naast dat we elkaar hielpen op inhoudelijk gebied, voorzag je me ook van advies ten aanzien van mijn carrière. Bedankt voor je support, ik heb er veel aan gehad!

Co-auteurs **Hans Burgerhof, Rob Coppes, Rudi Dierckx, Willemieke Dijksterhuis, Hette Faber, Andor Glaudemans, Johan de Jong, Arend Karrenbeld, Gursah Kats-Ugurulu, Kristel Muijs, Véronique Mul, Walter Noordzij, Elli Pfaehler, Michel Reijnen, Ben Saleem en Daisy Wang**, bedankt voor jullie bijdrage aan ons gezamenlijke wetenschappelijk onderzoek. Ik ben jullie dankbaar voor de vele discussies die ik met jullie heb gevoerd. Zonder jullie had dit proefschrift niet tot stand kunnen gekomen.

Kamergenoten **Andries Groen, Iris Koopmans, Ilona Schepel en Rinse Ubbink**, ik ben blij dat ik gedurende mijn hele promotie een kamer met jullie heb mogen delen. Als

kamergenoten kregen we nagenoeg elk hoogte- en dieptepunt van elkaars promotie en privéleven mee. Gezellig was het in ieder geval wel, en daarom hingen we trouw elke avond voor één van onze verjaardagen allemaal slingers door de kamer, waarop de jarige telkens weer heel “verrast” was. Op den duur besloten we dat de slingers ook wel konden blijven hangen, waardoor het stoffige karakter van onze kamer toch nog iets werd opgefleurd. Bedankt voor jullie gezelligheid.

Mede-onderzoekers van het Triadegebouw, regelmatig waren de dagen net even iets té gezellig tijdens een koffiebreek, lunch of vrijdagmiddag borrel. Gelukkig zorgde dat er dan wel weer voor dat de dagen voorbij vlogen en het werk een stuk luchtiger werd. Ik kijk alweer uit naar onze volgende vrijdagmiddag borrel!

Paranimfen **Tom Snuverink** en **Robin Wesselink**, één van de verbindende factoren van onze vriendschap is onze passie voor muziek. Talloze uren hebben we doorgebracht met muziek draaien en het bezoeken van muziekfestivals. Ik heb veel gehad aan de vele goede gesprekken, die zelfs tot in de vroege uurtjes onder het genot van een drankje reikten. Bedankt voor de ondersteuning voor vandaag en ik wens jullie allebei heel veel succes met jullie eigen carrière.

Vrienden voor het leven, **Marja Altenburg, Larissa Dijkstra, Mathilde Hermans, Bart Hoebe, Tom Janssen, Geert Kleinnibbelink, Chris van Lieshout, Floortje Lok, Willem van Schijndel, Feiko Tiessens, Juliette Velu** en **Tom van Vugt**, bedankt voor alle fantastische momenten die we samen hebben beleefd! Naast alle gezelligheid, hebben jullie ook altijd interesse getoond in mijn werk en was er ruimte voor het inhoudelijke gesprek. Maar eerlijk is eerlijk, vaker stond de gezelligheid centraal. Het is mooi om te zien dat, ondanks de grote geografische afstand tussen sommigen van ons, we allemaal moeite blijven doen om elkaar zoveel mogelijk te blijven zien. Ik hoop dat we dit in de toekomst nog lang voort kunnen blijven zetten!

Lieve **(schoon)familie**, dank voor jullie eeuwige steun en aanwezigheid! Ondanks dat het soms lastig uit te leggen is wat mijn werk precies inhoudt, waren jullie altijd ontzettend geïnteresseerd in alles wat ik doe. Onze familie is niet zo groot, maar daardoor kennen we elkaar door en door. *Klein, maar fijn* zal ik maar zeggen. Mijn familie werd fors uitgebreid toen ik met open armen werd ontvangen door mijn schoonfamilie. Jullie zijn er altijd voor mij geweest en ik weet dat jullie dat in de toekomst ook altijd blijven doen.

Beste **Julian**, ik ben dankbaar dat je mijn broer(tje) bent. Ondanks dat we soms zo verschillend kunnen zijn, konden we het altijd zo goed met elkaar vinden. Als kleine kinderen waren we al geïnteresseerd in techniek; we speelden urenlang met (technisch) Lego. Weet je nog dat we al die gebouwen en vliegtuigen bouwden? Maar gedurende we ouder werden,

gingen de gesprekken steeds vaker over medische techniek en dan met name over medische beeldvorming. Bedankt voor je interesse die je door de jaren heen hebt getoond.

