Quantification of functional imaging biomarkers in medicine

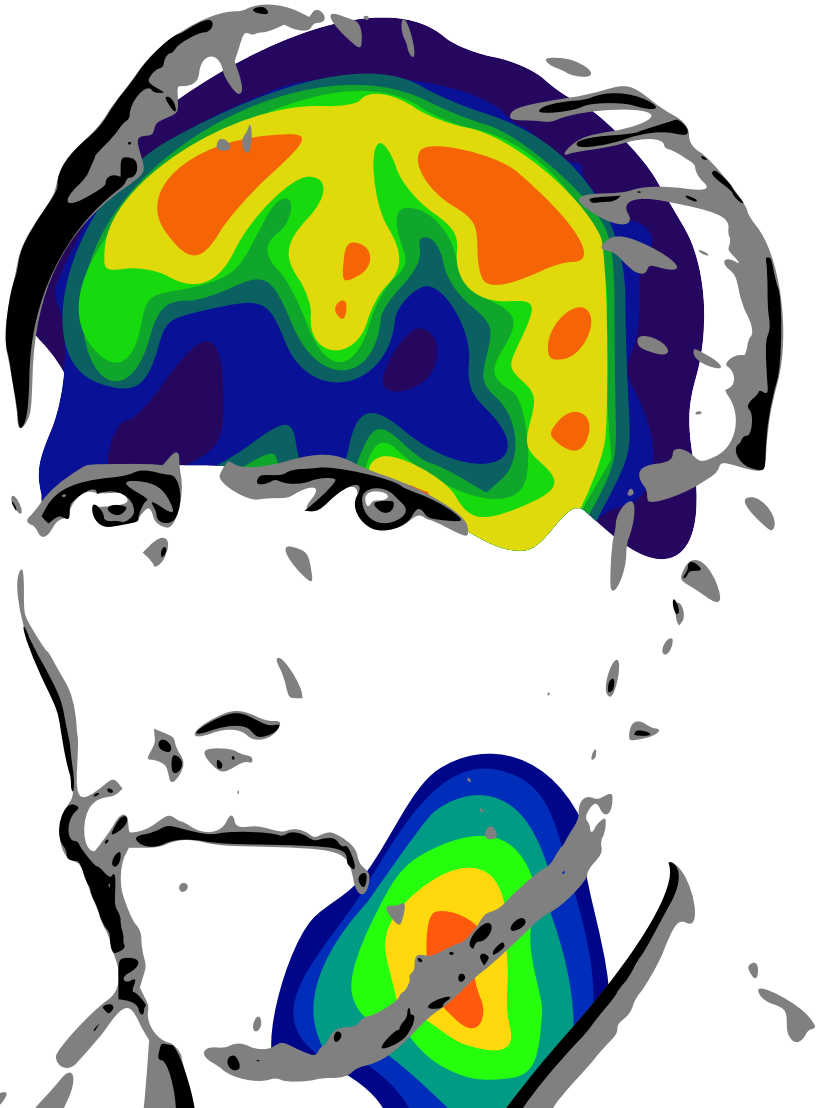
Koopman, Thomas

2022

Quantification of functional imaging biomarkers in medicine

Technical validation and simplification

Thomas Koopman



QUANTIFICATION OF FUNCTIONAL IMAGING BIOMARKERS IN MEDICINE

TECHNICAL VALIDATION AND SIMPLIFICATION

Cover: ThK
Lay-out: ThK

VRIJE UNIVERSITEIT

QUANTIFICATION OF FUNCTIONAL IMAGING BIOMARKERS IN MEDICINE

TECHNICAL VALIDATION AND SIMPLIFICATION

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geneeskunde
op vrijdag 8 april 2022 om 13.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Thomas Koopman

geboren te Huizen

promotoren:	prof.dr. R. Boellaard prof.dr. J.A. Castelijns
copromotoren:	dr.ir. M.M. Yaqub dr. J.T. Marcus
promotiecommissie:	prof.dr. E.F.I. Comans prof.dr.ir. H.W.A.M. de Jong dr. E.M. van de Giessen prof.dr.ir. A.J. Nederveen prof.dr. R. de Bree dr. P. de Graaf

Aan mijn ouders

Table of Contents

1

Introduction

Biomarkers are objective indicators of biological processes that can be measured accurately and precisely. They play a fundamental role in medical practice and research.[1] Evaluation of biomarkers in clinical research is aimed at providing evidence for the biomarker properties. Once validated—or rather, while constantly being evaluated[2]—biomarkers can be used for diagnosis of disease and for prediction of (or as a surrogate for) clinical endpoints, for example in risk assessment or treatment response prediction.[1-3] If these characteristics are extracted from medical images (e.g., CT, MRI, PET, ultrasound, etc), we speak of imaging biomarkers. These biomarkers can be qualitative, i.e., semantic descriptions of an image; or they can be quantitative, i.e., derived by measurement.

"A quantitative imaging biomarker is an objective characteristic derived from an in vivo image measured on a ratio or interval scale as an indicator of normal biological processes, pathogenic processes or a response to a therapeutic intervention."[4] Tumour volume is an example of a quantitative imaging biomarker, generally measured by computed tomography (CT) or magnetic resonance imaging (MRI), as these techniques can offer detailed, high resolution anatomical data.

Complementary to such anatomical information, functional imaging biomarkers aim to reflect underlying biological processes. Evidence shows that imaging biomarkers reflecting for example cell metabolism, cell density, and blood perfusion rate are predictive for treatment response and useful in risk assessment. The challenge is to reliably measure these characteristics.

This reliability is assessed by the abovementioned accuracy and precision. The former is the closeness to its true value, which is generally not available in clinical research. The latter construct is the closeness of replicate measurements and can be divided into two aspects: reproducibility and repeatability. Reproducibility represents precision under different measurement conditions (different imaging systems with different operators), whereas repeatability is precision under the same conditions (the same imaging system with the same operators).[4]

This work focusses on several imaging biomarkers. In the following sections I'll introduce these biomarkers in light of two relevant clinical challenges, while also providing a brief introduction to the underlying theory.

Cancer

Cancer is a disease with increasing incidence in our aging societies. According to global estimates, currently "1 in 5 men and 1 in 6 women will develop the disease and 1 in 8 men and 1 in 10 women will die from it."[5] However, thanks to improvements in medical care, the cancer mortality rate continues to decline. There are more than 100 different types of cancer, two of which are subject to research in this thesis.

Diffuse glioma

Diffuse gliomas are the most common brain tumours and more common in men than in women, with a respective ratio of 3 to 2 in the Netherlands. Named after their resemblance to glial cells, gliomas are aggressive, progressive and often diffuse, meaning the tumour infiltrates the surrounding tissue. Tumours are classified into four grades. The first grade describes non-diffuse tumours. Diffuse gliomas are classified as grade II, III or IV. Lower grade gliomas progress over time into higher, more aggressive grades, grade IV (glioblastoma) being the most aggressive. Patients with a primary diagnosis of glioblastoma are often older than patients with lower grade gliomas, the latter group often being younger than 50 years. Median survival ranges from 8 years for grade II to 15 months for grade IV.

Infiltration of the tumour into the peritumoral tissue makes it impossible to remove completely. On the one hand because the extent of infiltration is impossible to detect completely; on the other hand, because vital brain areas may be infiltrated, which cannot be removed. However, increased macroscopic resection is associated with better survival. If possible, this is the first step in treatment. Surgery also procures tumour tissue (otherwise obtained with a biopsy) which is used for further diagnosis through histology and molecular classification. Radiotherapy and chemotherapy are given to nearly all patients except those with certain low-grade gliomas that are slow growing and associated with longer survival.

MR imaging is currently indicated in the diagnosis and also during and after treatment. Standard MRI sequences include T1, T2, post contrast T1, FLAIR and diffusion weighted imaging. Contrast enhancement visible in the post contrast T1 image is important for classification, because contrast enhancing gliomas are often glioblastoma while non-enhancing tumours are often lower grade gliomas. Imaging is used primarily for estimation of the extent of tumour infiltration. This is used to determine the extent of resection in case of surgery and to define target volumes for radiotherapy.

Alternative imaging methods that measure functional imaging biomarkers are under investigation. These methods include dynamic contrast enhanced (DCE) MRI, which can measure perfusion related biomarkers, and amino acid PET, which aims to measure cell proliferation rates. The amino acid PET tracer, O-(2-[^{18}F]fluoroethyl)-L-tyrosine ([^{18}F]FET), has shown promising results for detection the tumour extent and its use has been recommended as a new standard for glioma delineation.

Head and neck squamous cell carcinoma

Advanced stage head and neck squamous cell carcinoma (HNSCC) refers to malignant tumours that arise from the mucosal surface of the upper aero-digestive tract including oral cavity, oropharynx, nasopharynx, hypopharynx and larynx. Approximately 4% of all

cancer cases are HNSCC.[5] The disease is more common in men than in women and major risk factors include tobacco and alcohol use and infection with the human papillomavirus (HPV).[6] The latter also provides a clinical distinction between two separate types of HNSCC.[7] Differences between HPV-positive and HPV-negative disease include survival and epidemiology. HPV-positive patients are generally younger and have a favourable prognosis. Patients with HPV-negative disease are generally over 50 years of age and five-year overall survival ranges from 32% to 58% for advanced stage disease.

About two thirds of the patients present with advanced stage disease and treatment will often consist of a combination of radiation and chemotherapy. When treatment fails (approx. 50% of the cases[8]), salvage surgery is often the last option. Due to initial treatment, wound healing is slowed and complications of salvage surgery are likely, causing high morbidity. With early treatment response prediction, patients can be selected for whom surgery is a better option and the side effects of radiation therapy can be avoided. Moreover, the response prediction may be used to create personalized (chemo)radiotherapy schemes and improve treatment success rate.

Treatment decisions are based on tumour type and characteristics. Tumour typing and prognosis prediction are performed with an array of tools. Imaging by FDG-PET and MRI sequences (T1, T2, post contrast T1, ADC) is prominent among those tools. Imaging is used not only for tumour localisation, but also for characterizing tumour properties through imaging biomarkers. DCE MRI is under investigation as a means to quantify tumour permeability characteristics. Extension of the DWI sequence is investigated to be able to correct for perfusion related effects by modelling intravoxel incoherent motions (IVIM). These biomarkers are potential candidates for early treatment response prediction to improve treatment success.

Quantification of tracer uptake and kinetics

Tracer kinetic modelling – temporal aspects

Tracer kinetic modelling is a method to estimate quantitative imaging biomarkers, in this case tissue pharmacokinetic properties, by monitoring an injected tracer. Tracer imaging is generally performed using PET; however, SPECT, MRI or CT can also be used. The premise involves simplifying the human body to a set of compartments: one or more blood or plasma compartments and one or more tissue compartments. Each compartment undergoes a change in tracer concentration and these changes are coupled between compartments. For example, a tracer is injected into the blood. When the blood transports the tracer throughout the body, the tracer gets into the tissue, the second compartment. The tissue concentration goes up, while the concentration in the blood goes down. Kinetic rate constants of transport between the compartments are estimated based on the measured tracer signal intensities in the compartments. These rate constants

are important, because they can be converted to a functional quantitative biomarker by interpretation of the model, e.g., tissue perfusion or receptor density.

An important aspect for accurate analysis is the measurement of the arterial input function, i.e., the concentration curve inside the blood compartment. Deriving this input function directly from the image is only possible when a large arterial volume, such as the aorta, is within the field of view. Without an alternative, the tracer concentration in the blood must be measured by (continuous) arterial blood sampling, which is burdensome for the patient.

Tracer kinetic modelling in human patient trials is focused on defining and applying the optimal clinical imaging protocol. This is generally a compromise between parameter precision and accuracy against patient burden and cost; a simple imaging protocol, while providing little burden to the patient, may provide poor parameter accuracy, i.e., results are not trustworthy and of little use. The same is possible for the opposite. For some patients the burden of a complicated protocol can be too high.

Radiomics – spatial aspects

Sometimes a phenomenon may seem unpredictable, or indeed chaotic, simply because we fail to note particularities. If we could, we might be able to predict treatment outcomes based on information that is not easy to see. Radiomics aims to extract such information from medical images. Whereas kinetic modelling is used to parametrise temporal signals, radiomics is used to parametrise spatial patterns.

Cancerous tissue growth can become so uncontrolled that the tissue structure becomes chaotic. Such chaos is then a clear sign that the order imposed by our biological systems is lost. This disorder can be quantified as entropy, which has a mathematical definition for digital data.[9] As such, this potential biomarker is easily derived from medical images.

Along with entropy, hundreds of other radiomic features can be extracted, ranging from shape based- to fractal features. These features are then subjected to statistical analysis. It is sometimes stated that radiomics is the combination of radiomic feature extraction and machine learning. One could argue, however, that radiomics *is* a form of machine learning, since feature extraction is an integral part of machine learning.

Outline

This thesis covers several studies on methods that quantify biomarkers by image processing of time varying signals and by means of radiomics.

- Chapter two looks into the necessity of arterial input measurement for brain perfusion measurements with $[^{15}O]H_2O$ PET.
- Chapters three and four describe the details of establishing the optimal kinetic model for $[^{18}F]$FET PET in glioma.
- Chapter five investigates the precision of image-derived arterial input functions obtained with DCE-MRI in head and neck cancer patients.
- Chapter six compares several methods for image processing of DWI-MRI data to estimate IVIM parameters in the head and neck region.
- Chapter seven deals with radiomics and the problem of combining parameters for prediction of treatment success in head and neck squamous cell carcinoma.

2

Semi-quantitative CBF parameters derived from non-invasive [^{15}O]H_2O PET studies

Thomas Koopman, Maqsood Yaqub, Dennis F.R. Heijtel, Aart J. Nederveen, Bart N.M. van Berckel, Adriaan A. Lammertsma, Ronald Boellaard

Published 13 September 2017
Journal of Cerebral Blood Flow & Metabolism 2019; 39: 163–172
DOI: 10.1177/0271678X17730654

Abstract

Quantification of regional cerebral blood flow (CBF) using [^{15}O]H_2O positron emission tomography (PET) requires the use of an arterial input function. Arterial sampling, however, is not always possible, for example in ill-conditioned or paediatric patients. Therefore, it is of interest to explore the use of non-invasive methods for the quantification of CBF. For validation of non-invasive methods, test–retest normal and hypercapnia data from 15 healthy volunteers were used. For each subject, the data consisted of up to five dynamic [^{15}O]H_2O brain PET studies of 10 min and including arterial sampling. A measure of CBF was estimated using several non-invasive methods earlier reported in literature. In addition, various parameters were derived from the time activity curve (TAC). Performance of these methods was assessed by comparison with full kinetic analysis using correlation and agreement analysis. The analysis was repeated with normalization to the whole brain grey matter value, providing relative CBF distributions. A reliable, absolute quantitative estimate of CBF could not be obtained with the reported non-invasive methods. Relative (normalized) CBF was best estimated using the double integration method.

Introduction

Regional cerebral blood flow (CBF) represents the amount of blood that perfuses a volume of tissue, i.e. mL blood per mL of tissue per min. To date, the accepted unit for CBF is $mL \cdot cm^{-3} \cdot min^{-1}$, where $mL \cdot cm^{-3}$ is used to indicate the transfer from blood to tissue.[10] Several modalities can be used to measure perfusion[11], but positron emission tomography (PET) with oxygen-15 labelled water is considered to be the reference standard method.

Over the years, various methods have been developed for deriving CBF from a dynamic $[^{15}O]H_2O$ PET scan.[12-23] Ultimately, all these methods are based on Kety's compartment model for (inert) H_2O.[24] Solving the differential equation leads to a convolution of the tissue response with the arterial input function (AIF). Quantitative studies therefore require the measurement of the AIF, which is obtained most reliably through continuous arterial sampling.[25] However, this is a somewhat invasive procedure, which is less suitable for routine clinical studies. In some cases, arterial sampling is clinically not feasible, for example in ill-conditioned patients or in children with Moyamoya disease. In case arterial sampling is not possible, it may be of interest to apply non-invasive methods that can estimate CBF or relative CBF distributions across the brain to identify regions with reduced perfusion. However, before using non-invasive methods, it is important to investigate their accuracy and precision against full-quantitative kinetic approaches.

One non-invasive approach is the use of an image derived input function (IDIF).[26] The main challenge for this approach is the limited spatial resolution inherent to PET. This affects the quality of the measured input function due to partial volume effects. This particularly affects CBF studies because, in contrast to myocardial blood flow studies, there are no large vascular structures within the field of view.[26] The IDIF approach seems very promising, but requires complex and accurate methodology for partial volume correction and delineation of the arteries. As a consequence, use of IDIF for CBF measurements is not widely used and had limited success so far.

Instead, this study focuses on validation of simplified methods independent of a measured AIF. As far as we know, the methods described below are all that have been reported for $[^{15}O]H_2O$ PET.[19,20,22] We also included the integral count approach from early brain activation studies.[27] In MR perfusion research often additional parameters describing the time intensity curve are reported, such as the time-to-peak (TTP), wash-in slope and the peak height. Their equivalents for PET have not been evaluated, because typically in dynamic PET studies, frame times are not shorter than 5 seconds and data are noisy, making it difficult to estimate these parameters reliably. These parameters were included to confirm this hypothesis.

The aim of this study was to select the best method based on their performance to estimate (relative) CBF, initiated by our interest in studying CBF changes in children suffering from Moyamoya disease for whom arterial sampling is clinically not feasible. In a head-to-head comparison with the reference kinetic method and including a report on the test–retest variability, this paper should provide clarity on the best non-invasive method for obtaining (relative) CBF.

Material and Methods

Subjects and study protocol

PET scans were acquired on a Gemini TF-64 PET/CT system (Philips Healthcare, Cleveland, TN, USA). The research participants were 16 healthy volunteers. All participants gave written informed consent for this study prior to inclusion. The study has been approved by the medical ethical review committees of both participating centers; the Amsterdam Medical Center and the VU University Medical Center. The study was conducted in accordance with the Declaration of Helsinki. Characteristics of the participants and scanning protocol have been described previously[28]. In brief, each person underwent five [^{15}O]H_2O scans in two separate scanning sessions. During the first scan session people were scanned twice under baseline cerebrovascular conditions and once during hypercapnia. During the second session, planned 28 days later, a single baseline scan was followed by a hypercapnia scan. Hypercapnia was induced using 5% CO_2 enriched air.

Each person received an arterial line for blood sampling and a venous line in the opposite arm for administration of [^{15}O]H_2O. Scanning sessions started with a 1 minute low-dose CT scan, which served for attenuation and scatter correction of the subsequent PET acquisitions. Emission scans were acquired in list mode for 10 minutes. A bolus of 800 MBq [^{15}O]H_2O was administered at the start of each scan. Arterial blood was measured continuously using an automatic blood sampler.[25] The resulting arterial input function was calibrated using three manual arterial samples collected at 5.5, 8 and 10 minutes post injection.

The scans were acquired in list mode and reconstructed into 26 frames of 1×10 s, 8×5 s, 4×10 s, 2×15 s, 3×20 s, 2×30 s, 6×60 s. The row action maximum likelihood algorithm (RAMLA) as provided by the scanner manufacturer was used for reconstruction of the scans with an isotropic voxel size of 2 mm. Thereafter, the dynamic images were smoothed using an isotropic 5 mm FWHM Gaussian kernel.

Brain region time activity curves

T1-weighted MR images were acquired on a Philips 3 T Intera System (Philips Healthcare, Best, The Netherlands) at the Amsterdam Medical Center. This anatomical reference scan

of each subject was co-registered to the emission scans using the SPM12 software package (Functional Imaging Laboratory, 2014, London, UK). For this purpose, emission scans were summed over all time frames. After co-registration, the anatomical scans were segmented using PVELab (Neurobiology Research Unit, 2010, Copenhagen, Denmark) into grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) and divided into 67 brain regions using the Hammers brain atlas.[29,30] Segmentations were then applied to the dynamic [^{15}O]H_2O images to generate regional GM Time Activity Curves (TACs). Furthermore, the union of GM and WM was used as whole brain mask for some methods; this will further be referred to as whole brain.

CBF reference methods

CBF was calculated using two reference methods: (a) full kinetic analysis of brain region TACs using non-linear regression (NLR) and (b) the basis function method (BFM)[23] for voxel wise (=parametric) calculations. Thus, NLR gives the CBF per region, and BFM gives a parametric map of CBF. Both methods used the single tissue compartment model with additional arterial blood volume parameter:

$$C_t(t) = V_a \cdot C_a(t) + (1 - V_a) \cdot f \cdot e^{(-f \cdot t \cdot V_T^{-1})} \otimes C_a(t) \tag{1}$$

Here $C_t(t)$ is the tissue concentration of the tracer over time, V_a the arterial blood volume fraction, $C_a(t)$ the arterial input function, f the cerebral blood flow (f=CBF) and V_T the volume of distribution. For NLR the arterial input function is estimated by the measured arterial tracer concentration corrected for delay. Dispersion correction was omitted in favour of fitting the arterial blood volume parameter. Blood flow estimated with this parameter is near equivalent to blood flow estimated with dispersion correction, as noted by Bol et al.[31] For BFM the measured arterial tracer concentration is corrected for both delay and dispersion as described previously.[23]

Simplified methods

Implementation of the methods described in this paper used the following calculations as published in the original papers. All methods are based on Kety's differential equation for the one-tissue reversible compartment model, see Equation 2, where $C_t(t)$ is the tissue concentration of the tracer over time, $C_a(t)$ is the arterial tracer concentration over time,f is the cerebral blood flow and V_T the volume of distribution.

$$\frac{dC_t(t)}{dt} = fC_a(t) - \frac{f}{V_T}C_t(t) \tag{2}$$

In 1994 Mejia et al. introduced the double integration method (DIM)[19], which eliminates the need for the arterial input function by using the whole brain as a reference and assuming the global CBF to be the normal average value. The method is based on the

double integration of Equation 2, leading to Equation 3, where end time T is 3 minutes (in accordance with the original method).

$$f = \frac{\int_0^T C_t(t)dt}{\int_0^T \int_0^t C_a(u)du\,dt - \frac{1}{V_T}\int_0^T \int_0^t C_t(u)du\,dt} \tag{3}$$

The double integration of the arterial tracer concentration is substituted by A, which is estimated using the time activity curve (TAC) of the whole brain $C_{t1}(t)$ and an assumed global CBF f_1, as described in Equation 4.

$$A = \int_0^T \int_0^t C_a(u)du\,dt = \frac{\int_0^T C_{t1}(t)dt}{f_1} + \frac{1}{V_{T1}}\int_0^T \int_0^t C_{t1}(u)du\,dt \tag{4}$$

A is calculated with $f_1 = 0.5$ mL·cm^{-3}·min^{-1} and $V_{T1} = 0.86$. With A substituted in Equation 3 the flow is calculated for every voxel, fixing V_T at 0.86. Note that in the original publication V_{T1} and V_T were fixed at unity, however a later (1992) recommended value is used here.[32]

In 1996 the DIM was extended by Watabe et al. using a two-step calculation strategy to estimate the global CBF and volume of distribution instead of fixing them to the normal average value.[20] In Watabe's method the same substitution is performed, and the TAC of a second reference region $C_{t2}(t)$ is introduced, yielding Equation 5. This second region was defined as the 10% of voxels with the highest number of total counts.

$$f_2 = \frac{\int_0^T C_{t2}(t)dt}{\frac{1}{f_1}\int_0^T C_{t1}(t)dt + \frac{1}{V_{T1}}\int_0^T \int_0^t C_{t1}(u)du\,dt - \frac{1}{V_{T2}}\int_0^T \int_0^t C_{t2}(u)du\,dt} \tag{5}$$

Fixing V_{T2} at 0.86, non-linear regression is performed using a trust-region reflective algorithm to fit the values for f_1, V_{T1} and f_2. A is then calculated using the fitted values for f_1 and V_{T1} and substituted in Equation 3 to calculate f for every voxel, again fixing V_T at 0.86.

Another approach was published by Treyer et al. in 2003. It is based on Alpert's weighted integration method[33] to estimate both CBF and the washout parameter k_2, but uses a standard AIF.[22] This standard input function is corrected for delay and dispersion (Meyer's method[17]). Because CBF depends on the amplitude of the AIF, and k_2 does not, estimated CBF values were then scaled using the estimated k_2 by making their averages in grey matter equal. However, as Treyer at al. note in their discussion, this causes a bias because the washout parameter is defined as $k_2 = \frac{f}{V_T}$. Therefore, in this study the estimated CBF values were scaled by making the estimate grey matter value equal to the

average k_2 times V_T, fixing V_T at 0.86. Unlike in the original study, we did not use a separate set of measured input functions to create a standard input function. Instead, the measured input functions of the subjects was used. However, to ensure that the used input function was independent of the subject, a 'standard' input function was produced per subject by averaging the input functions of all other subjects. Before averaging, the functions were normalized by their integral and the time of the peaks was aligned, as was done in the original study.[22]

For the above methods the first 3 minutes of scan data was used; the following TAC derived parameters used all or part of the 10 minutes scans. The definitions of the TAC derived parameters are illustrated in the supplementary material. These CBF related parameters were derived from the TAC: the area under the time activity curve (AUC); the AUC for a 60 second interval (AUC^{60}) after the beginning of the wash-in (t_0); peak height (peak); time between the beginning of the wash-in and the peak, or time-to-peak (TTP); the maximum wash-in slope (slope); the wash-out curve fitted with an exponential function ($washout^{EXP}$) and a power function ($washout^{powerlaw}$). The parameters were calculated using Matlab 7.10 (R2010a) (The MathWorks, Inc., Natick, MA, USA).

Evaluation of simplified methods

The various parameters associated with CBF were compared with the reference method using both correlation and agreement analysis. Firstly, using the simplified CBF estimates per brain region and NLR as reference, Pearson correlation coefficients were calculated. The five highest correlating simplified parameters were included for further investigation.

Secondly, Bland-Altman analysis[34] was used to investigate agreement. The results are reported as the mean difference (an estimate of the bias) and 1.96 times the standard deviation of the differences (an estimate of the precision). The results do not focus on individual regions, but rather report the average agreement for brain regions. To include the parameters with different units, the CBF estimates of each method were converted to a percentage of the sample average—the average over all subjects—and the standard deviation of the differences were reported in percentage points. Note that the mean difference is now zero and hence not reported. In addition, brain region values of all methods were normalized to their whole brain value providing relative CBF, and the agreement of the parameters relative to whole brain was also investigated. Using BFM as the reference, the agreement analyses were repeated using the parametric maps.

Repeatability of all methods was investigated using the brain region values of the repeated baseline measurements. The repeatability performance is reported by the repeatability index (RI), the repeatability coefficient[34] as percentage of the sample average,[28] again allowing to compare the repeatability of metrics with different units. The repeatability was re-calculated after normalizing each scan to the whole brain value. Intra-session test–

retest performance was investigated using the two consecutive baseline scans of the first scanning session. For inter-session repeatability, the baseline scan of the second session was used in combination with the first baseline scan of the first session. The 95% confidence interval of the repeatability indices is also reported.

Finally, agreement analysis was performed on hypercapnia-induced differences estimated by each method. The relative changes between the baseline and hypercapnia scans, see Equation 6, were calculated with each method for all brain regions. For this, differences between scan 3 and scan 1 and between scan 5 and scan 4 were used. These differences were compared to the differences found by the reference method (NLR) and the average difference between them is reported in percentage points.

$$\frac{hypercapnia - baseline}{baseline} \cdot 100\% \tag{6}$$

Simulations

Simulations were performed in order to better understand the behaviour of the CBF methods. TACs were generated without noise to investigate bias as function of simulated CBF. The TACs were generated with the single tissue compartment model, as described in Equation 1, and a representative input function. The representative input function was constructed from all measured input functions after normalization by their integral and aligning the peaks temporally. The volume of distribution V_T and the arterial blood volume fraction V_a were kept constant and simulated CBF parameter f ranged from 0.2 to 1.0 in 128 steps. In addition, noise was added to the TACs to investigate the noise characteristics of the methods. The noise level ranged from 0% to 16% coefficient of variation (COV) and was increased in 128 steps. Details of the noise simulation have been described earlier.[35] For every combination of noise and CBF 500 noisy TACs were generated.

The simulated data were then analysed with the BFM and simplified methods to estimate CBF. The whole brain reference TAC for the DIM was a noise free TAC generated with $f = 0.5$ mL·cm^{-3}·min^{-1}. Watabe's extension of the DIM was not investigated. The scale factor for Treyer's method was calculated with the average k_2 and K_1 of all generated TACs.

The mean observed CBF errors and their standard deviations are reported in error maps. The errors as percentage of simulated CBF are reported as relative error maps. Error plots are shown for the noise free TACs and the TACs with 8% noise level.

Results

Average time between sessions was 34 days (25-45 days). From the total of 80 scans, 70 scans were successfully evaluated. Acquisition failed for 10 scans of 5 volunteers: tracer production failure twice, inadequate arterial blood sampling twice, and acute nausea once. Details were reported elsewhere.[28]

The Pearson correlation coefficients per scan between regional CBF values of simplified metrics and the reference method (NLR) are presented in Figure 1. Results obtained with BFM were added for comparison. Highest correlations were found for DIM, Watabe, Treyer, AUC^{60}, and the peak height. The other measures ($washout^{EXP}$, $washout^{powerlaw}$, AUC, slope and TTP) showed lower correlation and were excluded from further analysis.

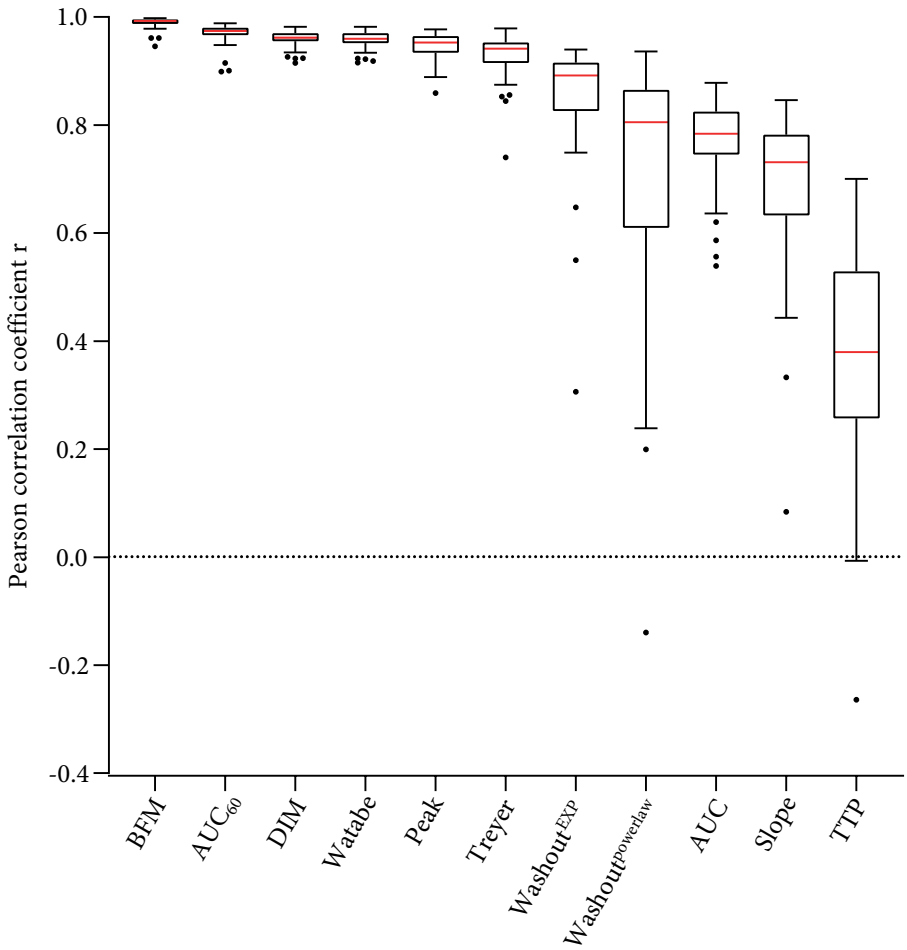


Figure 1: Correlation per scan of the investigated methods with full kinetic analysis using NLR.

The results of the agreement analysis using brain regions are summarized in Table 1. The mean difference was zero for the DIM and Watabe's method. Treyer's method showed a mean difference that was significantly different from zero ($\alpha<.001$), indicating bias. The standard deviations of the differences are an estimation of precision and show that for the absolute estimation of CBF, the DIM and Treyer's method are most precise, followed by the AUC^{60} and peak height. For estimation of relative CBF, the DIM and the AUC^{60} are most precise, followed by the Treyer's method and peak height. Watabe's method was most imprecise for estimation of both absolute and relative CBF.

Table 2 shows the same results, but for parametric comparison. The standard deviations of the differences are larger than for regional comparison. The DIM seems least affected by this.

Table 1: Regional agreement with NLR.

Method	Mean		1.96 SD		1.96 SD (% of sample average)		Relative CBF 1.96 SD (% of global CBF)	
BFM	0.00	(-0.01 – 0.01)	0.08	(0.06 – 0.10)	16.6	(13.1 – 20.1)	5.3	(5.1 – 5.4)
DIM	0.01	(-0.01 – 0.03)	0.18	(0.14 – 0.21)	36.6	(28.9 – 44.4)	10.4	(10.2 – 10.7)
Watabe	0.00	(-0.03 – 0.03)	0.27	(0.22 – 0.33)	56.7	(44.7 – 68.6)	15.4	(15.0 – 15.8)
Treyer	0.12	(0.09 – 0.14)	0.21	(0.17 – 0.25)	37.5	(29.6 – 45.4)	13.0	(12.7 – 13.3)
AUC60	N/A		N/A		44.0	(34.7 – 53.3)	10.2	(10.0 – 10.5)
peak	N/A		N/A		45.2	(35.7 – 54.7)	13.2	(12.8 – 13.5)

Note: Average over all brain regions and 95% confidence interval between brackets. The third column is converted to percentages by dividing over the average of all subjects, relative CBF is a percentage of the global CBF per scan.

Table 2: Parametric agreement with BFM.

Method	Mean		1.96 SD		1.96 SD (% of sample average)		Relative CBF 1.96 SD (% of global CBF)	
DIM	0.01	(-0.02 – 0.04)	0.23	(0.19 – 0.27)	45.4	(37.7 – 53.1)	27.0	(22.4 – 31.6)
Watabe	0.03	(-0.03 – 0.10)	0.52	(0.43 – 0.61)	96.7	(80.3 – 113.1)	29.0	(24.1 – 33.9)
Treyer	0.10	(0.06 – 0.13)	0.29	(0.24 – 0.34)	50.7	(42.1 – 59.3)	34.7	(28.8 – 40.6)
AUC60	N/A		N/A		52.0	(43.2 – 60.8)	28.1	(23.3 – 32.9)
peak	N/A		N/A		55.0	(45.6 – 64.3)	35.4	(29.4 – 41.4)

Note: Average over all brain regions and 95% confidence interval between brackets.

Figure 2 shows voxel wise scatter and Bland-Altman plots of relative CBF calculated with the DIM, Treyer's and AUC^{60} methods using BFM as reference. Mean differences were zero due to normalization to the whole brain average. An example of parametric maps calculated with the various methods can be found in the supplementary material.

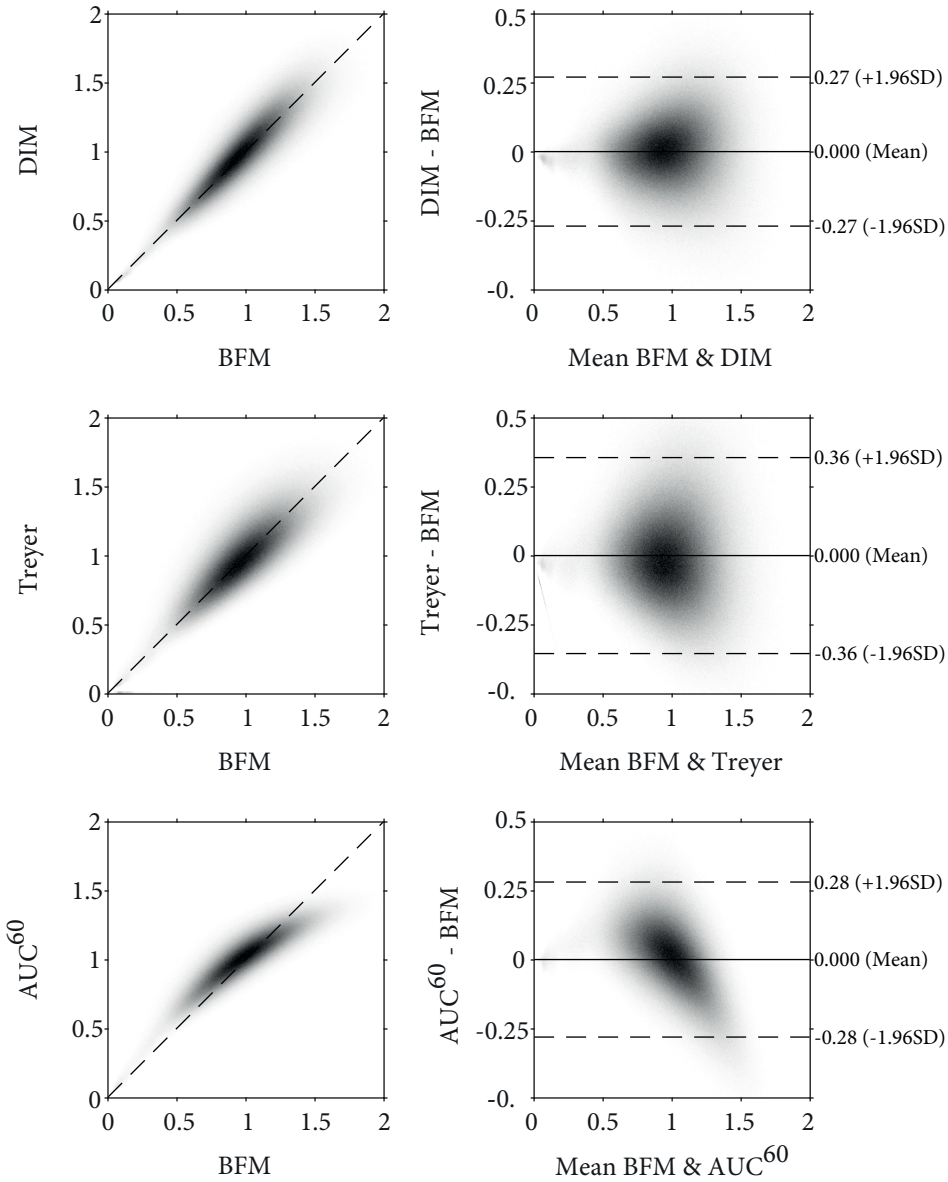


Figure 2: Voxel wise scatter and Bland-Altman plots of the methods vs BFM after normalization to whole brain (relative CBF).

Repeatability performance is shown in Table 3 for intrasession (n=14 subjects) and Table 4 for intersession (n=14 subjects). The DIM shows the same results, both with and without normalisation to the whole brain. Results of Watabe's method shows the largest RI for intrasession repeatability. The TAC derived parameters (AUC^{60} and peak height) show the best reproducibility indices for relative CBF estimation. The intersession RIs are approximately twice as large as the intrasessions RIs for Treyer's method, AUC^{60} and peak height.

Table 3: Intrasession test–retest repeatability.

		Relative CBF
Method	RI (%)	RI (%)
NLR	27.3 (25.7 – 28.9)	14.4 (13.6 – 15.3)
BFM	26.1 (24.6 – 27.6)	15.2 (14.3 – 16.0)
DIM	12.8 (12.1 – 13.5)	12.8 (12.1 – 13.5)
Watabe	37.2 (35.1 – 39.3)	14.4 (13.6 – 15.2)
Treyer	21.5 (20.2 – 22.7)	15.6 (14.7 – 16.5)
AUC^{60}	23.5 (22.2 – 24.8)	10.3 (9.7 – 10.8)
peak	23.2 (21.9 – 24.5)	12.3 (11.6 – 13.0)

Note: Data of 14 subjects, average over all brain regions and 95% confidence interval between brackets.

Table 4: Intersession test–retest repeatability.

		Relative CBF
Method	RI (%)	RI (%)
NLR	29.5 (27.8 – 31.3)	14.4 (13.5 – 15.2)
BFM	31.8 (30.0 – 33.6)	15.0 (14.1 – 15.8)
DIM	14.3 (13.5 – 15.2)	14.3 (13.5 – 15.2)
Watabe	36.9 (34.8 – 39.0)	17.2 (16.2 – 18.2)
Treyer	43.9 (41.4 – 46.4)	16.6 (15.7 – 17.5)
AUC^{60}	49.8 (46.9 – 52.6)	11.7 (11.0 – 12.3)
peak	47.5 (44.8 – 50.1)	12.9 (12.2 – 13.7)

Note: Data of 14 subjects, average over all brain regions and 95% confidence interval between brackets.

Agreement on hypercapnia-induced differences is presented in Table 5. Scatter and Bland-Altman plots of these differences are shown in the supplementary material. The DIM and Watabe's method show a clear disagreement with the reference method for estimating differences between the hypercapnia scans and the baseline scans. Treyer's method shows best performance among the simplified methods.

Table 5: Regional agreement with NLR of hypercapnia induced differences.

Method	Mean	1.96 SD
BFM	2.4 (-1.6 – 6.3)	32.7 (25.8 – 39.6)
DIM	-24.5 (-29.5 – -19.5)	41.1 (32.5 – 49.8)
Watabe	-31.9 (-37.7 – -26.2)	47.2 (37.3 – 57.2)
Treyer	1.6 (-3.0 – 6.3)	38.0 (30.0 – 46.1)
AUC^{60}	-3.7 (-11.0 – 3.6)	59.9 (47.2 – 72.5)
peak	-4.1 (-11.5 – 3.3)	60.9 (48.0 – 73.7)

Note: Average over all brain regions and 95% confidence interval between brackets.

The error maps and plots showing the simulation results are available in the supplementary material. BFM shows the least bias, only slightly overestimating CBF for simulated CBFs >0.8 $mL \cdot cm^{-3} \cdot min^{-1}$. The sawtooth pattern visible on the BFM error graph for the simulation without noise is caused by the use of the limited number of basis functions. The DIM shows slight overestimation at low CBF and slight underestimation

at high CBF. The AUC^{60} and Peak methods largely underestimate and overestimate CBF at low and high simulated CBF values. The bias of the peak method also shows dependence on noise, which is visible in the error maps, whereas the bias of all other methods are independent on noise. Treyer's method shows overestimation of CBF over the entire simulated CBF range, but increases with simulated higher CBF values

The error precision maps show that precision of the AUC^{60} method is constant at different CBF and all other methods show declining precision with increasing CBF. All methods show worse precision with increasing noise levels. For BFM and the DIM the precision is proportional to CBF, which can be seen in the relative precision maps.

Discussion

This study compared a wide range of simplified methods for estimating (absolute and relative) cerebral perfusion, independent of measurement of the AIF, in healthy volunteers. Their performance was investigated against reference kinetic methods, which use an arterial input function. Moreover, our study included assessment of repeatability performance of all metrics and methods tested; both intra- and inter-session.