Lieve **pap en mam**, ontzettend bedankt voor de support die ik altijd van jullie heb mogen ontvangen. Met jullie advies helpen jullie mij tot op de dag van vandaag met alle leuke en minder leuke dingen. Pap, mijn interesse voor zowel de geneeskunde als de techniek komt niet van een vreemde. Bedankt voor de vele inspirerende gesprekken die we hebben gevoerd. Mam, ik heb altijd op jou kunnen terugvallen dankzij jouw ongelooflijke steun. Dankzij jullie heb ik het kunnen maken tot waar ik nu sta.

En tot slot natuurlijk **Alette**, of we nou discussies voeren over inhoudelijk medisch-technische zaken, carrière maken of over alle dingen die spelen in ons leven; wat is het fijn dat we elkaar zo goed begrijpen. Uiteindelijk ben ik er trots op dat we zo'n goed team zijn en elkaar alleen maar sterker maken. Met jou zie ik de toekomst vol vertrouwen tegemoet. Ik hou van je!

A handwritten signature in black ink. It features a large, stylized capital letter 'J' with a horizontal line extending from its base. To the right of the vertical stroke of the 'J', the word 'Jorn' is written in a smaller, cursive script.

OVER DE AUTEUR

Roelof Jorn Beukinga werd geboren op 2 februari 1990 te Ruinen als zoon van Roelof en Anja Beukinga. In 2008 behaalde hij zijn VWO diploma aan het RSG Wolfsbos, locatie de Groene Driehoek te Hogeveen. Zijn interesse voor zowel geneeskunde als technologie bracht hem bij de opleiding Technische Geneeskunde aan de Universiteit Twente. Tijdens de bachelorfase werd zijn interesse gewekt voor medische beeldvorming. Daarom begon hij na het behalen van zijn bachelorgraad in 2011 aan een masterprogramma waarin hij zich specialiseerde in Medical Imaging & Interventions. Na meerdere klinische stages te hebben afgerond, begon hij aan een afstudeerstage bij de chirurgische oncologie in het Universitair Medisch Centrum Groningen. Dit resulteerde in een Masterthesis getiteld: “Prognostic and Predictive Value of PET Textural Features of Neoadjuvant Chemoradiation of Esophageal Cancer”. In 2015 verkreeg hij zijn Mastergraad en werd hem de kans geboden om zijn werk voort te zetten in een PhD traject onder begeleiding van prof. dr. R.H.J.A. Slart, prof. dr. J.Th.M. Plukker en prof. dr. ir. C.H. Slump. Gedurende dit PhD traject gaf hij meerdere presentaties op (internationale) congressen over zijn verrichtte werk op het gebied van “radiomics”, een kwantificatie techniek voor medische beeldvorming. In 2018 bood hij zijn proefschrift getiteld “Quantitative ^{18}F -FDG PET/CT in Esophageal Cancer and Vascular Graft Infections” aan ter verkrijging van de doctorsgraad. Momenteel is hij werkzaam op de afdeling radiologie in het Universitair Medisch Centrum Groningen.

LIST OF PUBLICATIONS

Ordered based on publication date, starting with the most recent publication.

1. [Beukinga RJ](#), Hulshoff JB, Mul VEM, Prediction of Response to Neoadjuvant Chemotherapy and Radiation Therapy with Baseline and Restaging ^{18}F -FDG PET Imaging Biomarkers in Patients with Esophageal Cancer. *Radiology*. 2018;287:983-992.
2. [Beukinga RJ](#), Hulshoff JB, van Dijk LV, et al. Predicting Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer with Textural Features Derived from Pretreatment ^{18}F -FDG PET/CT Imaging. *J Nucl Med*. 2017;58:723-9.
3. Saleem BR*, [Beukinga RJ](#)*, Boellaard R, et al. Textural features of ^{18}F -fluorodeoxyglucose positron emission tomography scanning in diagnosing aortic prosthetic graft infection. *Eur J Nucl Med Mol Imaging*. 2017;44:886-894.
4. van Dijk LV, Brouwer CL, van der Schaaf A, ... [Beukinga RJ](#), et al. CT image biomarkers to improve patient-specific prediction of radiation induced xerostomia and sticky saliva. *Radiother Oncol*. 2017;122:185-191.
5. Jager PL, [Beukinga RJ](#), de Haan TD, et al. Adequate diagnostic performance of combined ^{18}F -fluormethylcholine PET-CT with diffusion-weighted MRI in primary staging of high risk prostate cancer. *Adv J Mol Imaging*. 2016;06:1-9.