Most TAC derived parameters ($washout^{EXP}$, $washout^{powerlaw}$, AUC, slope and TTP) showed poor correlation with NLR derived CBF. As expected, these parameters are thus of little value for estimating CBF. Two TAC derived parameters, the peak height and the AUC^{60}, do show high correlation (see Figure 1). The AUC^{60} showed better results than the peak height: higher correlation, lower RIs and smaller standard deviation of the differences. Peak height does show a linear relationship, yet its performance is worse.

Relative CBF distributions can be estimated with reasonable precision using the DIM and AUC^{60} methods. However, it is known that the relationship between integral counts and CBF is nonlinear. This can also be seen in Figure 2 and this causes the lower contrast, which can be seen in the parametric maps (supplementary material). This is probably also why the AUC^{60} has a lower RI for both inter- and intra-session repeatability. Because the DIM does not show worse agreement performance, it should be the method of choice for estimating relative CBF. However, because the global CBF is always assumed to be the normal average, this method cannot estimate absolute CBF and should only be considered when studying relative CBF changes between subjects or longitudinally. This is exemplified by the disagreement of this method for absolute longitudinal changes, as presented in Table 5.

In our study, we observed that none of the non-invasive methods are able to estimate absolute CBF reliably. Watabe's method estimates global CBF using NLR. However, as they clearly explain in the original paper, there exists a very shallow error surface around the optimal solution. Hence, the method is very sensitive to getting trapped in local

minima, and Treyer et al. indeed report this as well. Furthermore, Watabe et al. mention in their discussion that "From the simulation it follows that a 2-min administration period performs better than a 1-min period."[20] Perhaps this explains the disparity in the results. In this study (and in the study of Treyer et al. too) a bolus injection was used, which had an even shorter administration period of 15 seconds (20 s in the study of Treyer et al.). Clearly, the method performs worse on bolus injection data and cannot be recommended for estimation of CBF for our imaging procedure.

In comparison with Watabe's method, Treyer's method shows better precision for estimating absolute CBF. However, Treyer's method showed overestimations of CBF in this dataset. The reason for this is unclear, but could have to do with the presumed volume of distribution. If we look at the precision in percentage points, it is clear that the precision of Treyer's method is similar to the DIM's precision, whereas the TAC derived parameters have a worse precision. For the assessment of CBF and relative CBF changes most simplified methods show similar RIs the reference methods (NLR or BFM with arterial sampling) for intra-session CBF and relative CBF data and somewhat worse RIs for inter-session CBF values. When short-term longitudinal changes in CBF need to be assessed Treyer's method may be considered.

The simulations largely confirmed the observations seen in the clinical data. Generally, the BFM provided most accurate and robust CBF estimates, while several simplified non-invasive methods suffer from substantial bias and poor precision. In line with the clinical studies the DIM seems to be able to estimate CBF accurately and with high precision over a large range of simulated CBF values and noise levels and comparable to those seen with BFM. It should be noted that we did not simulate deviations in volume of distribution or input functions and some observations for the simplified methods may therefore be more optimistic than seen in the clinical data. Yet, in general the simulations show the same trends as seen in the clinical studies.

Conclusion

In this study we evaluated the performance of a wide range of non-invasive methods for quantifying CBF and/or relative CBF which can be applied in studies were the collection of an arterial input function is clinically not feasible (e.g. in children with Moyamoya disease). Performance of these methods was compared with quantitative CBF derived using a kinetic model including an arterial input function. The double integration method showed the best performance for measuring relative cerebral perfusion (and its change) without arterial sampling. The main disadvantage of this method is the inability to estimate global CBF. Therefore, it is concluded that among the non-invasive methods tested the double integration method seems to be most optimal method for measuring relative CBF. None of the non-invasive methods were able to measure absolute CBF accurately, but Treyer's method may be considered when studying changes in CBF within the same subject in a longitudinal setting.

Supplementary material

Supplementary material for this paper is available at:

doi.org/10.1177/0271678X17730654

3

Quantification of O-(2-[^{18}F]fluoroethyl)-L-tyrosine kinetics in glioma

Thomas Koopman, Niels Verburg, Robert C. Schuit, Petra J.W. Pouwels, Pieter Wesseling, Albert D. Windhorst, Otto S. Hoekstra, Philip C. de Witt Hamer, Adriaan A. Lammertsma, Ronald Boellaard, Maqsood Yaqub

Published 31 July 2018
EJNMMI Research 2018; 8: 72
DOI: 10.1186/s13550-018-0418-0

Abstract

Background

This study identified the optimal tracer kinetic model for quantification of dynamic O-(2-[^{18}F]fluoroethyl)-L-tyrosine ([^{18}F]FET) positron emission tomography (PET) studies in seven patients with diffuse glioma (four glioblastoma, three lower grade glioma). The performance of more simplified approaches was evaluated by comparison with the optimal compartment model. Additionally, the relationship with cerebral blood flow—determined by [^{15}O]H_2O PET—was investigated.

Results

The optimal tracer kinetic model was the reversible two-tissue compartment model. Agreement analysis of binding potential estimates derived from reference tissue input models with the distribution volume ratio (DVR)-1 derived from the plasma input model showed no significant average difference and limits of agreement of -0.39 and 0.37. Given the range of DVR-1 (-0.25 to 1.5) these limits are wide. For the simplified methods, the 60-90 min tumour-to-blood ratio to parent plasma concentration yielded the highest correlation with volume of distribution V_T as calculated by the plasma input model (r=0.97). The 60-90 min standardized uptake value (SUV) showed better correlation with V_T (r=0.77) than SUV based on earlier intervals. The 60-90 min SUV ratio to contralateral healthy brain tissue showed moderate agreement with DVR with no significant average difference and limits of agreement of -0.24 and 0.30. A significant but low correlation was found between V_T and CBF in the tumour regions (r=0.61, p=0.007).

Conclusion

Uptake of [^{18}F]FET was best modelled by a reversible two-tissue compartment model. Reference tissue input models yielded estimates of binding potential which did not correspond well with plasma input derived DVR-1. In comparison, SUV ratio to contralateral healthy brain tissue showed slightly better performance, if measured at the 60-90 minute interval. SUV showed only moderate correlation with V_T. V_T shows correlation with CBF in tumour.

Background

Since its introduction in 1999[36] the amino acid tracer O-(2-[^{18}F]fluoroethyl)-L-tyrosine ([^{18}F]FET) is increasingly used to image glioma.[37] Because [^{18}F]FET is not incorporated into proteins, it is a tracer for amino acid transport rather than for protein synthesis rate.[36,38] [^{18}F]FET positron emission tomography (PET) has shown its added value to magnetic resonance imaging (MRI) for several clinical problems regarding brain tumours, such as prognosis assessment, delineation of tumour extent and glioma grading.[39]

The most extensive quantitative analysis of a PET tracer is based on dynamic PET scans in combination with plasma input based pharmacokinetic modelling.[40] For large clinical studies, such an extensive analysis is not feasible; tracer uptake needs to be quantified using simplified measures. For example, the standardized uptake value (SUV) interval of 20-40 minutes post injection is currently recommended for clinical reading in European Association of Nuclear Medicine and German guidelines[41,42]. Simplified approaches are not only affected by regulation of specific amino acid transporters—the primary parameter of interest—but also by the blood flow and plasma concentration, which is in turn affected by the biodistribution, tracer metabolism, and uptake in blood cells. It is of interest to quantify these effects to gain a better understanding of the accuracy of a simplified measure and its reliability.

In the current literature, we identified five studies which used pharmacokinetic modelling to quantify uptake of the tracer in the brain; two preclinical studies[43,44] and three human studies[45-47]. The human studies all used an image derived input function. Furthermore, we found only one study where metabolite concentration in plasma was measured.[48] The tracer kinetics of [^{18}F]FET in glioma patients are expected to be in line with preclinical research, but validation of kinetic models is needed. The aim of this study was therefore to identify the optimal metabolite corrected plasma input model for the quantification of [^{18}F]FET kinetics. In addition, reference tissue input models and several simplified methods were validated in terms of their agreement with full kinetic analysis results. Lastly, the relationship of the methods and parameters with blood flow were investigated using [^{15}O]H_2O PET data.

Methods

Subjects and study protocol

The study population consisted of seven patients with diffuse glioma from an ongoing patient study.[49] Each patient gave written informed consent prior to inclusion. This study has been performed in accordance with the Declaration of Helsinki, approved by the Medical Ethics Committee of the VU University Medical Center and registered in the

Netherlands National Trial Register (www.trialregister.nl, unique identifier NTR5354, registration date 4th of August 2015). The age of the patients ranged from 22 to 69 years. All gliomas were newly diagnosed and selected for resective surgery. Imaging was preoperatively performed. Based on histology of biopsies taken before surgery—but after imaging—each glioma was classified according to World Health Organization (WHO) criteria as lower grade (WHO II-III) or glioblastoma (WHO IV).[50] Four patients presented with glioblastoma, three with lower grade glioma. See supplemental Table S1 for more details.

The patients were required to fast at least 4 hours before undergoing the imaging protocol. T1-weighted gadolinium-enhanced (T1G) and FLAIR sequences were acquired on an Achieva whole-body 3.0T MR-scanner (Philips Healthcare, Best, the Netherlands). Details of the MR sequences are described in the supplemental material. Two dynamic PET scans were acquired on either a Gemini TF-64 PET/CT or an Ingenuity TF PET/CT (Philips Healthcare, Best, the Netherlands). Each scan started with a low dose computed tomography (CT) scan (30 mAs, 120 kVp) for attenuation and scatter correction purposes. A bolus of 800 MBq [^{15}O]H_2O was administered at the start of the first scan with a venous line and emission scans were acquired in list mode for 10 minutes. An arterial line in the opposite arm was used for continuous sampling using an on-line blood sampler (Comecer Netherlands, Joure, the Netherlands). Manual arterial samples were collected at 5, 7 and 9 minutes. A 90 minute dynamic scan was then acquired on the same system after a bolus of 200 MBq [^{18}F]FET. [^{18}F]FET was produced following the method earlier described.[51] The radiochemical purity was >98% and the specific radioactivity >18.5 GBq·μmol^{-1}. Arterial blood was continuously sampled and manual samples were taken at 5, 10, 20, 40, 60, 75 and 90 minutes. The line-of-response row-action maximum likelihood algorithm (LOR-RAMLA) algorithm as provided by the scanner manufacturer was used for reconstruction of the scans into 26 time frames (1 x 10, 8 x 5, 4 x 10, 2 x 15, 3 x 20, 2 x 30, 6 x 60 s) and 22 time frames (1 x 15, 3 x 5, 3 x 10, 4 x 60, 2 x 150, 2 x 300, 7 x 600 s), respectively, both with an isotropic voxel size of 2 mm.

The measured arterial whole blood curve was calibrated using manual arterial samples. Then, metabolite-corrected plasma curves were constructed from the whole blood curve by correcting for the plasma to whole blood ratio and labelled metabolites concentration. The parent fractions were fitted with a Hill function.[52] Concentration of both polar and non-polar metabolites was determined using solid phase extraction in combination with high performance liquid chromatography. More details on the blood measurements can be found in the supplemental material.

Image processing and segmentation

The reconstructed PET images were checked frame by frame for movement and corrected accordingly. Affected time frames were rigidly coregistered to the attenuation scan using

the generic multi-modality registration setup from Vinci (version 2.56.0, Max Planck Institute for Metabolism Research). However, if patient movement was more than 5 mm the affected time frames were reconstructed after re-aligning the attenuation scan. The newly reconstructed frames were coregistered to the original attenuation scan.

Tumour volumes were delineated on the MR images by a resident in neurosurgery with ample experience in imaging characteristics of patients with glial tumours. MR sequences were selected based on grade. Lower grade glioma was delineated using the FLAIR sequence; glioblastoma was delineated on T1G. These delineations were transferred to the dynamic PET scan after rigid coregistration—using the same registration setup—of the MR scan to the CT scan. Volume of the tumour delineations ranged from 25.2 to 100.8 cm^3. In order to minimize heterogeneity, the MR based delineations were divided into three volumes of interest (VOI) based on the 33rd and 67th percentiles of the 20-40 minutes [^{18}F]FET uptake value. These VOIs were labelled low, medium or high uptake. For the reference region, a spherical VOI with 14 mm radius was placed at the mirror location of the tumour on the contralateral side, encompassing white and grey matter tissue. In addition, two more spherical VOIs of the same volume were placed at the contralateral side, not overlapping the reference region. Together with the reference region, these form the VOIs of presumed non-tumour (healthy) brain tissue and were used to investigate the pharmacokinetics in healthy tissue.

Kinetic analysis of [^{15}O]H_2O

Parametric maps of cerebral blood flow (CBF) were constructed from the [^{15}O]H_2O PET scans and the plasma input functions using the basis function implementation of the standard single-tissue compartment model.[23]. The CBF maps were coregistered to the summed [^{18}F]FET image and the average value within each VOI was calculated. CBF was normalized to the same reference region to calculate the CBF-ratio.

Kinetic analysis of [^{18}F]FET

Time-activity curves (TACs) were generated by projecting the VOIs on the dynamic [^{18}F]FET PET images. These TACs were analysed with several pharmacokinetic plasma input models: the reversible single-tissue compartment model (1T2k_{Vb}), the irreversible two-tissue compartment model (2T3k_{Vb}) and the reversible two-tissue compartment model (2T4k_{Vb}).[53] All models included an additional fit parameter for fractional blood volume (Vb) and therefore included both the whole blood and the metabolite-corrected plasma curve as input functions. The input functions were corrected for delay using a whole brain TAC. All models were fitted using weighted non-linear regression.[35] Parameter errors were calculated as standard deviation, to estimate the reliability of the fitted kinetic parameter. To identify the optimal model, the fits of the pharmacokinetic plasma input models were evaluated visually and with the Akaike information criterion[54].

Main kinetic parameters of interest were the volume of distribution (V_T) for the reversible models, the influx rate constant (K_i) for the irreversible model and the rate constant from plasma to tissue (K_1). The relationship of these parameters with CBF was investigated using Pearson's correlation coefficient (r). A p-value less than 0.05 was considered significant. K_1 was also divided by CBF to calculate the extraction fraction. The distribution volume ratio (DVR) was calculated by normalizing the V_T using the V_T of reference region. The nondisplaceable binding potential, BP_{ND}[10], was then derived by BP_{ND} = DVR-1 and used to validate BP_{ND} obtained using reference tissue input models (next paragraph).

Performance of both the full reference tissue model (FRTM)[55,56] and the simplified reference tissue model (SRTM)[57] was investigated. The advantage of reference tissue input models is that no arterial input function is needed. Instead, a reference region is used as indirect input function, in this case the contralateral reference region. In this study, we assessed agreement between FRTM or SRTM derived BP_{ND} vs plasma input model derived DVR-1 and, similarly, R_1 vs plasma input model derived K_1-ratio (K_1 normalized to reference region) using Bland-Altman[58] analysis. The relationship of BP_{ND} and R_1 with the CBF-ratio was also investigated.

We calculated SUV for intervals 20-40 minutes (SUV^{20-40}), 40-60 minutes (SUV^{40-60}) and 60-90 minutes (SUV^{60-90}) and calculated correlation with V_T. We also calculated tumour-to-blood ratios (TBlR) to investigate whether this would be a possible surrogate of V_T. Two variants were considered: ratio to whole blood activity ($TBlR_{WB}$) and ratio to parent plasma activity ($TBlR_{PP}$). Furthermore, relationship with CBF for all the above parameters was investigated. The SUV ratio (SUVR, SUV normalized to reference region; also known as tumour-to-brain or tumour-to-normal ratio) was also calculated for these three intervals. Agreement with DVR was evaluated using Bland-Altman analysis and correlation with CBF-ratio was determined.

Results

One of the lower grade glioma patients, patient two, showed very little uptake in the tumour yet could be visually distinguished based on the SUV^{20-40}, see supplemental Figure S1. Figure 1 illustrates this and shows the SUV and SUVR over time for the high uptake VOIs. All except one tumour, from patient three, show the typical curve pattern generally associated with their grade[37]. During acquisition of the [^{15}O]H_2O PET scan of patient six there were problems with the measurement of the arterial blood activity. CBF could therefore not be quantified for this patient. Two patients had moved during the dynamic [^{18}F]FET PET scan, one had moved approximately 3 mm and the other 15 mm, both after at least 20 minutes. Both scans were corrected as described above.

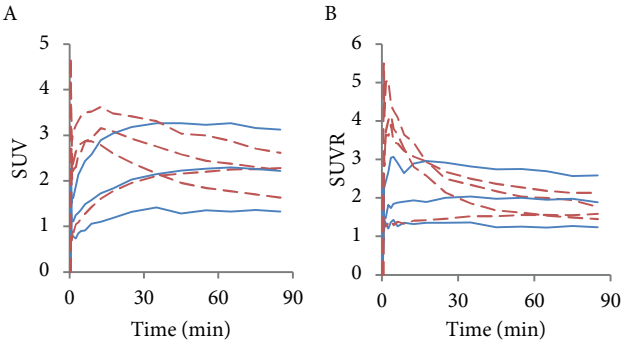


Figure 1. SUV (A) and SUVR (B) curves of the high [^{18}F]FET uptake VOI of each patient. Solid lines are lower grade gliomas, dashed lines are glioblastoma.

Figure 2 shows results from the manual blood sample measurements for the [^{18}F]FET scans. The plasma to whole blood ratio is stable at an average of 1.22 ± 0.05 (standard deviation between patients). The parent fraction of [^{18}F]FET was 79% ± 14% at time of the first manual blood sample (5 minutes post injection) and decreased slowly to 68% ± 13% at 90 minutes post injection.

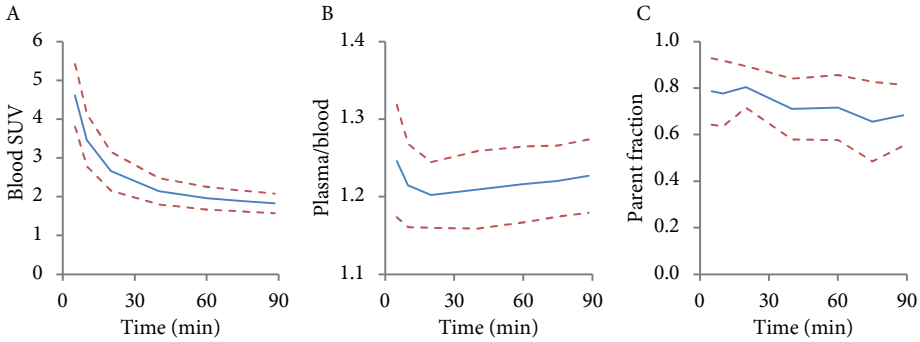


Figure 2. Data from manual blood samples, showing the whole blood activity concentration over time corrected for injected dose and patient weight (A), the ratio of activity concentration in plasma over activity concentration in whole blood (B), and the percentage parent compound in the samples (C). Solid lines are the average, dashed lines show the average ± SD over all patients.

Visual assessment of the fits showed that the irreversible model was not able to fit the tumour TACs. Figure 3 shows a typical example. The Akaike information criterion confirmed this finding and showed a preference for the 2T4k$_{Vb}$ model in 95% (20/21) of the fitted TACs; for the other 5% (1/21) the 1T2k$_{Vb}$ model was preferred. As such, the model preference seems independent of both uptake and grade as determined by histological assessment. In contralateral (healthy) brain tissue, the 2T4k$_{Vb}$ model was preferred in 52% (11/21) of the regions and the 1T2k$_{Vb}$ model in the other 48% (10/21).

Correlation for V_T in the tumour regions as derived from $2T4k_{Vb}$ and $1T2k_{Vb}$ was very high (r=0.99); however, agreement analysis showed a significant difference for estimated V_T of 0.08 (9%), as shown in the Bland Altman plot in supplemental Figure S2. The two-tissue reversible model was therefore used as reference for further analyses.

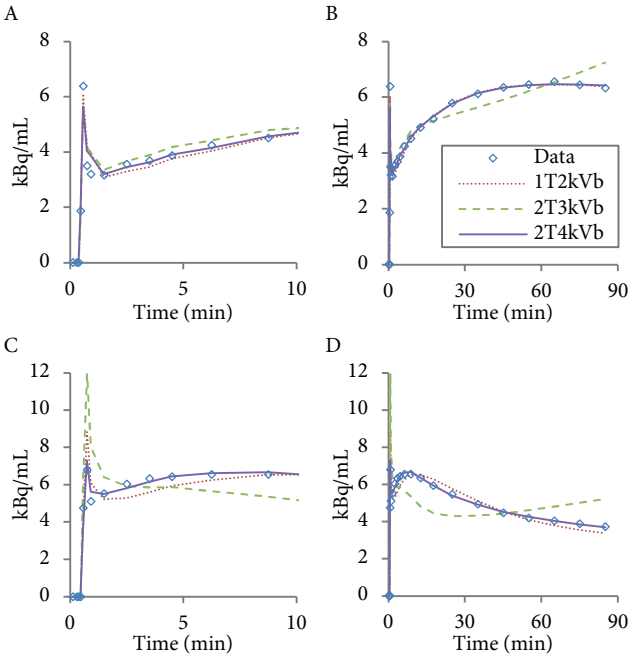


Figure 3. Typical example of a TAC with fits of the three models: $1T2k_{Vb}$ dotted line, $2T3k_{Vb}$ dashed line, $2T4k_{Vb}$ solid line. The TAC of the high uptake VOI of patient 5, lower grade glioma; the first 10 minutes of the TAC (A) and the whole 90 minutes (B). The TAC of the high uptake VOI of patient 6, glioblastoma; the first 10 minutes of the TAC (C) and the whole 90 minutes (D).

A significant but low correlation was found between V_T and CBF in the tumour regions (r=0.61, p=0.007), a scatter plot is shown in supplemental Figure S3. There was no correlation between K_1 values of [18F]FET and CBF in the tumour regions (r=-0.018, p=0.93), supplemental Figure S4. The calculated extraction fractions showed little variation in the non-tumour regions with a mean value of 0.071 and a standard deviation of 0.024. Extraction fraction in the tumour regions was higher with a mean value of 0.17 and a standard deviation of 0.13. A scatter plot of extraction fraction against CBF in both tumour and healthy regions is shown in supplemental Figure S5.

Agreement between the estimated BP_{ND} from SRTM and DVR-1 from the $2T4k_{Vb}$ is shown in Figure 4. Two outliers were identified, the low and medium uptake VOIs of patient two. The error of these BP_{ND} estimates was very high (standard deviations of 10.6 and 31.6). If we disregard these outliers the limits of agreement are -0.39 and 0.37

(range DVR-1: -0.25 to 1.5). Agreement of R_1 with K_1-ratio from $2T4k_{Vb}$ was poor with an average difference of -0.90 and limits of agreement of -3.23 and 1.44 (range K_1-ratio: 0.85 to 4.8). BP_{ND} showed significant correlation with the CBF-ratio (r=0.83, p<0.001), R_1 showed a significant but low correlation with the CBF-ratio (r=0.52, p=0.039); the scatterplots are shown in supplemental Figure S6. FRTM estimates of BP_{ND} mostly agreed with SRTM, however several additional outliers were seen with high parameter error of BP_{ND}.

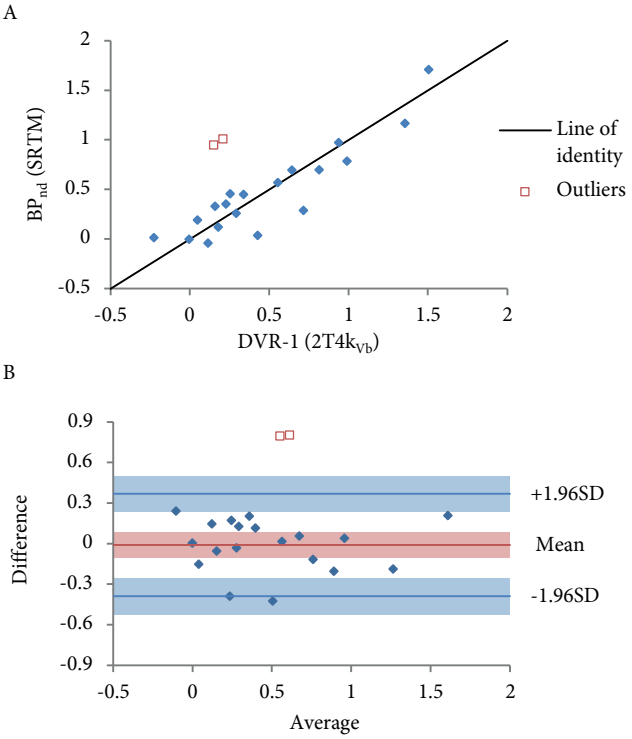


Figure 4. Agreement between BP_{ND} from SRTM and the DVR-1 from the $2T4k_{Vb}$ model. Scatter plot (A) and Bland Altman plot (B). Shaded areas are 95% confidence intervals.

Correlation between SUV^{20-40} and V_T was significant but low (r=0.62, p<0.001); the scatter plot is shown in supplemental Figure S7. Correlation with V_T was higher for later time intervals and this was also seen for $TBlR_{WB}$ and $TBlR_{PP}$ and for the correlations between SUVR and DVR. Correlation with K_1 was higher for earlier time intervals. Correlation coefficients are given in Table 1. The agreement between SUVR and DVR showed a similar pattern, where the SUVR for later time intervals show better agreement with DVR as calculated with the $2T4k_{Vb}$ model. $SUVR^{60-90}$ showed limits of agreement of -0.27 and 0.34, see Figure 5, while limits of agreement for $SUVR^{20-40}$ were -0.52 and 0.85 (range DVR: 0.75 to 2.5).

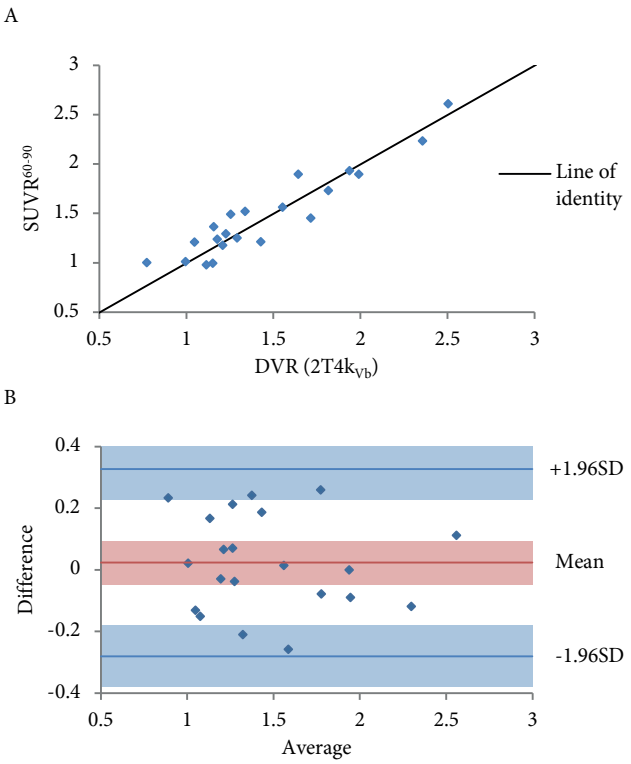


Figure 5. Agreement between $SUVR^{60-90}$ and the DVR from the $2T4k_{Vb}$ model. Scatter plot (A) and Bland Altman plot (B). Shaded areas are 95% confidence intervals.

Neither SUV nor $TBlR_{WB}$ showed significant correlation with CBF. In contrast, $TBlR_{PP}$ did show significant correlation with CBF and the correlation increased at later time intervals. For the 60-90 min interval the correlation coefficient was r=0.63, p=0.005. $TBlR_{PP}$ also showed agreement with V_T with limits of agreement of -0.17 and 0.19 (range V_T: 0.53 to 2.1) and without significant bias. SUVR showed significant correlation with the CBF-ratio, for all time intervals the correlation was higher than 0.85. It was highest for the 20-40 minute interval at 0.91, p<0.001.

Table 1: Pearson correlation r between SUV based measures and kinetic parameters from $2T4k_{Vb}$.

Interval	V_T			DVR	K_1			0.7
(min)	SUV	$TBlR_{WB}$	$TBlR_{PP}$	SUVR	SUV	$TBlR_{WB}$	$TBlR_{PP}$	
20-40	0.55	0.79	0.85	0.78	0.76	0.48	0.55	
40-60	0.70	0.84	0.94	0.88	0.69	0.41	0.45	
60-90	0.77	0.86	0.97	0.94	0.63	0.39	0.42	1.0

Discussion

The aim of this study was to derive the optimal plasma input kinetic model for dynamic [^{18}F]FET PET studies and to validate performance of simplified methods. Therefore, various metabolite corrected plasma input models were evaluated and the optimal model was determined. Next, the optimal model was used to assess the agreement of various simplified methods with the optimal model including approaches often used in [^{18}F]FET PET studies in glioma.

The optimal plasma input kinetic model was found to be the reversible two-tissue compartment model with fitted blood volume fraction. The model preference based on the Akaike criterion was clear for the tumour regions, where only 5% could be better fitted with the single-tissue compartment model. These data indicate that the model preference is independent of tumour grade or curve pattern, although there are too few data to substantiate this in this study. Healthy tissue regions were best fitted by the reversible two-tissue compartment model in half of the cases and by a single-tissue compartment model in the other half. Use of the single-tissue compartment model resulted in systematically lower estimates of V_T: in tumour regions with an average difference of -9%, in healthy regions with an average difference of -7%. Based on the fits of all target and reference tissue TACs, we concluded that the two-tissue compartment model is most suitable for the further evaluations.

Fully quantitative pharmacokinetic models require arterial plasma input functions. In this study manual arterial samples were used to correct for the labelled metabolite concentration. In an earlier report, results of metabolite measurements showed low fractions (5% at 5 minutes post injection, 13% at 120 minutes post injection), suggesting rapid excretion of labelled metabolites by the kidneys.[48] In our study the results from the manual arterial blood samples showed a larger fraction of metabolites in blood (21% at 5 minutes post injection, 32% at 90 minutes post injection). In an effort to investigate the effect of correction for the labelled metabolites, we fitted a $2T4k_{Vb}$ model with a whole plasma input function. Estimates of V_T were on average 39% lower. Yet, estimates of DVR were the same on average. Therefore, the impact of using metabolite corrected input functions versus whole plasma input function on the validation of reference region based models or simplified methods is minimal.

The results on the relationship with blood flow showed a significant correlation of V_T with CBF, but correlation was low. As V_T represents a perfusion independent estimate of tracer uptake, the observed correlation is likely due to physiological coincidence of both increased amino acid utilisation and perfusion. This makes it impossible to draw conclusions about perfusion dependence of the simplified methods. The absence of correlation between K_1 and CBF suggests that the extraction fraction is highly variable

between patients. Indeed, the variation in the calculated extraction fractions is relatively high in the tumour regions across the patients. This could be the consequence of different levels of transporter expression or may be due to differences in blood brain barrier breakdown.

Agreement analysis on the simplified reference tissue model BP_{ND} vs plasma input derived DVR-1 showed wide limits of agreement. As such, BP_{ND} seems a poor surrogate for this parameter. Agreement for R_1 vs the K_1-ratio was poor as well. The full reference tissue model showed no different results from the simplified reference tissue model, except for a few additional outliers. The poor performance of the reference tissue input model might be due to violated assumptions, making the model invalid. One of the assumptions is that both reference and target regions can be represented by a single-tissue compartment model. For half of these data, both regions are better described by a two-tissue compartment model; for the other half the target region is better described by two tissue compartments while the normal regions are best described by a single tissue compartment. The expected error from the first violation is minor, while the second violation can lead to a 10% bias.[59] Another possible source of error is non-negligible blood volume contribution. Moreover, use of reference tissue input models requires that the transport across the blood-brain barrier, represented by K_1/k_2 ratio, is equivalent between target (tumour) and reference regions. In case of gliomas, tracer uptake in the tumour can be affected by disruptions of the blood-brain barrier. Consequently, use of reference tissue input models may not be valid for dynamic [^{18}F]FET brain studies.

The $TBlR_{PP}^{60\text{-}90}$ showed good agreement with V_T. A disadvantage of the $TBlR_{WB}$ and $TBlR_{PP}$ is the requirement of blood samples and, for $TBlR_{PP}$, the need for metabolite measurements. However, their correlation results suggest that plasma clearance effects (and thus variability in input functions between subjects) seem the largest contributor to SUV variability. If we convert the correlation results to coefficients of determination we see that 94% of the variability in $TBlR_{PP}^{60\text{-}90}$ can be explained by the variability in V_T. This is encouraging for the use of SUVR, which largely corrects for variability of the input functions between patients.

For SUV, $TBlR_{WB}$ and $TBlR_{PP}$ uptake intervals later than the currently recommended 20-40 minutes show better correlation with V_T. Correlation was lowest for $SUV^{20\text{-}40}$ and highest for $TBlR_{PP}^{60\text{-}90}$. Furthermore, from the time activity curves it becomes clear that the uptake value of the tumours is still changing during the 20-40 minute interval, see Figure 1. A possible downside of early static imaging might be that variability in uptake time will lead to variability in SUV. In contrast, the SUVR curves of four patients are relatively stable during this period. Three patients, however, show a variable SUVR at the 20-40 minute interval, which becomes more constant at later times. The agreement of SUVR with DVR also improves at later time intervals. The size of this improvement is

clearly illustrated by the limits of agreement, which are more than twice as wide for the 20-40 minute interval. In terms of limits of agreement $SUVR^{60-90}$ showed a slightly better agreement with DVR than SRTM. Just like for SRTM, a possible source of error is the blood-volume fraction, especially in case of blood-brain-barrier disruption. To conclude, early time point imaging (20-40 min post injection) is usually applied and preferred in a clinical setting. A downside to static imaging is that the time activity curve pattern cannot be assessed, which has been shown to be helpful in determining the grade of glioma. Furthermore, when non-invasive quantification is required, it is recommended to use SUVR at later time points (60-90 min post injection). When studies are designed to measure changes (longitudinally or after intervention), use of $TBlR_{WB}$ and $TBlR_{PP}$ would be recommended, because of the better agreement with plasma input derived V_T and the ability of compensating for inter-subject variability of the input function. Further studies are needed to investigate whether this improved quantification also improves the clinical value.

It must be noted that the small sample size of this study requires appropriate caution in the interpretation of the results presented here. The complexity of compartmental modelling with metabolite corrected plasma input function do not enable large study cohorts, yet compartmental modelling is an important step in the evaluation of tracer kinetics and its implications for more simplified approaches. The results of this study only apply to regional analyses, i.e. based on the mean signal of a VOI. Thus, relationships between parameters within a scan cannot be adequately investigated, because the number of data points (VOIs) per scan was limited. Voxel-based methods enable such analysis, but require further evaluation due to higher noise levels in voxel-based signals.

Conclusion

In this study we derived that the two-tissue reversible plasma input model with fitted blood volume fraction is the optimal plasma input model to describe the kinetics of [^{18}F]FET in glioma patients. Furthermore, use of reference tissue input models and simplified methods, such as SUV and SUVR, was validated. BP_{ND} results obtained with reference tissue input models did not correspond well with plasma input derived DVRs, possibly due to violation of the reference tissue model assumptions. SUVR showed slightly better agreement with DVR than SRTM derived BP_{ND}. SUV only moderately correlated with V_T with the best correspondence at later uptake time intervals (60-90 min post injection). The results of the study suggest that later time point imaging (60-90 min post injection) outperforms currently recommended uptake time (20-40 min post injection) in terms of quantitative value, i.e. correlation with V_T and DVR.

Supplementary material

Supplementary material for this paper is available at:

doi.org/10.1186/s13550-018-0418-0

4

Quantitative parametric maps of O-(2-[^{18}F]fluoroethyl)-L-tyrosine kinetics in diffuse glioma

Thomas Koopman, Niels Verburg, Petra J.W. Pouwels, Pieter Wesseling,
Otto S. Hoekstra, Philip C. De Witt Hamer, Adriaan A. Lammertsma, Maqsood Yaqub,
Ronald Boellaard

Published 24 May 2019
Journal of Cerebral Blood Flow & Metabolism 2020; 40(4): 895–903
DOI: 10.1177/0271678X19851878

Abstract

Quantitative parametric images of O-(2-[^{18}F]fluoroethyl)-L-tyrosine kinetics in diffuse gliomas could be used to improve glioma grading, tumour delineation or the assessment of the uptake distribution of this positron emission tomography tracer. In this study, several parametric images and tumour-to-normal maps were compared in terms of accuracy of region averages (when compared to results from nonlinear regression of a reversible two-tissue compartment plasma input model) and image noise using 90 min of dynamic scan data acquired in seven patients with diffuse glioma. We included plasma input methods (the basis function implementation of the single-tissue compartment model, spectral analysis and Logan graphical analysis) and reference tissue methods (basis function implementations of the simplified reference tissue model, variations of the multilinear reference tissue model and non-invasive Logan graphical analysis) as well as tumour-to-normal ratio maps at three intervals. (Non-invasive) Logan graphical analysis provided volume of distribution maps and distribution volume ratio maps with the lowest level of noise, while the basis function implementations provided the best accuracy. Tumour-to-normal ratio maps provided better results if later interval times were used, i.e. 60–90 min instead of 20–40 min, leading to lower bias (2.9% vs. 10.8%, respectively) and less noise (12.8% vs. 14.4%).

Introduction

Diffuse gliomas exhibit increased uptake and retention of O-(2-[^{18}F]fluoroethyl)-L-tyrosine ([^{18}F]FET), an amino acid tracer that can be visualised with positron emission tomography (PET). In a previous study the optimal plasma input model for describing [^{18}F]FET kinetics was identified.[60] However, VOIs have to be defined beforehand and tracer uptake distributions cannot be assessed. The currently recommended[41] [^{18}F]FET PET standardized uptake value (SUV) image at 20–40 min shows good contrast between lesions and healthy tissue. Interpatient differences are reduced by normalizing tumour uptake to that in a contralateral healthy region. Indeed, a tumour-to-normal ratio at 20–40 min is widely used for tumour delineation.[39] At the same time, many other studies have used a dynamic scanning protocol, mostly for discriminating different tumour types based on uptake patterns.[39] Several methods exist for "catching" tracer kinetics into parametric images. In theory, parametric images are more accurate than SUV images or tumour-to-normal maps, and may be better for glioma grading or delineation. Yet Logan graphical analysis has been the only parametric method for quantifying [^{18}F]FET uptake so far.[43,44,46]

The aim of this study was to determine the accuracy of parametric images and tumour-to-normal maps for quantifying [^{18}F]FET uptake. Results obtained using the previously identified plasma input model were used as reference. In addition, image noise characteristics of the maps were taken into account.

Methods

Subjects

Data were derived from a study that has been reported previously.[49,60] In short, the study population consisted of seven patients with a diffuse glioma (age range, 22 – 69 y; four glioma WHO[50] grade IV and three grade II). This study has been performed in accordance with the Declaration of Helsinki, approved by the Medical Ethics Committee of the VU University Medical Center and registered in the Netherlands National Trial Register (www.trialregister.nl, unique identifier NTR5354, registration date 4 August 2015). Written, informed consent was obtained from all subjects prior to inclusion.

Scanning protocol

Magnetic resonance (MR) sequences were acquired on an Achieva whole body 3.0T MR scanner (Philips Healthcare, Best, the Netherlands), equipped with a standard head coil. Each patient was scanned using a sagittal 3D fluid-attenuated inversion recovery (FLAIR) sequence (repetition time(TR)/echo time(TE)/inversion time(TI) 4800/279/1650 ms, acquired voxel size 1.12×1.12×1.12 mm^3, reconstructed voxel size 1.04×1.04×0.56 mm^3), and a sagittal 3D T1-weighted gadolinium-enhanced (T1G) sequence (TR/TE/TI/flip

angle 7/3/950 ms/12°, acquired voxel size 0.98×0.98×1.0 mm^3, reconstructed voxel size 0.87×0.87×1.0 mm^3). A dynamic PET scan was acquired on either a Gemini TF-64 or an Ingenuity TF PET/computed tomography (CT) scanner (Philips Healthcare, Cleveland, Ohio, USA). Each scan started with a 1 min low dose CT scan for attenuation correction purposes. Next, a 90 min PET scan was acquired after administration of 200 MBq [^{18}F]FET. The tracer was injected using a venous line, while an arterial line in the opposite arm was used for continuous sampling using an on-line blood sampler (Comecer Netherlands, Joure, the Netherlands). In addition, manual arterial samples were collected at 5, 10, 20, 40, 60, 75 and 90 min post injection of [^{18}F]FET. Using the LOR-RAMLA algorithm, as provided by the manufacturer, scans were reconstructed into 22 frames (1 x 15, 3 x 5, 3 x 10, 4 x 60, 2 x 150, 2 x 300, 7 x 600 s), with an isotropic voxel size of 2 mm. Reconstructions included all usual corrections, i.e. normalization, decay, dead time, attenuation, randoms and scatter correction. The manual blood samples were used to calibrate the on-line blood curve and to correct it for plasma-to-whole blood concentration ratios and labelled metabolite fractions, thereby generating a metabolite corrected, arterial plasma input function.

Data analysis

Glioblastomas were delineated on T1-weighted gadolinium-enhanced MRI images (T1G) and lower grade gliomas on FLAIR MRI images. As described elsewhere[60], tumour segmentations were divided into three equal sized volumes of interest (VOI) using the 33rd and 67th percentiles of the activity concentrations of [^{18}F]FET at 20 to 40 min. A spherical reference region with a radius of 14 mm was placed in the middle of the contralateral homologous brain region.[60] Time activity curves were extracted from these regions, which were fitted to the reversible two-tissue compartment plasma input model with additional blood volume fraction using nonlinear regression. In earlier work[60] we found that reversible models were always preferred over the irreversible model in both tumour and reference regions and that the reversible two-tissue compartment model was preferred over the reversible single-tissue compartment model in most cases. The total volume of distribution (V_T) was used as outcome measure. The distribution volume ratio (DVR) was calculated by normalizing the V_T to the V_T of the reference region. Results for both parameters served as reference standard for the agreement analysis.

Parametric V_T images were created using a basis function implementation of the reversible single-tissue compartment model (BFM)[23], plasma input-based Logan graphical analysis (Logan)[61] and Spectral Analysis (SA)[62]. Using the contralateral reference region, reference input-based Logan analysis (RLogan)[63] was used to create a DVR map. Non-displaceable binding potential (BP_{ND}) maps were generated with basis function implementations of the simplified reference tissue model (receptor parametric mapping (RPM) and $SRTM_2$)[64,65] and using several variations of the multi-linear

reference tissue model ($MRTM_O$, MRTM, $MRTM_2$, $MRTM_3$ and $MRTM_4$)[66-68]. $MRTM_2$, $MRTM_3$, $MRTM_4$ and $SRTM_2$ are all methods using a fixed k_2' (the clearance rate of the reference tissue) based on the median value from a first run. They are based on MRTM, $MRTM_O$, $MRTM_O$ and RPM, respectively. In $MRTM_4$ uses a different model in the first run where the fixed k_2' is based on MRTM. The BP_{ND} maps were converted to DVR maps using DVR = BP_{ND} + 1. Each method was applied using only the first 60 min of the acquired data to investigate the possibility of shortening scanning times, indicated in the results by 60 in superscript. Finally, standardized uptake value ratio (SUVr, also known as tumour-to-normal ratio) maps were created for three intervals: 20–40, 40–60 and 60–90 min with intervals indicated by superscripts. SUVr was calculated by normalizing to the average uptake value in the reference region.