*These authors contributed equally.

Appendix I

Supplementary Data

SUPPLEMENTAL DATA

Morphological Features

Notation:

$N_{v,m}$ Number of voxels in the region of interest morphological mask.

$N_{v,gl}$ Number of voxels in the region of interest intensity mask.

X_c The morphological mask voxel set.

$X_{c,gl}$ The intensity mask voxel set.

X_{gl} Intensities in the intensity mask voxel set.

μ Mean of X_{gl} .

SUPPLEMENTAL TABLE 1

Morphological Features

Variable	Equation	Description
Volume	$V = \left \sum_{k=1}^{N_{fc}} \frac{a \cdot (b \times c)}{6} \right $	The volume of the region of interest mesh is calculated by the summation over N_{fc} tetrahedrons which are defined by face k and the origin. Here a , b , and c are the vertex points of face k .
Approximate volume	$\sum_{k=1}^{N_{v,m}} V_k$	Here V_k is the volume of voxel k .
Surface area	$A = \sum_{k=1}^{N_{fc}} \frac{ ab \times ac }{2}$	The surface area is also calculated from the region of interest mesh by summing over the triangular face surface areas.
Surface to volume ratio	A/V	Surface to volume ratio.
Compactness 1	$\frac{V}{\sqrt{\pi A^2}}$	The degree to which a region of interest is compact or sphere-like. The most compact shape is a perfect sphere.
Compactness 2	$\frac{36\pi \cdot V^2}{A^3}$	The degree to which a region of interest is compact or sphere-like. The most compact shape is a perfect sphere.
Spherical disproportion	$\frac{A}{(36\pi V^2)^{1/3}}$	The degree to which a region of interest is compact or sphere-like. The most compact shape is a perfect sphere.
Sphericity	$\frac{(36\pi V^2)^{1/3}}{A}$	The degree to which a region of interest is compact or sphere-like. The most compact shape is a perfect sphere.
Asphericity	$\left(\frac{1}{36\pi} \frac{A^3}{V^2} \right)^{1/3} - 1$	Describes how much the region of interest deviates from a perfect sphere, with perfectly spherical volumes having an asphericity of 0.
Centre of mass shift	$\left\ \frac{1}{N_{v,m}} \sum_{k=1}^{N_{v,m}} \vec{X}_{c,k} - \frac{\sum_{k=1}^{N_{v,gl}} X_{gl,k} \vec{X}_{c,gl,k}}{\sum_{k=1}^{N_{v,gl}} X_{gl,k}} \right\ $	The distance between the region of interest volume centroid and the intensity-weighted region of interest volume.
Maximum 3D diameter	$\max \left(\left\ \vec{X}_{vx,k_1} - \vec{X}_{vx,k_2} \right\ _2 \right)$	The distance between the two most distant vertices in the region of interest mesh vertex set X_{vx} .
Major axis length	$4 \sqrt{\lambda_{major}}$	A measure of how far the region of interest extends along the largest axis of the triaxial ellipsoid with the largest eigenvalue λ_{major} .

SUPPLEMENTAL TABLE 1 - cont.