All maps were visually inspected for artefacts. After extracting average regional values from the parametric images, Bland-Altman analysis[58] was used to determine the accuracy, i.e. the agreement with the reference, described above. Relative differences were calculated by dividing the difference by the reference. Results were summarized by both mean and standard deviation of these relative differences.

The 3D T1G sequence was used for segmenting grey matter with SPM12.[69] The grey matter probability map of the whole brain, including cerebellum, was converted to a binary mask using an intensity cut-off of 0.9. The tumour VOI was excluded from the grey matter mask to obtain a mask with only normal appearing brain tissue. This region was used to estimate image noise in the parametric maps by means of the coefficient of variation (COV, the standard deviation divided by the mean) of the voxel values within the region. These image noise estimates were used to rank the methods with respect to image quality.

Logan, RLogan and MRTM variations are linearization methods and require a start time (t*) representing the time beyond which the linear fit can be applied. The other methods are basis function implementations and require a range and number of basis functions. Settings were optimized for each method in preliminary analysis, selecting the settings producing the best accuracy. The settings used for each method are listed in Table 1.

Table 1: Parametric methods and settings.

Method	Parameter	Start time (min)	Basis function range (min^{-1})	Number of basis functions
BFM	V_T		0.01 – 0.5	50
SA	V_T		0.01 – 4	50
Logan	V_T	10		
RLogan	DVR	30		
RPM	BP_{ND}+1 = DVR		0.01 – 4	50
$SRTM_2$	BP_{ND}+1 = DVR		0.01 – 0.1	50
$MRTM_O$	DVR	30		
MRTM	DVR	10		
$MRTM_2$	DVR	10		
$MRTM_3$	DVR	30		
$MRTM_4$	DVR	50		

Results

Typical parametric maps of all methods are shown in Figure 1, using the three intervals for the SUVr images and 90 min of data for the other methods. Upon visual inspection, it became evident the BFM maps contained an artefact: boundaries appeared due to sudden steps in V_T values, forming patches throughout the brain. We will refer to this as patchiness. The RPM maps showed a similar effect and the $SRTM_2$ maps showed some patchiness mostly in white matter. These patches can sometimes be situated near or inside the tumour region. MRTM maps suffered from 'dot artefacts'—isolated voxels showing very high or very low values—resulting in high estimated image noise. The SUVr maps showed a decreasing contrast between tumour and normal brain for later intervals for most glioblastoma patients. The glioblastoma patient where this effect was strongest is shown in Figure 1. All results are summarized in Table 2.

Table 2: Results.

Method		Relative accuracy		Noise	Tumour-to-normal ratio	
		SD (%)	Mean (%)	COV (%)	Mean	SD
BFM	90	5.7	-4.9	15.9	1.48	0.45
	60	7.9	-9.2	21.1	1.47	0.46
Logan	90	7.5	-12.1	13.2	1.51	0.46
	60	10.3	-20.7	16.2	1.52	0.50
SA	90	9.4	19.4	14.2	1.45	0.43
	60	12.3	24.8	16.1	1.45	0.45
RLogan	90	18.3	7.3	12.1	1.54	0.46
	60	21.8	9.3	13.7	1.57	0.50
RPM	90	7.8	0.9	20.8	1.46	0.40
	60	8.2	-0.5	26.5	1.44	0.44
$SRTM_2$	90	12.0	6.7	12.7	1.54	0.44
	60	15.2	9.3	14.4	1.58	0.47
$MRTM_O$	90	15.6	4.4	12.4	1.50	0.45
	60	19.0	6.0	54.2	1.53	0.48
MRTM	90	11.6	4.3	85.7	1.51	0.46
	60	19.0	6.4	74.6	1.53	0.49
$MRTM_2$	90	139.9	67.5	229.8	2.37	1.85
	60	44.0	2.8	146.5	1.44	0.89
$MRTM_3$	90	16.0	4.8	12.3	1.51	0.44
	60	21.1	3.5	25.7	1.49	0.48
$MRTM_4$	90	36.2	3.1	24.1	1.46	0.54
	60	440.7	433.6	34.4	6.76	5.70
SUVr	60–90	12.4	2.9	12.8	1.48	0.43
	40–60	17.9	6.0	13.5	1.53	0.47
	20–40	27.1	10.8	14.4	1.59	0.54

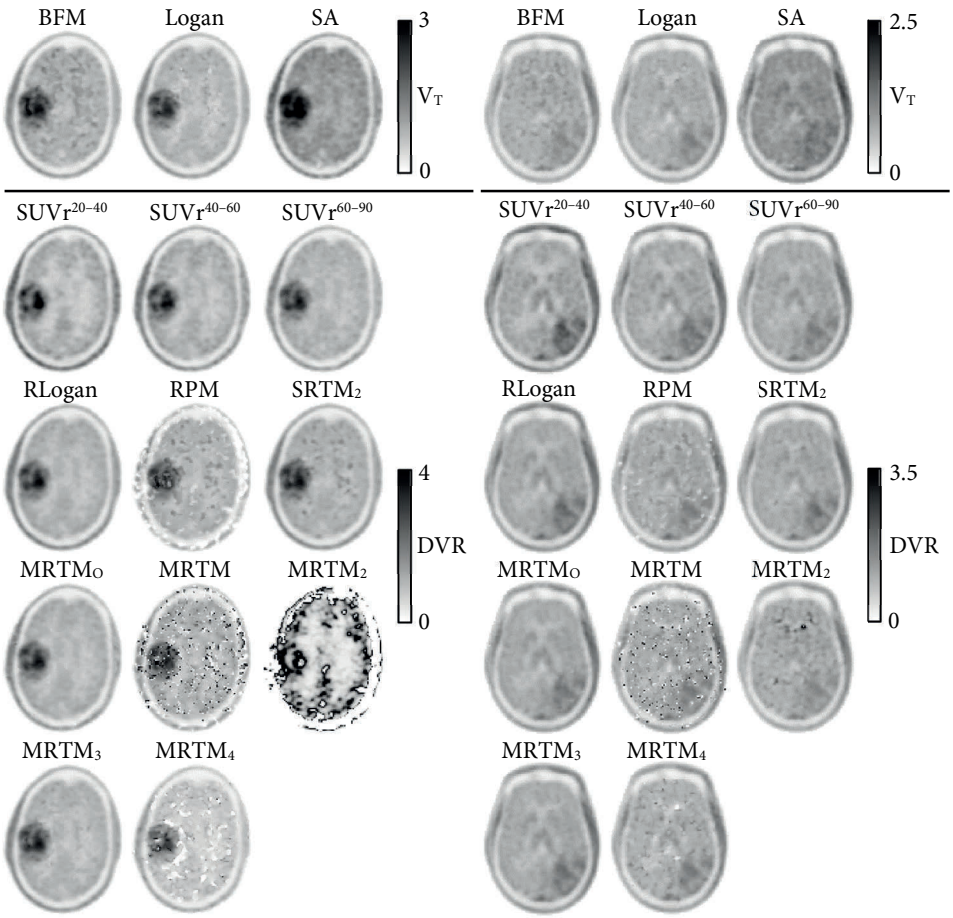


Figure 1: Typical parametric and SUVr (tumour-to-normal) maps. Left is a patient with an oligodendrocytoma, right is a glioblastoma patient.

The results on accuracy for V_T are shown in Figure 2A, which shows the relative agreement with the reference standard. The highest accuracy when using 90 min of data was observed for BFM with a standard deviation of 5.7% and a small average underestimation of −4.9%. Logan shows a larger standard deviation, 7.5%, and a larger and consistent underestimation, −12%. SA had the lowest accuracy with a standard deviation of 9.4% and an average overestimation of 19%. The measured image noise, i.e. COV of every V_T map is visualized in Figure 2B. In terms of image noise, BFM was found to be the worst of the three, with an average COV of 15.9%. This is in line with visual inspection, as described above. Logan showed the lowest level of image noise with an average COV of 13.2%. SA showed an average COV of 14.2%. When using 60 min of data, the accuracy became worse for all methods, but their ranking remained the same, and the average image noise COV rises to more than 16% for all methods.

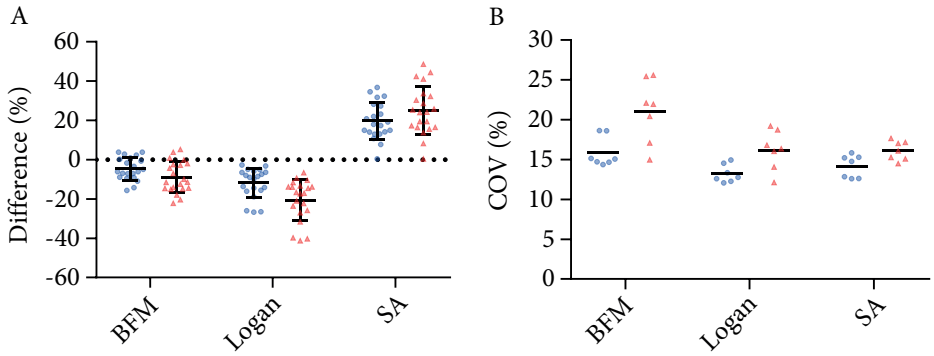


Figure 2: Circles represent the full 90 min dataset, triangles the first 60 min. (A) Accuracy; bars represent mean and standard deviation. Please note that the data points are from three regions inside the tumour for each subject, thus data can be correlated. (B) Noise estimated in V_T maps; bars represent mean.

Results on accuracy and the measured COVs for DVR maps are shown in Figure 3. Using 90 min of data, RLogan provided the best maps in terms of image noise with a COV of 12.1%. In terms of agreement with results from the reference standard, however, it showed a wide range of differences with a standard deviation of 18.3% and an average overestimation of 7.3%. RPM provided the best accuracy with a standard deviation of 7.8% and a mean overestimation of 0.9%, but showed poor performance in terms of image noise. Observed image noise was less for $SRTM_2$ maps. However, the accuracy of $SRTM_2$ maps was poorer with a standard deviation of 12.0% and an average overestimation of 6.7%.

When using 90 min of data, $MRTM_O$ showed little noise, yet the standard deviation of the differences was higher than for RPM, $SRTM_2$, MRTM and $SUVr^{60-90}$. $MRTM_3$, where the k_2' in $MRTM_O$ is fixed, was comparable to $MRTM_O$ in terms of noise, but poorer in accuracy. MRTM performed better than $MRTM_O$ in terms of accuracy, but showed poor performance in terms of noise, agreeing with visual inspection described above. Both $MRTM_2$ and $MRTM_4$ showed inconsistent results: for most patients the maps showed large offsets, negative or positive, resulting in high standard deviations of differences (36.3% to 440%). Note that $MRTM_2$ and $MRTM_4$ were not included in Figure 3A to more clearly show the differences between the other methods. For the same reason, RPM^{60}, $MRTM_O^{60}$, MRTM, $MRTM_2$, $MRTM_3^{60}$ and $MRTM_4$ were not included in Figure 3B. These data can be found in the supplemental material.

Amongst the SUVr maps, the 60–90 min interval was the best in terms of accuracy as well as image noise. $SUVr^{60-90}$ showed accuracy comparable with MRTM and $SRTM_2$ and in terms of image noise it was comparable to $SRTM_2$, although $SRTM_2$ shows some abnormal patches mostly in white matter, which was not included in noise estimation.

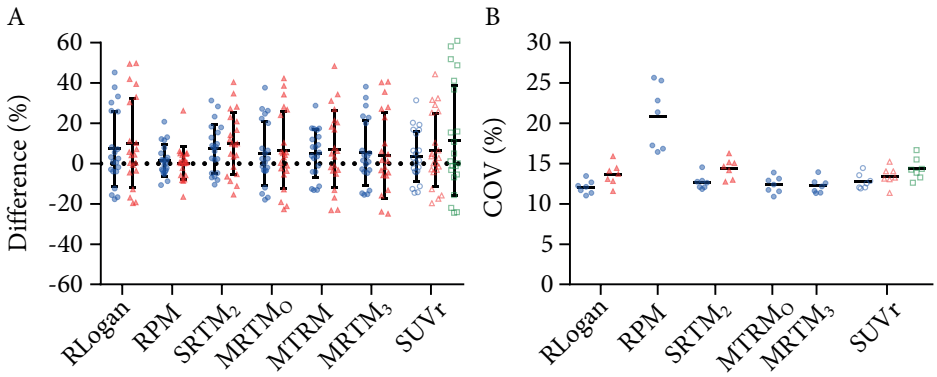


Figure 3: Filled circles represent the full 90 min dataset, filled triangles the first 60 min, open circles the time interval of 60–90 min, open triangles 40–60 min, open squares 20–40 min. (A) Accuracy; bars represent mean and standard deviation. Please note that the data points are from three regions inside the tumour for each subject, thus data can be correlated. $MRTM_2$ and $MRTM_4$ were excluded from this figure. (B) Noise estimated in the DVR or BP_{ND}+1 maps; bars represent mean. RPM^{60}, $MRTM_O^{60}$, MRTM, $MRTM_2$, $MRTM_3^{60}$ and $MRTM_4$ were excluded from this figure to more clearly show differences between the remaining methods.

Discussion

An important finding of this study is that, in general, less noise in the images (COV of voxel values) is associated with poorer accuracy at region level. In other words, the optimal parametric method depends on the specific application where it is used for. Some methods, however, showed better performance than others and can be recommended for further research. For estimation of V_T, BFM showed the best accuracy, while in terms of noise, Logan plots show the best performance. For estimation of DVR, $MRTM_O$, $MRTM_3$ and RLogan plots showed good results in terms of image noise, but performed relatively poor in terms of accuracy, i.e. these methods showed some larger variance in differences with the reference. RPM showed the best accuracy, followed by MRTM, but both methods showed relatively high image noise levels. $SRTM_2$ and $SUVr^{60-90}$ showed comparable results both in terms of estimated image noise and accuracy.

Patchiness in BFM V_T maps can be seen especially in areas with low tracer uptake. The rate constants are difficult to determine in these areas because k_2 reaches the lower limit. Although lowering the limit results in fewer and smaller patches, it also results in more prominent patches because contrast with surrounding tissue becomes higher. Because some of the patches are inside or near the tumour region, BFM is ill-suited for delineation purposes. Logan V_T maps show an expected systemic underestimation mainly caused by noise, as previously reported for other tracers.[70] SA does not show patches, but in terms of noise and accuracy of V_T, it is inferior to the Logan maps in this study. Therefore, Logan is the most precise method for measuring V_T at the voxel level. This conclusion also holds if shorter (60 min) dynamic scans are used.

The basis function implementations RPM and $SRTM_2$ showed patchiness similar to BFM. Possibly, the patches arise from fit instability due to low tracer uptake or from the violated assumption of single tissue compartment models in both target and reference regions. $SRTM_2$ is less affected than RPM, which indicates that the effect in RPM is partly caused by an unstable k_2' estimation. Investigating estimated k_2 values showed that for most voxels RPM chooses either the upper or the lower limit, thus k_2' compensation is needed to ensure good fits. When k_2' is fixed to a global brain estimate in $SRTM_2$, most patches disappear, although some patches persist in areas with relatively low rate constants. Again, these patches can be near or in the tumour region. Therefore, the use of both RPM and $SRTM_2$ for delineation is questionable while they perform well for assessing tracer uptake within (regions of) the tumour.

The main purpose of MRTM is not the parametric map itself, but providing a reliable k_2' estimate. The noise in the MRTM maps was expected: as described in the original paper the variability of the method increases compared to $MRTM_O$, but a better accuracy is achieved, which is in line with the results here. Although a better accuracy for DVR is achieved, the k_2' estimation is unstable, causing large differences in the $MRTM_2$ and $MRTM_4$ maps. Ichise et al. recommend to use regional TACs where $k_2' \neq k_2$ for MRTM's k_2' estimation because the method is not only sensitive to noise, but also becomes unstable when the clearance rates become identical.[71] We fixed k_2' for both $MRTM_2$ and $MRTM_4$ using a threshold on MRTM $BP_{ND}>0$—which has worked well in the past[68,72]—but, given the sensitivity to noise, it might be better to use region based signal(s) for the k_2' estimation. The data shows, however, that clearance rates using a single tissue compartment model can be very similar in both tumour and reference region, especially in the lower grade diffuse gliomas. Thus, finding a suitable reference region is problematic. Although some optimization is possible, use of $MRTM_2$ or $MRTM_4$ is not promising for FET in glioma.

RLogan plots showed maps with the lowest noise levels, but also with relatively low accuracy. $MRTM_O$ showed better accuracy, and only a small increase in noise. $MRTM_3$ is comparable to $MRTM_O$. $SUVr^{60-90}$ shows the best accuracy among the remaining methods and is not much poorer in terms of noise. When only 60 min data is available, $SUVr^{40-60}$ is the best method in terms of noise and only RPM^{60} and $SRTM_2^{60}$ show better accuracy. If 60 min data is used, all MRTM variations show more noise than the other methods.

SUVr is the easiest method to implement and it is used in most studies since it is the currently recommended method, although with an earlier tracer uptake interval. The present results indicate, however, that a later interval shows better agreement with DVR derived using a two-tissue compartment model with blood volume fraction. SUVr also showed less noise at later intervals. From visual inspection of the images, it is clear that the contrast between grey and white matter also decreases. Although we have tried to

minimize partial volume effects by using a relatively high cut-off value for the grey matter mask, the higher contrast for earlier intervals might (partly) explain the higher image noise estimates. Although some methods show better results in terms of accuracy or image noise, the SUVr maps show relatively good results in both.

Inherent to SUVr images at later intervals is a decreased tumour-to-normal ratio in high-grade gliomas; these tumours typically show decreasing activity concentrations after an early peak, while the activity concentration in the reference region is constant after 30 min, approximately. This decreasing contrast over time can make it harder to see and delineate the tumour. In case of threshold-based delineation, the decrease can pose a problem when the ratio approaches noise levels in the image. An example of this is found in Figure 1, where the extent of the tumour is increasingly difficult to determine in the later SUVr images compared to the SUVr$^{20\text{-}40}$ image. Although SUVr images at a later interval provided better quantitative performance, their application will prove problematic in some glioblastoma patients. Future research should investigate whether changing the time interval of SUVr images shows improvement in clinical applications, such as improved sensitivity or specificity in distinguishing between tumour and normal tissue, and whether or not this outweighs the problem of (too) low contrast in some patients.

Conclusion

In this study, we evaluated the performance of several parametric methods for the analysis of dynamic brain ^{18}F-FET PET studies. It was found that the optimal method depends on the intended application. If a region-based approach is used, BFM and RPM are recommended for most accurate estimation of V_T and DVR, respectively, despite patchy artefacts in the images. If quantitative maps are required for accurate estimates on voxel level, e.g. for assessing the location of tumour boundaries or assessing tracer uptake distribution, Logan graphical analysis and SUVr$^{60\text{–}90}$ (tumour-to-normal maps at interval 60–90 min) are the most suitable methods for deriving V_T and DVR, respectively. For tumour-to-normal maps, longer or, in case of static imaging, later scans provided better quantitative performance. Assessment of the clinical relevance of these findings is needed. Because of the good performance of SUVr, future studies could focus on the clinical evaluation of SUVr, obtained at several tracer uptake intervals.

Supplementary material

Supplementary material for this paper is available at:

doi.org/10.1177/0271678X19851878

5

Repeatability of arterial input functions and kinetic parameters in muscle obtained by dynamic contrast enhanced MR imaging of the head and neck

Thomas Koopman, Roland M. Martens, Cristina Lavini, Maqsood Yaqub,
Jonas A. Castelijns, Ronald Boellaard, J. Tim Marcus

Published 21 January 2020
Magnetic Resonance Imaging 2020; 68: 1–8
DOI: 10.1016/j.mri.2020.01.010

Abstract

Background: Quantification of pharmacokinetic parameters in dynamic contrast enhanced (DCE) MRI is heavily dependent on the arterial input function (AIF). In the present patient study on advanced stage head and neck squamous cell carcinoma (HNSCC) we have acquired DCE-MR images before and during chemo radiotherapy. We determined the repeatability of image-derived AIFs and of the obtained kinetic parameters in muscle and compared the repeatability of muscle kinetic parameters obtained with image-derived AIF's versus a population-based AIF.

Materials and methods: We compared image-derived AIFs obtained from the internal carotid, external carotid and vertebral arteries. Pharmacokinetic parameters (v_e, K^{trans}, k_{ep}) in muscle—located outside the radiation area—were obtained using the Tofts model with the image-derived AIFs and a population averaged AIF. Parameter values and repeatability were compared. Repeatability was calculated with the pre- and post-treatment data with the assumption of no DCE-MRI measurable biological changes between the scans.

Results: Several parameters describing magnitude and shape of the image-derived AIFs from the different arteries in the head and neck were significantly different. Use of image-derived AIFs led to higher pharmacokinetic parameters compared to use of a population averaged AIF. Median muscle pharmacokinetic parameters values obtained with AIFs in external carotids, internal carotids, vertebral arteries and with a population averaged AIF were respectively: v_e (0.65, 0.74, 0.58, 0.32), K^{trans} (0.30, 0.21, 0.13, 0.06), k_{ep} (0.41, 0.32, 0.24, 0.18). Repeatability of pharmacokinetic parameters was highest when a population averaged AIF was used; however, this repeatability was not significantly different from image-derived AIFs.

Conclusion: Image-derived AIFs in the neck region showed significant variations in the AIFs obtained from different arteries, and did not improve repeatability of the resulting pharmacokinetic parameters compared with the use of a population averaged AIF. Therefore, use of a population averaged AIF seems to be preferable for pharmacokinetic analysis using DCE-MRI in the head and neck area.