Variable	Equation	Description
Minor axis length	$4\sqrt{\lambda_{minor}}$	A measure of how far the region of interest extends along the second largest axis of the triaxial ellipsoid with the second largest eigenvalue λ_{minor} .
Least axis length	$4\sqrt{\lambda_{least}}$	A measure of how far the region of interest extends along the least extended axis of the triaxial ellipsoid with the smallest eigenvalue λ_{least} .
Elongation	$\sqrt{\frac{\lambda_{minor}}{\lambda_{major}}}$	The ratio of the major and minor principal axis lengths could be viewed as the extent to which a volume is longer than it is wide, i.e. is eccentric.
Flatness	$\sqrt{\frac{\lambda_{least}}{\lambda_{major}}}$	The ratio of the major and least axis lengths could be viewed as the extent to which a volume is flat relative to its length.
Volume density - axis-aligned bounding box	$\frac{V}{V_{aabb}}$	The ratio of the region of interest volume and the axis-aligned bounding box volume (V_{aabb}) of the region of interest mesh.
Area density - axis-aligned bounding box	$\frac{A}{A_{aabb}}$	The ratio of the region of interest surface area and the surface area of the axis-aligned bounding box (A_{aabb}) enclosing the region of interest mesh.
Integrated intensity	$V \frac{1}{N_{v,gl}} \sum_{k=1}^{N_{v,gl}} X_{v,gl}$	Integrated intensity or total legion glycolysis is the average grey level multiplied by the volume.
Moran's I index	$\frac{N_{v,gl}}{\sum_{k_1=1}^{N_{v,gl}} \sum_{k_2=1}^{N_{v,gl}} w_{k_1,k_2}} \frac{\sum_{k_1=1}^{N_{v,gl}} \sum_{k_2=1}^{N_{v,gl}} w_{k_1,k_2} (X_{gl,k_1} - \mu)(X_{gl,k_2} - \mu)}{\sum_{k=1}^{N_{v,gl}} (X_{gl,k} - \mu)^2}$ $, k_1 \neq k_2$	An indicator of spatial autocorrelation. Here, w_{k_1,k_2} is a weight factor. Values of Moran's I close to 1.0, 0.0 and -1.0 indicate high spatial autocorrelation, no spatial autocorrelation, and high spatial anti-autocorrelation, respectively.
Geary's C measure	$\frac{N_{v,gl} - 1}{2 \sum_{k_1=1}^{N_{v,gl}} \sum_{k_2=1}^{N_{v,gl}} w_{k_1,k_2}} \frac{\sum_{k_1=1}^{N_{v,gl}} \sum_{k_2=1}^{N_{v,gl}} w_{k_1,k_2} (X_{gl,k_1} - X_{gl,k_2})^2}{\sum_{k=1}^{N_{v,gl}} (X_{gl,k} - \mu)^2}$ $, k_1 \neq k_2$	An indicator of spatial autocorrelation. In contrast to Moran's I index, Geary's C measure directly assesses grey level differences between voxels and is more sensitive to local spatial autocorrelation.

Local Intensity Features

SUPPLEMENTAL TABLE 2
Local intensity features

Variable	Equation	Description
Local intensity peak	N/A	The mean intensity in a 1 cm ³ spherical volume for the voxel(s) with the maximum grey level. It was devised for reducing variance in determining the maximum grey level.
Global intensity peak	N/A	Calculates the mean intensity within a 1 cm ³ spherical neighborhood for every voxel in the region of interest intensity mask and selects the highest intensity peak value.

Intensity-Based Statistical Features

Notation:

- N_v Number of voxels in the region of interest morphological mask.
 N_{gl} The intensity values of the three dimensional image matrix with N voxels.
 X_{gl} The set of grey levels of the N_v voxels included in the region of interest intensity mask.
 μ Mean grey level of X_{gl} .
 σ Grey level standard deviation of X_{gl} .

SUPPLEMENTAL TABLE 3
Intensity-based Statistical Features

Variable	Equation	Description
Mean	$\mu = \frac{\sum_{k=1}^{N_v} X_{gl,k}}{N_v}$	Mean grey level of X_{gl} .
Variance	$\frac{1}{N_v} \sum_{k=1}^{N_v} (X_{gl,k} - \mu)^2$	Grey level variance of X_{gl} .
Skewness	$\frac{\frac{1}{N_v} \sum_{k=1}^{N_v} (X_{gl,k} - \mu)^3}{\left(\frac{1}{N_v} \sum_{k=1}^{N_v} (X_{gl,k} - \mu)^2 \right)^{3/2}}$	A measure of the asymmetry of the data around the sample mean.
Kurtosis	$\frac{\frac{1}{N_v} \sum_{k=1}^{N_v} (X_{gl,k} - \mu)^4}{\left(\frac{1}{N_v} \sum_{k=1}^{N_v} (X_{gl,k} - \mu)^2 \right)^2} - 3$	A measure of peakedness in the grey level distribution X_{gl} .
Median intensity	$\text{median}(X_{gl})$	Median grey level of X_{gl} .
Minimum	$\min(X_{gl})$	Lowest grey level present in X_{gl} .
10 th percentile	N/A	10th percentile of X_{gl} .
90 th percentile	N/A	90th percentile of X_{gl} .
Maximum	$\max(X_{gl})$	Highest grey level present in X_{gl} .
Interquartile range	75 th percentile – 25 th percentile	Interquartile range of grey levels of X_{gl} .
Range	$\max(X_{gl}) - \min(X_{gl})$	Range of grey levels of X_{gl} .
Mean absolute deviation	$\frac{1}{N_v} \sum_{k=1}^{N_v} X_{gl,k} - \mu $	A measure of dispersion from the mean of X_{gl} .
Robust mean absolute deviation	$\frac{1}{N_{v,10-90}} \sum_{k=1}^{N_{v,10-90}} X_{gl,N_{v,10-90}} - \bar{X}_{gl,N_{v,10-90}} $	A more robust measure of dispersion from the mean of X_{gl} . $X_{gl,N_{v,10-90}}$ is the set of voxels in X_{gl} whose grey levels lie between the values corresponding to the 10 th and 90 th percentiles of X_{gl} .
Median absolute deviation	$\frac{1}{N_v} \sum_{k=1}^{N_v} X_{gl,k} - \text{median} $	A measure of dispersion from the median of X_{gl} .
Coefficient of variation	$\frac{\sigma}{\mu}$	A measure of dispersion of the X_{gl} distribution.
Quartile coefficient of dispersion	$\frac{75^{\text{th}} \text{ percentile} - 25^{\text{th}} \text{ percentile}}{75^{\text{th}} \text{ percentile} + 25^{\text{th}} \text{ percentile}}$	A robust alternative to coefficient of variance.
Energy	$\sum_{k=1}^{N_v} X_{gl,k}^2$	The summation of all squared grey levels.

Textural Features - Grey Level Co-Occurrence Based Features

Notation:

p_{ij} $(i,j)^{th}$ entry in a normalized grey level co-occurrence matrix.

N_g Number of distinct grey levels in the quantized image.

$$p_i \text{ and } p_j \quad \sum_{j=1}^{N_g} p_{i,j} \text{ and } \sum_{i=1}^{N_g} p_{i,j}$$

$$p_{|i+j|,k} \quad \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{ij} [k = |i + j|]$$

$$p_{|i-j|,k} \quad \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{ij} [k = |i - j|]$$

μ_i and μ_j The mean of p_i and p_j .

σ_i and σ_j The standard deviation of p_i and p_j .

$$HX \quad - \sum_{i=1}^{N_g} p_i \log_2 p_i$$

$$HY \quad - \sum_{j=1}^{N_g} p_j \log_2 p_j$$

$$HXY \quad - \sum_{i=1}^{N_g} p_{ij} \log_2 p_{ij}$$

$$HXY1 \quad - \sum_{i=1}^{N_g} p_{ij} \log_2 (p_i p_j)$$

$$HXY2 \quad - \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{ij} \log_2 (p_i p_j)$$

SUPPLEMENTAL TABLE 4

Grey Level Co-occurrence Based Features

Variable	Equation	Description
Joint maximum	$\max(p_{i,j})$	Determines the grey level with the maximum probability in the grey level co-occurrence matrix. The maximum probability is expected to be high if the occurrence of the most predominant voxel pair is high.
Joint average	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} i p_{ij}$	The grey level weighted sum of joint probabilities.
Joint variance	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - \mu)^2 p_{ij}$	A measure of heterogeneity, which characterizes the distribution of grey levels around the mean. This feature puts relatively high weights on the elements that differ from the average value of the grey level co-occurrence matrix.
Joint entropy	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{ij} \log_2 p_{ij}$	A measure of disorder. Entropy is small when the image is texturally uniform.
Difference average	$\sum_{k=0}^{N_g-1} k p_{i-j,k}$	Variance of the difference histogram.
Difference variance	$\sum_{k=0}^{N_g-1} (k - \mu)^2 p_{i-j,k}$	Variance of the difference histogram.

SUPPLEMENTAL TABLE 4 - cont.