Introduction

Dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) can be used to estimate tissue perfusion and micro vessel permeability. The rate constants estimated using Tofts pharmacokinetic analysis of DCE-MR images (i.e. K^{trans} and k_{ep}) [73] and their ratio (v_e) reflect physiological parameters such as perfusion, permeability and cellular density, and can therefore be used to quantitatively assess these tissue properties. As reviewed by Bernstein et al., quantitative DCE-MRI biomarkers are potential predictors of prognosis and treatment response in head and neck squamous cell carcinoma (HNSCC) [74]. The validation of these biomarkers is, however, still ongoing, both in the head-and-neck region as well as in other body parts.[75]

One essential requirement for the Tofts pharmacokinetic analysis is the knowledge of the arterial input function (AIF). Since the obtained rate constants are heavily dependent on the AIF[76-80], an accurate and precise measurement is necessary for their absolute and reliable quantification. Alternatively, a simplified approach, such as a population averaged AIF can be used. However, (large) variabilities in cardiac output—between patients and within patients over time—are no longer taken into account with this approach. If this variability in cardiac output can be accounted for by precise measurement of the AIF, the accuracy and repeatability of the kinetic parameters should be superior over use of a population averaged AIF. Some authors have shown that a population averaged AIF can result in better repeatability[81,82], whereas others report the opposite.[83,84] It is possible that repeatability depends on the imaged body part and imaging sequence parameters, but also on the choice of the artery for AIF measurement.

As recently indicated by the quantitative imaging biomarkers alliance (QIBA)[75], the literature lacks studies on repeatability of quantitative (pharmacokinetic model derived) DCE-MRI parameters. This is especially true in the head and neck region. The repeatability of the AIF used as input for the model is also only sporadically reported[85,86]. We therefore sought to investigate both the dependence of the AIF repeatability on the choice of the artery, as well as the dependence of repeatability of the pharmacokinetic parameters on the chosen AIF.

In the present patient study on advanced stage head and neck squamous cell carcinoma (HNSCC) we have acquired DCE-MR images before and during chemo radiotherapy. Because the second MRI examination occurs during treatment, while the first occurs before, we are not able to report on the repeatability of kinetic parameters in HNSCC tumor tissue. Instead, we chose a neck muscle (left semispinalis capitis muscle) outside the radiation zone assuming that this muscle would be unaffected by the treatment and its pharmacokinetic parameters would remain unchanged between the first and the second examination.

We assessed the repeatability of the parameters describing the image-derived AIFs, measured in the internal carotids, external carotids and vertebral arteries, both on the left and right side. At the same time we assessed the repeatability of the pharmacokinetic parameters in the muscle using image-derived AIFs obtained from the internal carotids, external carotids and vertebral arteries, respectively, and compared it to that obtained using a population averaged AIF.

Methods

Subjects

The study population consisted of 29 patients with advanced stage squamous cell carcinoma who successfully underwent two MRI examinations in an ongoing prospective study. This prospective, single-center study was approved by the Medical Ethics Committee of the university and has been performed in accordance with the Declaration of Helsinki. Informed consent was acquired from all patients after full explanation of the procedures. Previously untreated patients with histologically proven HNSCC, planned for curative (chemo) radiotherapy were consecutively included from 2013 until 2018. Treatment consisted of radiotherapy (70 Gy in 35 fractions in a seven week period) with or without concomitant chemotherapy (cisplatin or cetuximab). Exclusion criteria were: nasopharyngeal tumors, age <18 and inadequate image quality.

Baseline imaging was performed before treatment. Two weeks after start of treatment a second imaging session was performed with exactly the same MRI protocol on the same MRI scanner. The basic assumption in this study is that between both MRI examinations, there was no systematic effect of treatment on the AIF and on the contrast enhancement properties of muscle tissue outside the radiation zone. The validity of this assumption might seem questionable, because weight and muscle mass loss is a general effect of the treatment and of the disease itself.[87] However, bodyweight is a factor that is accounted for in the administration dose of the contrast agent. Moreover, given the relatively short amount of time between scans, measurable changes in healthy muscle tissue were not expected. Thus the comparison between baseline and during treatment imaging gives the opportunity to assess repeatability of the AIF and the DCE parameters in muscle.

Imaging protocol

The DCE MRI acquisition was preceded by a variable flip angle (VFA) measurement for T_1 map estimation and followed by a B1 mapping acquisition. Sequences were acquired on a 3.0T Ingenuity TF PET/MR-scanner (Philips Healthcare, Best, the Netherlands) equipped with a 16-channel neuro-vascular coil. Dotarem® (Guerbet, Roissy, France) was used as a gadolinium-based contrast agent. The specifications of the DCE sequence were: 3D T1-FFE (T1-weighted 3D spoiled gradient echo sequence), TR 3.1 ms, TE 1.48 ms, flip angle 12°, acquired matrix size 184×169×17, acquired voxel size 1.30×1.30×4.40 mm^3,

reconstructed matrix size 320×320×17, reconstructed voxel size 0.75×0.75×4.40 mm^3, 75 time frames, frame duration 4.1 s. A SENSE factor of two was applied in the anterior-posterior direction. After at least four time frames, the contrast agent (0.2 ml/kg, concentration 0.5 mmol/ml) was injected at a speed of 3 ml/s using a Medrad® Spectris Solaris® power injector. A flush of 15 ml saline water was injected at 3 ml/s following the contrast bolus. The VFA measurement was acquired prior to contrast injection with settings nearly identical to the DCE protocol and five flip angles (2°, 5°, 10°, 15° and 20°). B1 mapping was performed using the method described by Yarnykh[88] (3D T1-FFE, TR_1 20 ms, TR_2 100 ms, TE 3.2 ms, flip angle 50°, acquired matrix size 176×177×17, acquired voxel size 1.31×1.30×4.40 mm^3, reconstructed matrix size 320×320×17, reconstructed voxel size 0.72×0.72×4.40 mm^3). The B1-map was resliced to the voxel size of the DCE image using linear interpolation and used for flip angle correction of both the DCE and VFA image. The VFA image was converted to a T_1 map using a linear least squares fit of Equation 1 as described by Gupta.[89]

The signal intensity equation for a spoiled gradient echo sequence, assuming steady state and ignoring T_2^* effects[90], is given by

$$S = M_0 \frac{\left(1-e^{-TR/T_1}\right)\sin\theta}{1-e^{-TR/T_1}\cos\theta}, \tag{1}$$

where S is the signal intensity, M_0 is the thermal equilibrium magnetization and θ is the flip angle. By assuming a fast exchange regime (i.e. $T_1^{-1} = T_{10}^{-1} + r_1 C$) the contrast concentration dependent signal intensity expression (Equation 1) becomes

$$S(t) = M_0 \frac{\left(1-e^{-TR\left(T_{10}^{-1}+r_1 C(t)\right)}\right)\sin\theta}{1-e^{-TR\left(T_{10}^{-1}+r_1 C(t)\right)}\cos\theta}, \tag{2}$$

where T_{10} is the pre-contrast longitudinal relaxation time, r_1 is the relaxivity of the contrast medium and C is the contrast concentration. Defining the pre-contrast signal intensity as S_0, signal enhancement can be defined as

$$\frac{S(t)-S_0}{S_0} = \frac{\left(e^{-TR\left(T_{10}^{-1}+r_1 C(t)\right)}-1\right)\left(e^{-TR/T_{10}}\cos\theta-1\right)}{\left(e^{-TR/T_{10}}-1\right)\left(e^{-TR\left(T_{10}^{-1}+r_1 C(t)\right)}\cos\theta-1\right)} - 1, \tag{3}$$

such that the contrast concentration can be expressed as

$$C(t) = \frac{1}{TR \cdot r_1} \ln\left(\frac{1-\frac{S(t)}{S_0}\cos\theta\left(\frac{1-e^{-TR/T_{10}}}{1-e^{-TR/T_{10}}\cos\theta}\right)}{1-\frac{S(t)}{S_0}\left(\frac{1-e^{-TR/T_{10}}}{1-e^{-TR/T_{10}}\cos\theta}\right)}\right) - \frac{1}{T_{10} \cdot r_1}. \tag{4}$$

Equation 4 is identical to equation 5 from Heilmann et al. and equivalent to equation 7 from Schabel and Parker.[91,92]

Image-derived arterial input functions

The delineated neck arteries were: vertebral arteries, internal carotids and external carotids (see Figure 1). Each artery was manually delineated on the left and right side separately on the third most cranial slice of the image to minimize the effect of in-flow, while avoiding inaccuracies at the outer edges of the field of view. Delineation was performed using in-house developed software by a single observer in one session. The time frame of maximum enhancement in the arteries was used during delineation. The later time frames were used to identify veins, as these show a later time of contrast arrival. If the identified veins showed overlap with the delineated artery, the delineation was edited to exclude the vein from the arterial regions of interest. Images were visually inspected for movement and artefacts. Data was excluded for further analysis when movement in the arterial regions of interest was >2 mm during the DCE image acquisition.

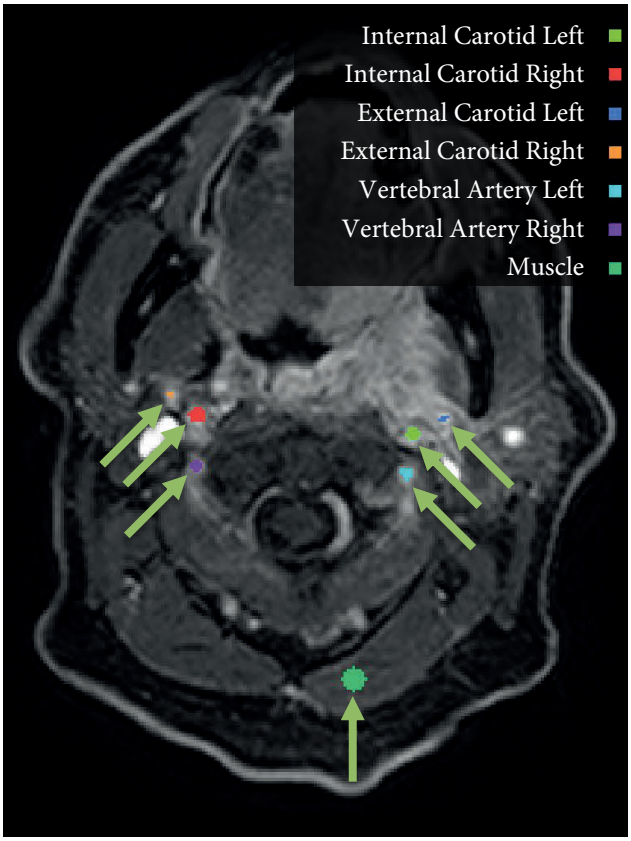


Figure 1: Volumes of interest shown on the last dynamic frame of the DCE image. Arteries were separately delineated on the third most cranial slice, currently shown. The circular region of interest of 6 mm in diameter was placed manually in left semispinalis capitis muscle tissue on all slices except the two most cranial and two most caudal slices.

Signal intensities from the arteries were extracted from the dynamic contrast enhanced images by taking the average over the cross-section for each arterial region of interest. This was done for left and right regions of interest separately and for both combined, i.e. considering left and right regions as one region of interest and taking the average signal intensity of all voxels within this combined region. Enhancement of these signals over time was converted to tracer concentration using Equation 4, defining S_0 as the average of the first four time frames and assuming a T_{10} value in blood of 1932 ms taken from literature.[93] A correction for flip angle θ was performed by using the average value of the B1-map in each arterial region of interest. A fixed hematocrit level of 0.42 was used to convert to plasma concentration as described by Parker et al.[81]

Similarly to Klawer et al.[94], the resulting concentration-time curves were fitted to the model of Parker et al.[81] to extract parameters describing the magnitude and shape of the AIFs. An example of this fit is shown in Figure 2. Several parameters were defined: maximal concentration (peak), time to peak, area under the curve (AUC), full width half maximum (FWHM), concentration at 180 seconds (C_{180}) and the exponential decay constant of the sigmoid modulated exponential in the Parker model, describing the tail of the concentration-time curve (washout). To ensure the FWHM only described the width of the first peak, the FWHM value was considered invalid if the value was >30 seconds.

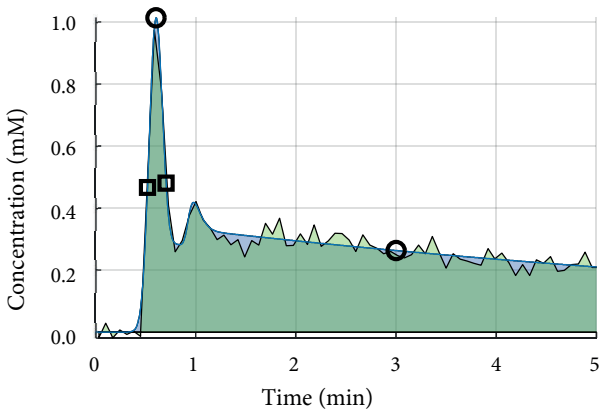


Figure 2: Example of fitting the Parker model[81] to the image-derived AIF. The area under the fit curve is filled with blue, the area under the data is filled with green. The identified peak value and the value of the fit at 180 seconds are circled. The time points used for the FWHM are indicated by squares.

Kinetic parameters in muscle

A circular region of interest of 6 mm diameter (see Figure 1) was placed in the left semispinalis capitis muscle on the DCE image on all slices except the two most cranial and two most caudal slices, to avoid inaccuracies at the edges of the field of view. In some patients the muscle did not lie completely in the field of view and hence less slices were

included in the volume of interest. To minimize spatial mismatch, delineation of the muscle region on during treatment scans was performed while also showing the pre-treatment delineation. Signals of all voxels within the volume of interest were extracted from the DCE image and converted to concentration time curves with correction for the transmitted radiofrequency field using the B1map and using the T_{10} values from the T_1map. Mean concentration was calculated after converting to concentration for each voxel independently. Mean concentration time curves were fitted to the standard Tofts model[73] (without the vascular space) as given by

$$\begin{aligned} C_t(t) &= K^{trans} \cdot e^{-k_{ep} \cdot t} \otimes C_p(t - \Delta T) \\ &= K^{trans} \cdot \int_{\tau=0}^{t} C_p(\tau - \Delta T) \cdot e^{-k_{ep} \cdot (t-\tau)} d\tau, \end{aligned} \tag{5}$$

where C_t is the tissue concentration time curve, C_p is the plasma concentration time curve and ΔT is the time delay between the plasma curve and arrival time in tissue. The kinetic parameters K^{trans} (rate constant from plasma to the interstitial space), k_{ep} (rate constant from the interstitial space to plasma) and v_e (fractional volume of the interstitial space and the ratio of K^{trans} and k_{ep}) were estimated. The fit was performed using a nonlinear least squares fitting procedure, constraining the kinetic parameters to positive values and using multiple starting values.[95] The model was fitted numerically using each image-derived AIF, and the population averaged AIF as described by Parker et al.[81] All data processing was performed in Matlab, version R2017b.

Statistical analysis

Before repeatability assessment, we checked if there were significant differences between the repeated measurements by calculating the average difference and its 95% confidence interval.[96] The repeatability of the kinetic parameters was then assessed using the within-subject coefficient of variation (wCV)[97], as calculated by Equation 6

$$wCV = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{(x_{i,1} - x_{i,2})^2 / 2}{((x_{i,1} + x_{i,2})/2)^2}}, \tag{6}$$

where n is the number of patients and $x_{i,1}$ and $x_{i,2}$ are parameter values for patient i in session 1 and 2, respectively. A low value of wCV represents a high repeatability. Differences between the AIFs in descriptive parameters of the AIFs and in kinetic parameters of the muscle, and differences between wCVs were tested for significance using the nonparametric Wilcoxon matched-pairs signed ranks test. These tests were performed for left-right comparison of the three arteries, comparison between the arteries (left and right region combined of internal carotids, external carotids and vertebral arteries were compared to each other) and each artery (left and right region combined) with the population averaged AIF. The significance level was set to 0.05, after

Bonferroni correction this level was 0.000397. The analyses were performed with GraphPad Prism, version 7.04.

Results

Data from 10 of 29 patients were excluded because movement in the arterial regions of interest was >2 mm. Signal from the right vertebral artery of one patient could not be used because the measured enhancement was too high and conversion to concentration was not possible for the peak signal because enhancement was higher than the relationship in Equation 3 permits. The signal of left and right vertebral artery combined was also excluded for this patient. Plots of image-derived AIFs of the internal carotids, of the tissue time-concentration curves with model fit and p-values of all tests can be found in the supplemental materials.

Image-derived arterial input functions

FWHM values were invalid for the AIF from the left internal and external carotids in one patient, from the left external carotid in another patient and from all but the left external carotid in a third patient. These AIFs showed a relatively low peak and the FWHM therefore did not describe the width of the peak. Figure 3 shows boxplots of the parameters describing the image-derived AIFs from the pre-treatment datasets. Boxplots from the during-treatment data can be found in the supplemental material.

Left-right differences were small and not significant for any of the arteries. The arterial plasma concentrations measured in this study were generally lower than the population averaged AIF measured by Parker et al., which gives an approximate plasma concentration of 10 mM at the peak and 1 mM at 180 seconds. The concentrations found in the current study showed median peak concentrations below 2 mM and median concentrations at 180 seconds below 0.5 mM for all image-derived AIFs. The parameters describing the magnitude of the AIF (i.e. peak, AUC and C180) were lower in the external carotids, higher in the vertebral arteries and intermediate in the internal carotids. These differences were significant only between the external carotids and vertebral arteries. Differences in TTP and FHWM between the different arteries were not significant. Washout was significantly different between the internal and the external carotids.

Repeatability of image-derived arterial input functions

Figure 4 shows the repeatability of the AIF describing parameters. The parameters describing the internal carotids generally showed the best repeatability, except for the peak and TTP. No significant differences were found between arteries.

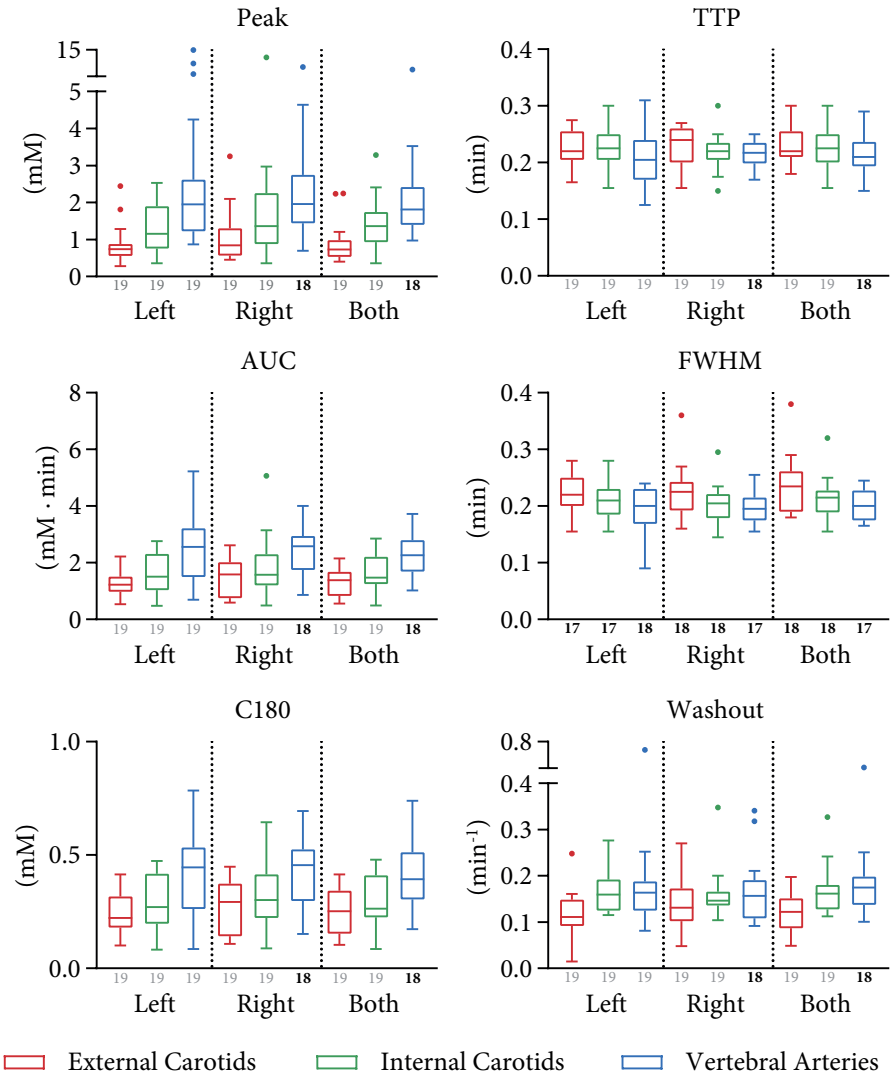


Figure 3: Boxplots with Tukey whiskers of the parameter values describing the image-derived AIFs before treatment. Number of subjects for each boxplot is indicated by the number below it.