Variable	Equation	Description
Difference entropy	$-\sum_{k=0}^{N_g-1} p_{l-j}(k) \log_2 p_{l-j,k}$	Entropy of the difference histogram.
Sum average	$\sum_{k=2}^{2N_g} k p_{l+j,k}$	A measure of the relation between clear and dense areas in an image.
Sum variance	$\sum_{k=2}^{2N_g} (k - \mu)^2 p_{l+j,k}$	Variance of the sum histogram.
Sum entropy	$-\sum_{k=2}^{2N_g} p_{l+j,k} \log_2 p_{l+j,k}$	Entropy of the sum histogram.
Angular second moment/ Energy	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_{ij}^2$	Emphasizes local homogeneity. Homogeneous images have few dominant grey tone transitions, which results into a higher energy.
Contrast/Inertia	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - j)^2 p_{ij}$	A measure of local variations present in the image. A high contrast value indicates a high degree of local variation.
Dissimilarity	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} i - j ^2 p_{ij}$	A measure that defines the variation of grey level pairs.
Inverse difference/Homogeneity 1	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p_{ij}}{1 + i - j }$	A measure of local homogeneity. Grey level co-occurrences with a large difference in levels are weighed less, thus lowering the total feature value. The feature score is maximal if all grey levels are the same.
Inverse difference normalized	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p_{ij}}{1 + \frac{ i - j }{N_g}}$	The normalised version of inverse difference to improve classification ability.
Inverse difference moment	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p_{ij}}{1 + (i - j)^2}$	Inverse difference moment is similar in concept to the inverse difference feature, but with lower weights for elements that are further from the diagonal.
Inverse difference moment normalized	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p_{ij}}{1 + \frac{(i - j)^2}{N_g^2}}$	The normalised version of inverse difference moment to improve classification ability.
Inverse variance	$2 \sum_{i=1}^{N_g} \sum_{j>i}^{N_g} \frac{p_{ij}}{(i - j)^2}$	
Correlation	$\frac{1}{\sigma_i \sigma_j} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - \mu_i)(j - \mu_j) p_{ij}$	A measure of grey tone linear dependency of neighbouring cells. For an image with large areas of similar intensities, correlation is higher than for an image with noisier, uncorrelated intensities.
Autocorrelation	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} ij p_{ij}$	A measure of the magnitude of the fineness and coarseness of texture.
Cluster tendency	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i + j - \mu_i - \mu_j)^2 p_{ij}$	Indicates the number of potential clusters present in the image.
Cluster shade	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i + j - \mu_i - \mu_j)^3 p_{ij}$	A measure of the skewness and uniformity of the grey level co-occurrence matrix. A higher cluster shade implies greater asymmetry.
Cluster prominence	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i + j - \mu_i - \mu_j)^4 p_{ij}$	A measure of the skewness and asymmetry of the grey level co-occurrence matrix. A higher values implies more asymmetry about the mean value while a lower value indicates a peak around the mean value and less variation about the mean.

SUPPLEMENTAL TABLE 4 - cont.

Variable	Equation	Description
Information measure of correlation 1	$\frac{HXY - HXY1}{HX}$	
Information measure of correlation 2	$\sqrt{1 - e^{-2(HXY2 - HXY)}}$	

Textural Features - Grey Level Run Length Based Features

Notation:

N_g The number of discretised grey levels present in the region of interest intensity mask.

N_r The maximal possible run length along direction x.

r_{ij} Matrix element r_{ij} of the GLRLM is the occurrence of grey level i with run length j .

N_r The total number of voxels in the region of interest intensity mask.

$$N_s = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} r_{ij}$$

r_i The marginal sum of the runs over run lengths j for grey value i , i.e. $\sum_{j=1}^{N_r} r_{ij}$.

r_j The marginal sum of the runs over the grey values i for run length j , i.e. $\sum_{i=1}^{N_g} r_{ij}$.