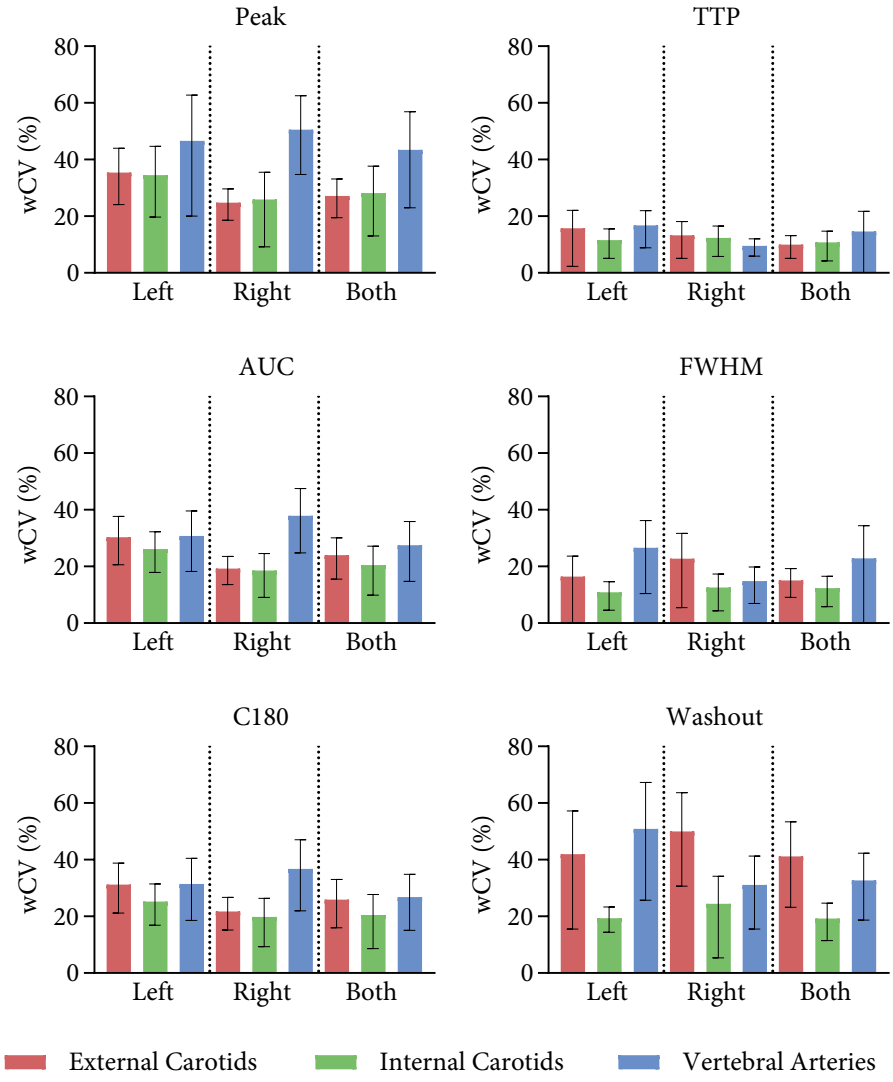


Figure 4: Bar plots of the wCV of the descriptive parameters of the image-derived AIFs. The error bars indicate the 95% confidence interval.

Kinetic parameters in muscle

Five out of the 38 fits to the Tofts Model (Equation 5)—three from pre-treatment imaging, two from imaging during treatment, none of which referring to the same patient—provided unrealistic results when the population averaged AIF was used: the fitted v_e values were above 1 (range 5–160). These data were excluded from the results below. When an image-derived AIF was used these same data often—though not always—led to v_e values above 1 as well. However, because the v_e values fitted with the image-derived AIFs were in general higher than those fitted with the image-derived AIF, possibly due to underestimation of the arterial concentration as result of flow and T2 shortening, the results were only excluded if the v_e was above 3. This criterion led to exclusion of data of three patients for all image-derived AIFs and of one patient for all AIFs except those derived from the right vertebral artery and combined vertebral arteries. These four patients were also excluded when using the population averaged AIF. One additional patient was excluded for the AIF derived from the left vertebral artery.

Figure 5 shows boxplots of the fitted pharmacokinetic parameters in the muscle before treatment. The boxplots from the during-treatment data can be found in the supplemental material. No significant differences between pre- and during-treatment data were observed. No significant differences arising from using either a left or right location of the AIF were observed in any of the parameters for any of the arteries. The values of all parameters—but most notably K^{trans}—were higher when an image-derived AIF was used, compared to those obtained using the population averaged AIF. These differences were significant for all comparisons except k_{ep} and K^{trans} between the vertebral arteries and the population averaged AIF. Comparisons between the different arteries from which the AIFs were derived showed significant differences between the vertebral arteries and the external carotids for K^{trans} and k_{ep}, but not for v_e. Values for K^{trans} and k_{ep} were the largest when using the external carotids, followed by the internal carotids and the vertebral arteries. The fitted v_e values were sometimes larger than 1 when image-derived AIFs were used, most often when using the external carotids.

Repeatability of kinetic parameters in muscle

Figure 6 shows the repeatability of the pharmacokinetic parameters when using the various image-derived AIFs and the population averaged AIF. The wCV for each of the three parameters was lowest (i.e. highest repeatability) when a population averaged AIF was used, and highest when the AIF was derived from the vertebral arteries. However, no significant difference was observed for any of the comparisons.

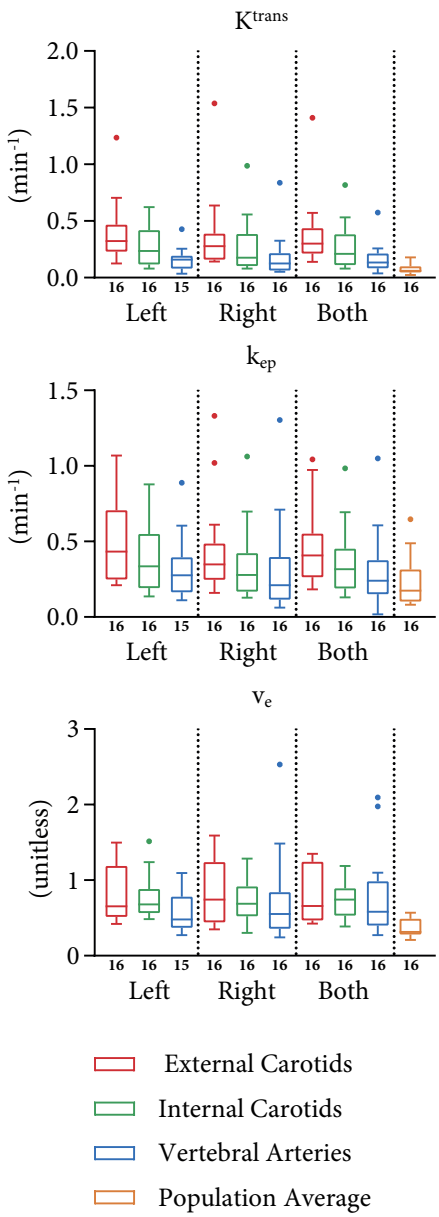


Figure 5: Boxplots with Tukey whiskers of the pharmacokinetic parameter values in the muscle before treatment. Number of subjects for each boxplot is indicated by the number below it.

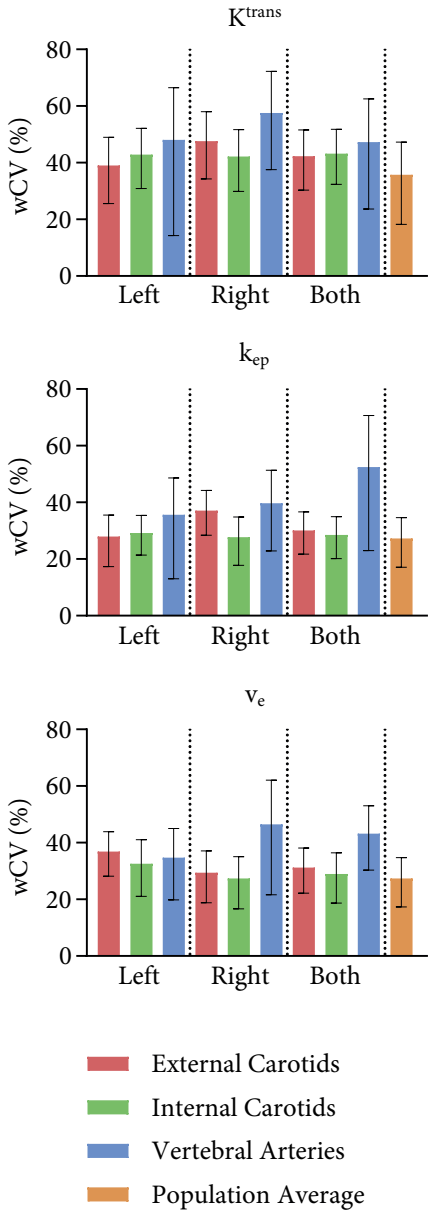


Figure 6: Bar plots of the wCV of the muscle pharmacokinetic parameters. The error bars indicate the 95% confidence interval.

Discussion

This study shows that image-derived AIFs obtained from different arteries in the head and neck region in the same patient differ in both magnitude and shape. Pharmacokinetic parameters in muscle, obtained using AIFs originated from different arteries, also showed significant differences. Moreover, use of a population averaged AIF led to significantly lower values of K^{trans}, k_{ep} and v_e and slightly better repeatability, although differences in repeatability between different AIF methods were not significant.

Image-derived arterial input functions

The image-derived AIFs from this study seem to underestimate the arterial plasma concentration when compared to the population averaged AIF or AIFs obtained by DCE-CT.[98] Keil et al. have observed similar results in the comparison of the internal carotid, superior sagittal sinus, arteries closest to brain lesions and Parker's population averaged AIF, where arterial regions provided markedly lower concentration curves.[79] This is likely caused by blood flow artefacts and partial volume and T_2^* effects.[99-101,92] These effects are dependent on the arterial region of interest, which may be why Parker's AIF, measured in the descending aorta, provides higher concentrations. Moreover, sensitivity for blood flow and other artefacts is dependent on the sequence settings.[92] More accurate measurements in the head and neck region might be achieved with different settings; however, this generally leads to inferior temporal and spatial resolution.[92] Additional use of phase images has been shown to lead to more repeatable AIFs which are less affected by flow.[94] In the current study, however, phase images were not available.

While there are no significant differences in magnitude between left and right measured AIFs, the data indicate that the magnitude and shape of an image-derived AIF are dependent on the choice of the artery. These differences are larger than what physiological differences might suggest. They might be partly explained by differences in artery diameter. Smaller arteries, such as the external carotids, are likely to be more influenced by partial volume effects, possibly resulting in lower measured concentration. Moreover, differences in flow velocities can be responsible for the differences in concentration due to the in-flow effect, possibly explaining the higher concentration in the vertebral arteries which exhibit lower flow velocity.[102]

Also the repeatability of the image-derived AIF seems to be affected by the choice of artery. The internal carotids seem to give the most repeatable AIFs, especially in terms of washout. The vertebral arteries tend to have higher signal enhancement than the other arteries, suggesting that they provide more accurate concentration values. However, this may also be the reason for the poorer repeatability of the AIFs from the vertebral arteries:

because the relationship between signal enhancement and concentration in Equation 3 is nonlinear and flattens at higher concentrations of contrast agent, the value of the estimated concentration at low T1 (high concentrations) is more sensitive to noise. This results in increased concentration variability. Use of a low dose pre-bolus scan can (partially) resolve this saturation effect and may lead to more repeatable AIFs.[103-107] One other study measured AIFs in the carotids at multiple time points (to generate a population averaged AIF); however, repeatability of the individual measurements was not investigated.[108]

Kinetic parameters in muscle

The differences in the parameters describing the image-derived AIFs seemed to propagate into the resulting pharmacokinetic parameters. Use of an image-derived AIF leads to significantly higher parameter values compared with using a population-based AIF, especially for K^{trans}. This is caused by the lower amplitude of the image-derived AIFs, as discussed above.

Although the resulting pharmacokinetic values are different, in terms of repeatability no significant differences between them were found when using the different AIFs. The repeatability of k_{ep} and v_e was comparable when using either a population averaged AIF or an AIF derived from the internal carotids. This result differs from the study of Rijpkema et al.[83], who found that repeatability of k_{ep} was better if individual AIFs were used. In their dataset, 6 patients are included with a tumor in the head and neck region; however, a different sequence is used (the flip angle is particularly different) and this may explain the disparity with our study. Peled et al.[77] also found that k_{ep} repeatability improved by using individual AIFs, although their study covers the prostate. In accordance with some literature[81,82], but contradicting other[84], the repeatability of K^{trans} seems to improve when a population averaged AIF is used. Ideally, use of an image-derived AIF corrects for variability in cardiac output within the patients over time, thus leading to a better repeatability of the pharmacokinetic parameters. Apparently, however, the variability introduced by the AIF measurement counteracts this effect. Variability could be caused by partial volume effects, B1 errors and flow enhancement artefacts. Because this is different for other acquisition settings, the generalizability of our results is limited.

When the population averaged AIF was used, the tissue curves in five subjects were fitted with v_e above 1, indicating that the data do not adhere to the theoretical model, either because the population averaged AIF cannot lead to the tissue time-concentration curve, or the tissue time-concentration curve is incorrect. The latter might be explained by inaccurate T_1 estimation or errors in B1 that are not accounted for by the B1 correction. This would explain why the model also produced outliers when the image-derived AIFs were used in four of the five cases. In one case the image-derived AIFs showed a dispersed

shape and fitting led to reasonable pharmacokinetic parameter values and a better fit, indicating that use of a population averaged AIF was inappropriate in this case.

Limitations

More than one third of the data could not be evaluated. Deriving the AIF from the image is problematic when the patient moves or swallows during acquisition. Motion correction for these, often small and quick, movements is not straightforward and was not performed in this study. Use of a population averaged AIF largely overcomes this problem, although movement can also affect signals from the tissue.

Repeatability estimates within tumor and lymph nodes are necessary for biomarker validation of DCE-MRI pharmacokinetic parameters in the head and neck cancer, such as HNSCC. However, due to the setup of this study such estimates could not be investigated, because between baseline and follow-up the patients underwent chemo radiotherapy. Moreover, the repeatability estimates reported for muscle cannot be extrapolated to, for example, tumor tissue; the different tissue characteristics in tumor (or tissues in which they arise) lead to different pharmacokinetic parameter values and their repeatability is likely different.[109] The semispinalis capitis muscle was chosen as it is located outside the radiated area. However, the combination of (chemo) radiotherapy and the ongoing disease might have some systemic effects, even within this short period of time between scans. As reported, no significant differences were found between pre- and during-treatment kinetic parameters in muscle; however, if the effects of disease and therapy caused an increased variability, the wCV's reported here may be overestimated. Nonetheless, we believe that the repeatability of the pharmacokinetic parameters in the muscle region can still be a useful tool for comparison of the use of different AIFs as input of the model.

Conclusion

Significant variations were found in the AIFs obtained from different arteries in the head and neck region. Image-derived AIFs measured in the internal carotids show a trend to better repeatability for both the AIF itself and for the pharmacokinetic parameters estimated in muscle tissue. However, the image-derived AIF does not improve repeatability of the pharmacokinetic parameters compared to a population averaged AIF. Moreover, patient movement during acquisition, which can be common in the head and neck region, is likely to disturb AIF measurement. For these reasons, the use of a population averaged AIF in this patient population seems to be preferable for pharmacokinetic analysis of DCE-MRI when absolute PK parameter values are not of major concern.

Supplementary information

Supplementary material for this paper is available at:

doi.org/10.1016/j.mri.2020.01.010

6

Repeatability of IVIM biomarkers from diffusion-weighted MR imaging in head and neck: Bayesian probability versus neural network

Thomas Koopman*, Roland M. Martens*, Oliver J. Gurney-Champion, Maqsood Yaqub, Cristina Lavini, Pim de Graaf, Jonas Castelijns, Ronald Boellaard, J. Tim Marcus

* These authors contributed equally to this manuscript.

Published 26 January 2021
Magnetic Resonance in Medicine 2021; 85: 3394–3402
DOI: 10.1002/mrm.28671

Abstract

Purpose: The intravoxel incoherent motion (IVIM) model for diffusion-weighted imaging might provide useful biomarkers for disease management in head and neck cancer. This study compared the repeatability of three IVIM fitting methods to the conventional nonlinear least-squares regression: Bayesian probability estimation, a recently introduced neural network approach, IVIM-NET, and a version of the neural network modified to increase consistency, IVIM-NET$_{mod}$.

Methods: Ten healthy volunteers underwent two imaging sessions of the neck, two weeks apart, with two DWI acquisitions per session. Model parameters (apparent diffusion coefficient ADC; diffusion coefficient, D_t; perfusion fraction f_p; pseudo-diffusion coefficient D_p) from each fit method were determined in the tonsils and in the pterygoid muscles. Within-subject coefficients of variation (wCV) were calculated to assess repeatability. Training of the neural network was repeated 100 times with random initialization to investigate consistency, quantified by the coefficient of variance (CV).

Results: The Bayesian and neural network approaches outperformed nonlinear regression in terms of wCV. Inter-session wCV of D_t in the tonsils was 23.4% for nonlinear regression, 9.7% for Bayesian estimation, 9.4% for IVIM-NET and 11.2% for IVIM-NET$_{mod}$. However, results from repeated training of the neural network on the same dataset showed differences in parameter estimates: CV over the 100 repetitions for IVIM-NET were 15% for both D_t and f_p, and 94% for D_p; for IVIM-NET$_{mod}$ these values improved to 5%, 9% and 62%, respectively.

Conclusion: Repeatabilities from the Bayesian and neural network approaches are superior to that of nonlinear regression for estimating IVIM parameters in the head and neck.

Introduction

Magnetic resonance diffusion-weighted imaging (DWI) is used for diagnostic and prognostic purposes in head and neck cancer.[110-113] In DWI, signal decreases with diffusion-weighting as result of Brownian motion of water molecules and other intravoxel incoherent motions (IVIMs), i.e. "microscopic translational motions that occur in each image voxel".[114,115] By fitting the DWI signal from different diffusion-weightings to an exponential model, its parameters can be estimated. A mono-exponential can be used to estimate the apparent diffusion coefficient (ADC). A bi-exponential model (the IVIM model[115]) can be used to additionally model pseudo-diffusion component (D_p) and perfusion fraction (f_p)—both related to the microcirculation of blood—resulting in the corrected or "true" diffusion coefficient (D_t). Because the restriction of diffusion is related to the microstructure of tissue—e.g. cellular density—this can characterize tumors and provide early information on changes due to (or despite) treatment occurring before detectable tumor growth or shrinkage.[116]

The IVIM model is appealing as it allows the assessment of the additional biomarkers D_p and f_p. However, IVIM parameter estimation tends to be very sensitive to noise. As a result, parametric maps are often noisy and show poor repeatability.[117] Poor repeatability limits the use of IVIM in practice because precision is required for patient-specific clinical use of IVIM.

Recently, novel fitting methods with a Bayesian probability approach[118-120] and a neural network[121] have shown promising results in terms of reduced noise in the parameter maps based on simulations, and they reduced inter-observer variability in vivo. If these techniques also help improve test-retest repeatability in vivo, they could help introduce IVIM into clinical workflows.

Therefore, in this study, we investigate these new methods in terms of test-retest repeatability. We compare the intra- and intersession repeatability of the least squares fitting method, the Bayesian inference fitting method and two neural network-based fitting methods for in vivo IVIM data in the head and neck region in healthy volunteers. We hypothesize that the new Bayesian and neural network approaches will outperform the conventional least squares fitting approach.

Methods

This study was approved by the local Medical Ethics Committee and written informed consent was obtained from all subjects. Ten healthy volunteers were included; male/female 7/3, mean age 33 years, range 22–50 years. Each volunteer underwent two MR imaging sessions (at least two weeks apart) with two examinations per session. The subject was taken out of the MR scanner between examinations. Sequences were acquired on a 3.0T Ingenuity TF PET/MR-scanner (Philips Healthcare, Best, the Netherlands) equipped with a 16-channel neuro-vascular coil. Each examination consisted of an axial stack of 29 T1-weighted turbo-spin-echo images followed by a stack of DWI acquisitions in the same 29 imaging planes, covering the neck from the larynx until the base of the skull. DWI was acquired with a single-shot spin-echo echo-planar imaging sequence with 12 b-values (0, 2, 5, 25, 50, 75, 100, 150, 300, 500, 700, and 1000 s/mm^2). Only the DWI images with b=1000 s/mm^2 were averaged over 2 acquisitions. Diffusion-weighting was performed in three orthogonal directions with: bipolar gradients, echo time 57 ms, repetition time 3242 ms, gradient time interval 28 ms, gradient duration 18 ms. Further scan parameters were: acquired matrix size 128×111×29, acquired voxel size 1.88×1.95×4 mm^3, reconstructed voxel size 1×1×4 mm^3, Short TI Inversion Recovery (STIR) was used for fat suppression, with a 230 ms inversion time. The DWI scan duration was 6 minutes. Motion correction of the DWI images was applied by image registration, as provided by the scanner software.

Analysis

The DWI data were processed voxelwise to generate parametric maps of the ADC and IVIM parameters. Parameter estimates were extracted for two tissues: tonsil and medial pterygoid muscle. Volumes of interest (VOIs) were defined on the images without diffusion weighting ($b = 0$ s/mm^2) while using the T1-weighted image for anatomical reference. The T1 images were not co-registered to the diffusion weighted images. Delineation was performed using in-house developed software by a single observer in one session. Spherical VOIs of 5 mm radius were placed in each tonsil and spherical VOIs of 6 mm radius were placed in the medial pterygoid muscle on each side. These VOIs were small enough to always fit inside the tissues of interest. The VOIs were projected onto the parametric maps, all voxels with (partial) overlap were extracted, and the median values of the parameters were then calculated.

The signal at $b = 0$ s/mm^2 was excluded (except for calculating fit boundary of S_0, see below, and for normalization purposes in the neural network) for the parameter estimations described below, reducing the number of b-values to 11. The reason for this is to reduce attenuation effects of macroscopic flow at small b-values[122,123]. This additional accelerated decay between $b = 0$ s/mm^2 and the first non-zero b-value is not accounted for in the conventional IVIM and ADC models.[124,125]

The mono-exponential model used to estimate the ADC is given by:

$$S(b) = S_0 \cdot e^{-b \cdot ADC}\,, \tag{1}$$

where S_0 is the signal intensity without diffusion weighting ($b = 0$ s/mm^2). ADC and S_0 are estimated by performing a linear least squares fit on the log-transformed data, as implemented by the scanner manufacturer.