$$p_{ij} = \frac{r_{ij}}{N_s}$$

$$\mu = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} j p_{ij}$$

SUPPLEMENTAL TABLE 5
Grey Level Run Length Based Features

Variable	Equation	Description
Short run emphasis	$\frac{1}{N_s} \sum_{j=1}^{N_r} r_j$	Is highly dependent on the occurrence of short runs and is expected large for fine textures.
Long run emphasis	$\frac{1}{N_s} \sum_{j=1}^{N_r} j^2 r_j$	Is highly dependent on the occurrence of long runs and is expected large for coarse textures.
Low grey level run emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} r_i$	Is highly dependent on the occurrence of runs with low grey levels.
High grey level run emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} i^2 r_i$	Is highly dependent on the occurrence of runs with high grey levels.
Short run low grey level emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{r_{ij}}{i^2 j^2}$	Is highly dependent on the occurrence of short runs with low grey levels.
Short run high grey level emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{i^2 r_{ij}}{j^2}$	Is highly dependent on the occurrence of short runs with high grey levels.
Long run low grey level emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{j^2 r_{ij}}{i^2}$	Is highly dependent on the occurrence of long runs with low grey levels.
Long run high grey level emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} i^2 j^2 r_{ij}$	Is highly dependent on the occurrence of long runs with high grey levels.
Grey level non-uniformity	$\frac{1}{N_s} \sum_{i=1}^{N_g} r_i^2$	Assessment of the distribution of runs over the grey values. The feature value is low when runs are equally distributed along grey levels.
Grey level non-uniformity normalised	$\frac{1}{N_s^2} \sum_{i=1}^{N_g} r_i^2$	This is a normalised version of grey level non-uniformity.

SUPPLEMENTAL TABLE 5 - cont.

Variable	Equation	Description
Run length non-uniformity	$\frac{1}{N_s} \sum_{j=1}^{N_r} r_{\cdot j}^2$	Assessment of the distribution of runs over the run lengths. The feature value is low when runs are equally distributed along run lengths.
Run length non-uniformity normalised	$\frac{1}{N_s^2} \sum_{j=1}^{N_r} r_{\cdot j}^2$	This is a normalised version of run length non-uniformity.
Run percentage	$\frac{N_s}{N_v}$	The fraction of the number of realised runs and the maximum number of potential runs. Strongly linear or highly uniform region of interest volumes produce a low run percentage.
Grey level variance	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} (i - \mu)^2 p_{ij}$	The variance in runs for the grey levels.
Run length variance	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} (j - \mu)^2 p_{ij}$	The variance in runs for run lengths.
Run entropy	$-\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p_{ij} \log_2 p_{ij}$	The entropy in runs.

Textural Features - Grey Level Size Zone Based Features

Notation:

N_g	The number of discretised grey levels present in the region of interest intensity mask.
N_r	The maximal possible zone size of any group of linked voxels.
s_{ij}	Matrix element s_{ij} of the GLSZM is the number of zones with discretized grey level i and size j .
N_v	The total number of voxels in the region of interest intensity mask.
N_s	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_z} s_{ij}$
s_i	The marginal sum of the runs over run lengths j for grey value i , i.e. $\sum_{j=1}^{N_r} s_{ij}$.
r_j	The marginal sum of the runs over the grey values i for run length j , i.e. $\sum_{i=1}^{N_g} s_{ij}$.
p_{ij}	$\frac{s_{ij}}{N_s}$
μ	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_z} j p_{ij}$

SUPPLEMENTAL TABLE 6
Grey Level Size Zone Based Features

Variable	Equation	Description
Small zone emphasis	$\frac{1}{N_s} \sum_{j=1}^{N_z} s_j$	Is highly dependent on the occurrence of small zones and is expected large for fine textures.
Large zone emphasis	$\frac{1}{N_s} \sum_{j=1}^{N_z} j^2 s_j$	Is highly dependent on the occurrence of large zones and is expected large for coarse textures.
Low grey level zone emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} s_i$	Is highly dependent on the occurrence of zones with low grey levels.
High grey level zone emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} i^2 s_i$	Is highly dependent on the occurrence of zones with high grey levels.
Small zone low grey level emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_z} \frac{s_{ij}}{i^2 j^2}$	Is highly dependent on the occurrence of small zones with low grey levels.
Small zone high grey level emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_z} \frac{i^2 s_{ij}}{j^2}$	Is highly dependent on the occurrence of small zones with high grey levels.
Large zone low grey level emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_z} \frac{j^2 s_{ij}}{i^2}$	Is highly dependent on the occurrence of large zones with low grey levels.
Large zone high grey level emphasis	$\frac{1}{N_s} \sum_{i=1}^{N_g} \sum_{j=1}^{N_z} i^2 j^2 s_{ij}$	Is highly dependent on the occurrence of large zones with high grey levels.
Grey level non-uniformity	$\frac{1}{N_s} \sum_{i=1}^{N_g} s_i^2$	Assessment of the distribution of zones over the grey values. The feature value is low when zones are equally distributed along grey levels.
Grey level non-uniformity normalised	$\frac{1}{N_s^2} \sum_{i=1}^{N_g} s_i^2$	This is a normalised version of grey level non-uniformity.

SUPPLEMENTAL TABLE 6 - cont.

Variable	Equation	Description
Zone size non-uniformity	$\frac{1}{N_s} \sum_{j=1}^{N_z} r_{.j}^2$	Assessment of the distribution of zones over the zones sizes. The feature value is low when zones are equally distributed along zones sizes.
Zone size non-uniformity normalised	$\frac{1}{N_s^2} \sum_{j=1}^{N_z} r_{.j}^2$	This is a normalised version of zone size non-uniformity.
Zone percentage	$\frac{N_s}{N_v}$	The fraction of the number of realised zones and the maximum number of potential zones. Strongly linear or highly uniform region of interest volumes produce a low zone percentage.
Grey level variance	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_z} (i - \mu)^2 p_{ij}$	The variance in zones for the grey levels.
Zone size variance	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_z} (j - \mu)^2 p_{ij}$	The variance in zones for zone sizes.
Zone size entropy	$-\sum_{i=1}^{N_g} \sum_{j=1}^{N_z} p_{ij} \log_2 p_{ij}$	The entropy in zones.

Textural Features - Neighbourhood Grey Tone Difference Based Features

Notation:

N_g	The number of discretised grey levels present in the region of interest intensity mask.
n_i	The total number of voxels with grey level i which have a valid neighbourhood.
$N_{v,c}$	$\sum n_i$, the total number of voxels that have at least one neighbour.
p_i	$\frac{n_i}{N_{v,c}}$
s_i	$\sum_k i - \bar{X}_k $, the summation of the neighbourhood grey tone difference.
$N_{g,p}$	The number of discretised grey levels with $p_i > 0$.
p_{i_1} and p_{i_2}	Grey level probabilities p_{i_1} and p_{i_2} are copies of p_i with different iterators.

SUPPLEMENTAL TABLE 7

Neighbourhood Grey Tone Difference Based Features

Variable	Equation	Description
Coarseness	$\frac{1}{\sum_{i=1}^{N_g} p_i s_i}$	The reciprocal of normalized sum of the deviations of pixel intensities from their neighborhood average intensities. Grey level differences in coarse textures are generally small due to large-scale patterns.
Contrast	$\left(\frac{1}{N_{g,p}(N_{g,p}-1)} \sum_{i_1=1}^{N_g} \sum_{i_2=1}^{N_g} p_{i_1} p_{i_2} (i_1 - i_2)^2 \right) \left(\frac{1}{N_{v,c}} \sum_{i=1}^{N_g} s_i \right)$	Contrast depends on the dynamic range of the grey levels as well as the spatial frequency of intensity changes.
Busyness	$\frac{\sum_{i=1}^{N_g} p_i s_i}{\sum_{i_1=1}^{N_g} \sum_{i_2=1}^{N_g} i_1 p_{i_1} - i_2 p_{i_2} }, p_{i_2} \neq 0 \text{ and } p_{i_2} \neq 0$	A measure of the spatial frequency of intensity changes. Textures with large changes in grey levels between neighbouring voxels are said to be busy.
Complexity	$\frac{\sum_{i=1}^{N_g} p_i s_i}{\sum_{i_1=1}^{N_g} \sum_{i_2=1}^{N_g} i_1 p_{i_1} - i_2 p_{i_2} }, p_{i_2} \neq 0 \text{ and } p_{i_2} \neq 0$	Refers to the visual information content of a texture. A texture is considered complex if the information content is high and rapid changes in grey levels are common.
Strength	$\frac{\sum_{i_1=1}^{N_g} \sum_{i_2=1}^{N_g} (p_{i_1} + p_{i_2})(i_1 - i_2)^2}{\sum_{i_1=1}^{N_g} s_i}, p_{i_2} \neq 0 \text{ and } p_{i_2} \neq 0$	A measure of the size of the primitives and the differences between their average intensities. This feature is increased when the primitives are easily definable and clearly visible.