The IVIM model extends the ADC model with a second exponential. The bi-exponential equation of the model is given by:

$$S(b) = S_0\left(f_p e^{-bD_p} + (1 - f_p)e^{-bD_t}\right) \tag{2}$$

where f_p is the perfusion fraction, D_p the pseudo-diffusion coefficient, D_t is the diffusion coefficient, and S_0 is the fitted signal intensity for $b = 0$ s/mm^2. The IVIM model parameters were estimated using four different approaches: a nonlinear least squares fit, a Bayesian approach[118] and two neural network-based fitting approaches[121]. The two neural network approaches consisted of a network nearly identical to the original publication (IVIM-NET)[121] and a modified network (IVIM-NET$_{mod}$), as detailed later.

Nonlinear least squares

The nonlinear least squares (NLS) fit was performed using the trust-region reflective algorithm as implemented in MATLAB R2019a, with the following fit boundaries: $0 < f_p < 1$, $0 < D_t < 0.005$ mm^2/s, $0.005 < D_p < 1$ mm^2/s, and $0 < S_0 < 5 \cdot \max S(b)$. Starting values were selected randomly in the range [0,1] as provided by Matlab functionality.

Bayesian probability

The Bayesian approach was also performed in MATLAB R2019a and was based on a previous publication.[118] In short, the method gives a maximum a-posteriori estimate of each parameter by maximizing the marginal posterior probability density functions, which are acquired by means of slice sampling[126] the joint posterior probability.[118,126,127] A multiparametric Gaussian likelihood function was used, i.e.

$$P(S|D_t, D_p, f_p, S_0) \propto \left(\frac{1}{2}\sum_{\{b\}}\left(S(b) - S_0\left(f_p e^{-bD_p} + (1 - f_p)e^{-bD_t}\right)\right)^2\right)^{-n/2} \tag{3}$$

where n is the number of b-values. The constraint $D_t < D_p$ was implemented in the joint prior distribution.[128] Lognormal distribution priors were used for D_p and D_t, a beta distribution prior was used for f_p and a uniform distribution prior was used for S_0. The priors for D_p, D_t and f_p were estimated by fitting these distributions to results of a pre-run of the same Bayesian approach using bounded uniform priors ($0 < f_p < 1$, $0 < D_t < 1$ mm^2/s, $0 < D_p < 1$ mm^2/s and $0 < S_0$.)

IVIM-NET

The IVIM-NET approach was carried out in Python 3.7.4 and PyTorch 1.3.0 using the open access code from the original publication[121,129] and code obtained from the repository of co-author OGC (currently shared on request, will be public in near future and maintained as IVIM-NET evolves). Our source code with the network definitions and training methods is available on GitHub.

The network, depicted in Figure 1, consists of an input layer with a number of neurons equal to the number of b-values used to analyze the data, three fully connected hidden layers—each with the same number of neurons, each using the exponential linear unit activation function—and an output layer with a neuron for each parameter. Background voxels were excluded by manually thresholding the $b = 0$ mm²/s images. Training was performed on the entire dataset for each epoch, combining and shuffling the voxels from all patients. Data normalization, which is standard for neural networks, was performed using S(0). The mean squared error between the fitted and actual, normalized, signal ($S(b)/S(0)$) was used as loss function. An early stopping criterium (patience) of 10 bad epochs was used: meaning training was stopped when no improvement was found during the last 10 epochs. Different from the original publication, we included an output neuron for $S_0/S(0)$, where S_0 and $S(0)$ are the estimated and measured signal intensity at $b = 0$ mm²/s, respectively.

Additional to the above implementation, we made a few modifications in a new implementation IVIM-NET$_{mod}$. IVIM parameters were constrained by g(x). In the original network, the predicted IVIM parameters were constraint by taking the absolute:

$$g(x) = |x| \tag{4}$$

In the presented modified network, a sigmoid function was applied to the output as constraint instead:

$$g(x) = \min + \frac{1}{1 + e^{x}}(\max - \min) \tag{5}$$

which rescaled the output between the following fit boundaries (min < parameter < max): $0 < f_p < 0.7$, $0 < D_t < 0.005$ mm²/s, $0.005 < D_p < 0.5$ mm²/s, and $0.8 < S_0/S(0) < 1.2$. Second, with the aim of preventing overfitting, we split the dataset into two parts: one for training (80%) and one for validation (20%). For the same reason, we reduced the patience (early stopping criterion, see above) from 10 to 4. Furthermore, as we had a substantially larger dataset than Barbieri et al, we limited the number of iterations during each training epoch to 1024, such that we regularly validate how well the network is performing even for large datasets. Because the batch size of an iteration is fixed (128 voxels), each epoch no longer processes the entire dataset but a random selection of the

training set, in our case, approximately 1.5%. Each epoch does evaluate on the entire validation set.

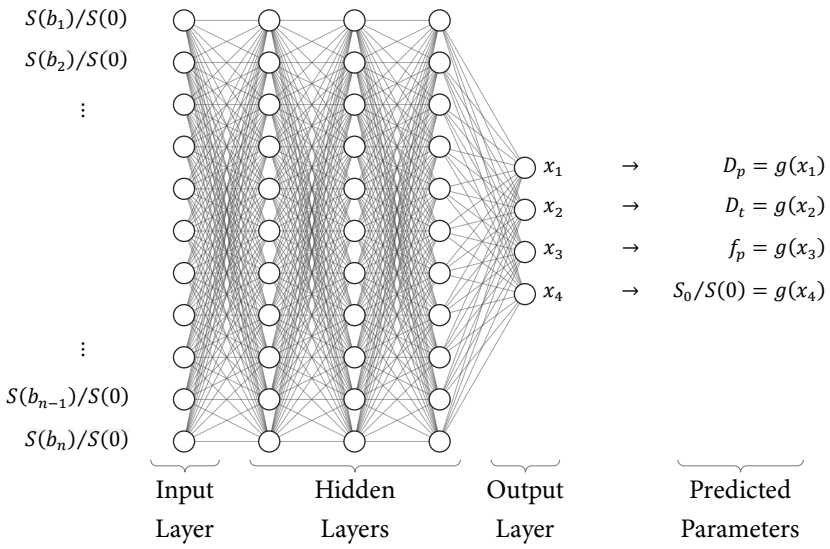


Figure 1: Neural network architecture, figure created with NN-SVG.[130] The network predicts x_1 to x_4, which are converted to the IVIM parameters by the constrain function g(x) via equations 4 (original network) and 5 (modified network) to add parameter constraints.

Network consistency

To investigate the consistency of the IVIM-NET approaches as a whole—i.e. whether the network converges to consistent estimates—we repeated the complete process of training 100 times. Each time, the network was initialized with new random weights and shuffling (and splitting in case of IVIM-NET$_{mod}$) of the dataset. We compared the runs qualitatively by visual inspection of the parametric maps. We investigated consistency by calculating the average parameter values for both tissues over all subjects and sessions in each run, and then calculating the coefficient of variance (CV) over the 100 runs.

Statistics

The intrasession repeatability was calculated by considering the two measurements within a session as paired measurements. Conversely, the intersession repeatability was calculated by considering the first measurements in each session as one pair, and the second measurements in each session as another pair. Moreover, the left and right measured values were considered measurements for the same tissue of interest, i.e. tonsil and pterygoid muscle. Thus, each subject had four pairs of observations for the calculation of repeatability and pairs were either between sessions or within sessions.

We used 95% confidence intervals of the mean difference between paired measurements over all subjects (for both intra- and intersession pairing) to verify that the repeated measurements were not systematically different.[96] We then calculated the within-subject coefficient of variation (wCV), which is a relative measure of repeatability.[97] An overview of the concepts repeatability and consistency can be found in Table 1.

We compared the repeatability of the four methods with paired Wilcoxon signed rank tests of the wCV estimates. For the IVIM-NET methods we calculated the median wCV of the 100 runs for each subject, and used these wCV estimates in the paired tests between the four methods. A p-value below 0.05 was considered significant.

Table 1: Explanation of analysis concepts used in this study.

Concept	Description	Quantification	Applicable for
Repeatability	Variation between repeated measurements	Within-subject coefficient of variation, wCV	All methods
Consistency	Variation between training runs of IVIM-NET on same measurements	Coefficient of variance, CV	IVIM-NET

Results

Figure 2 shows examples of the parametric maps calculated with the different methods. Parametric maps calculated by nonlinear regression were most noisy, followed by the Bayesian probability approach. IVIM-NET showed the least noise and most anatomical detail and was in these terms comparable to the ADC map. The f_p maps estimated with nonlinear regression showed systematically higher values than the other methods, as can be seen in Figure 2. The D_p maps of nonlinear regression and IVIM-NET$_{mod}$ showed many regions with very high values.

None of the methods showed a systematic difference between the repeated measurements for any of the parameters. The calculated wCV estimates are shown in Figure 3. Notably, intra- and intersession wCV was comparable and both VOIs show the same patterns when comparing the IVIM methods. The methods differ in terms of repeatability and, in general, wCV was highest (worst) when parameters were estimated using NLS (except for f_p in the pterygoid).

This difference was often significant, especially for D_p. Significance is indicated in Figure 3, comparing the wCV values of each of the methods for each of the parameters. Tables of the p-values are available in the supplemental materials. The median repeatability results of IVIM-NET and IVIM-NET$_{mod}$ were mostly comparable to the repeatability of the Bayesian approach, except for f_p where the wCV of IVIM-NET was significantly better; IVIM-NET$_{mod}$ was only significantly better for f_p in the tonsils. The median repeatability of IVIM-NET was better than for IVIM-NET$_{mod}$, although the difference was rarely significant.

Network consistency

Visual comparison of the parametric maps of the repeated network instances showed inconsistencies for both IVIM-NET and IVIM-NET$_{mod}$, examples of this can be found in supplemental figures 2S, 3S and 4S. 100 figures of the maps of each method are included in the supplemental material as illustration. The network instances generally produced D_t maps with a similar distribution but different offset/scaling values. Comparing the average parameter values for IVIM-NET, the instances showed a CV of 15% for D_t and f_p, and 94% for D_p. The CV for the pterygoids and tonsils were equal. The D_p maps sometimes showed a visually different distribution. The wCV values for the IVIM-NET instances were also inconsistent, as shown by the boxplots in Figure 4. For IVIM-NET$_{mod}$, the average parameter values were more consistent, with CV of 5%, 9% and 62% for D_t, f_p and D_p respectively (again equal for pterygoids and tonsils). The wCV values for D_t and D_p were also more stable for IVIM-NET$_{mod}$, as reflected by the smaller intervals of the boxplots in Figure 4.

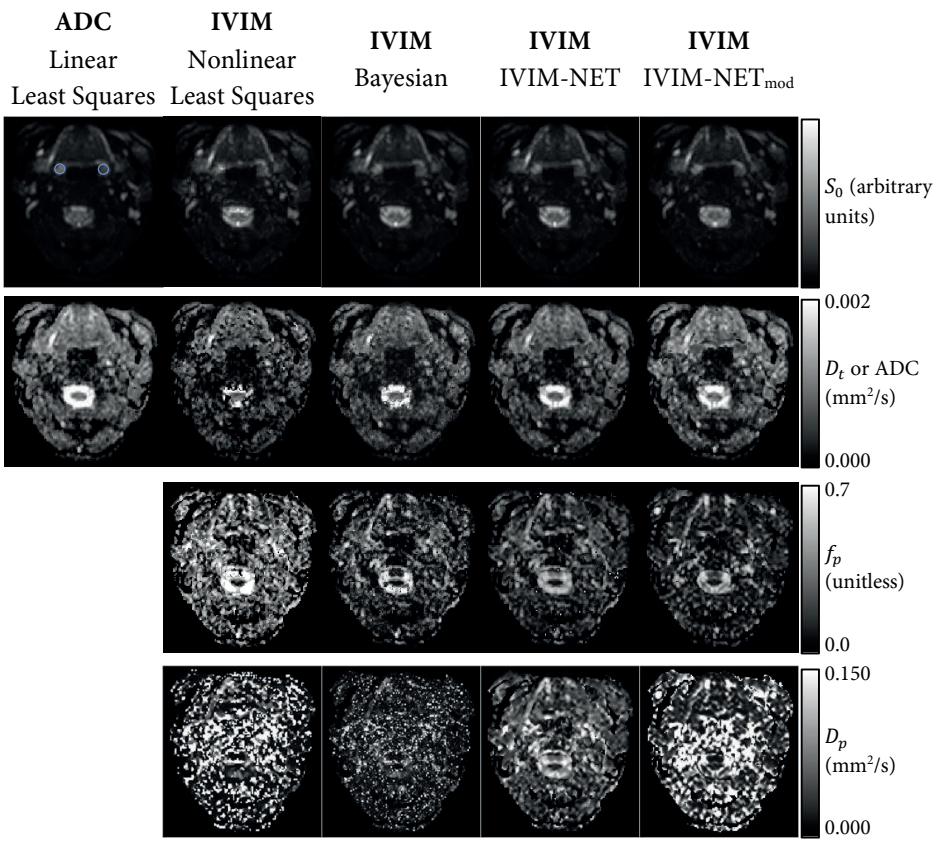


Figure 2: Typical parametric maps of the estimated S_0, apparent diffusion coefficient ADC, true diffusion coefficient D_t. pseudo diffusion coefficient D_p and perfusion fraction f_p. Regions of interest delineating the tonsils are shown in the ADC map. The pterygoids are not situated at this level, examples of regions of interest can be found in the supplemental material. Parametric maps of other IVIM-NET instances can be found in the supplemental material.

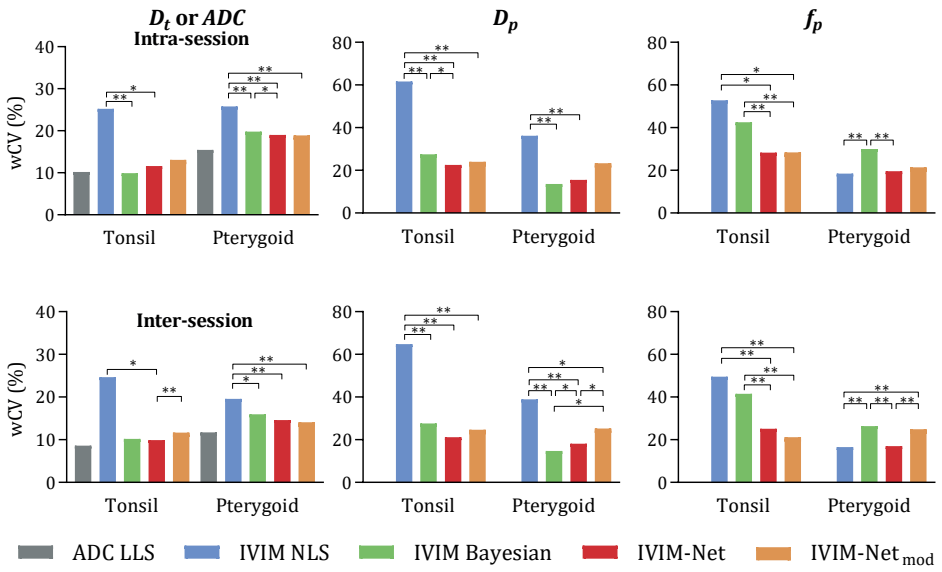


Figure 3: wCV of the parameters for each method. The median value of 100 training runs is displayed for the neural network methods. * $p \leq 0.05$, ** $p \leq 0.01$

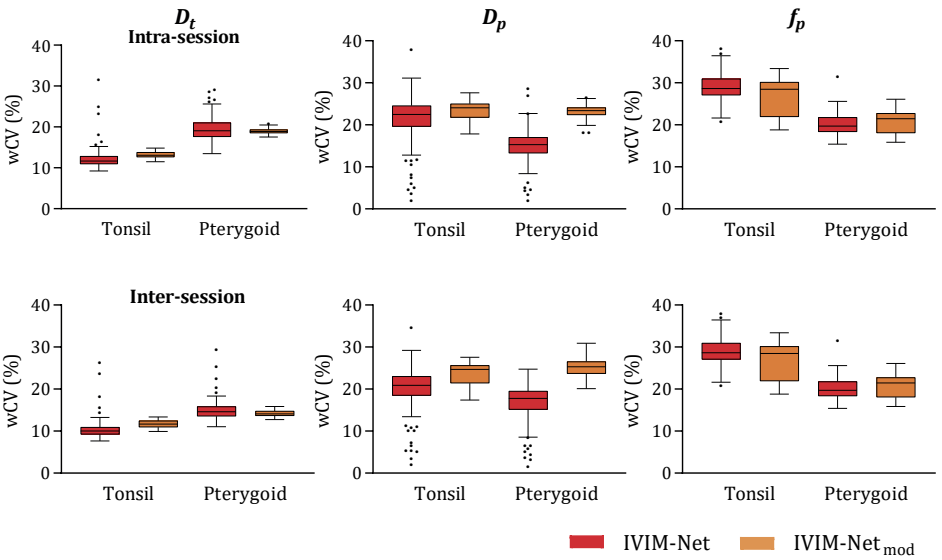


Figure 4: Boxplots (with Tukey type whiskers) of wCV values for 100 runs of the neural networks.

Discussion

In this study, we quantified the test-retest repeatability of nonlinear regression, neural network-based and Bayesian IVIM in the head-and-neck region. Our results show that these latter two fit approaches substantially outperform the conventional nonlinear regression approaches commonly used for IVIM fitting. Furthermore, although IVIM-NET has an improved test-retest repeatability, it has an additional uncertainty in that repeated training of networks gives inconsistent results on identical data.

Repeatability estimates of ADC in the tonsils using linear regression fit reported in this study are similar to those reported by Kang et al.[131] (also in the tonsils). Two other studies mainly focused on the primary lymph nodes, which makes it hard to compare results directly. Hoang et al. report a repeatability coefficient in percentages (15%), which is equivalent to a wCV of 5.3%.[132] The wCV reported by Paudyal et al. is 2.38% and much lower than repeatability values found in this study.[133] The generally larger volumes of metastatic lymph nodes might partly explain why the reported estimates are lower. In case of the study of Paudyal et al., no repositioning of the subject in the magnet seems to have occurred between scans, which could be a major source of measurement variability. This might also explain the difference in reported wCV between the two studies. In our present study, the subject was taken out of the scanner between scans for the intra-session repeatability estimates. No differences were seen between intra- and intersession repeatability. This indicates that long-term (order of weeks) physiological variability over time was secondary to the measurement error and short-term (~30 minutes) physiological variability.

The neural network approach for calculating IVIM maps was introduced only recently. Visual interpretation of the images suggests that more realistic parametric maps are produced by both neural networks compared to the other methods; these maps do not show isolated high or low pixels and thus seem to be least affected by noise in the acquisitions. This is in line with earlier observations from Barbieri et al.[121] Our study now has quantified the test-retest repeatability and shows that the network also outperforms linear regression regarding this aspect.

Network training for the entire dataset took up to one hour for IVIM-NET and up to 5 minutes for IVIM-NET$_{mod}$. Application of the network took only a couple of seconds for the entire dataset. Barbieri et al. had substantially less training data and hence had training times of 5 minutes using the unmodified network. The large difference in training time in our data was mainly the result of decreasing the amount of data seen each epoch in the IVIM-NET$_{mod}$. This major advantage of analysis speed, compared to the other methods investigated in this study (around half an hour per scan with nonlinear regression, and

multiple hours per scan using Bayesian probability fitting), makes it viable for use in clinical practice.

Although IVIM-NET showed promising test-retest repeatability, consistency of the neural network approach is currently still an issue. Our results show that, after renewed training, the parameter values and repeatability estimates vary. IVIM-NET$_{mod}$ showed more consistent results, although the method is still unstable for D_p. Consistency of the approach might be improved by optimizing the starting point of the network, for instance by choosing different weight initialization or by training on a set of simulated data first. Avoiding to fit D_p—fixing it instead to an a priori estimate—has been shown to improve repeatability[117,119] and might also improve network consistency.

A challenge for further research is to identify an acceptable neural network that does not only give estimates with good repeatability, but is also consistent after retraining. Until such consistency is achieved it is imperative that a single network instance is used for comparative applications, for example in longitudinal studies. Use of separately trained networks will otherwise lead to biased results.

Although other DWI models[134-136] are available, this study has been limited to the ADC and the IVIM model. Another limitation of our study is that we could not compare the methods in terms of accuracy, because a ground truth was unavailable in our study. We hope, therefore, that these methods will be included in future phantom studies. Lastly, the choice of b-values was probably not optimal; b-value optimization may improve IVIM estimates.[137,138]

Conclusion

The processing speed of the neural network makes it viable for use in clinical practice. However, the inconsistency of training results is challenging. Our presented modifications in the neural network make this approach more consistent, although the output still shows some inconsistency between different training runs on the same dataset. Thus, the neural network approach needs to be further improved to identify neural networks that are both consistent and precise. Nonetheless, repeatability from the Bayesian and neural network approaches are superior to that of nonlinear regression for estimating IVIM model parameters.

Acknowledgements

The authors would like to thank Dr. S. Barbieri for sharing his code for implementation of the Bayesian probability approach and for making the code of the neural network approach publicly available.

Data Availability Statement

The source code that supports the findings of this study are openly available at GitHub:

IVIM-NET and IVIM-NET$_{mod}$: https://github.com/koopmant/ivim-net, reference number 8943f073.

IVIM Bayesian probability: https://github.com/koopmant/ivim-bp, reference number a6d8f94a.

IVIM-NET main repository: https://github.com/oliverchampion/IVIMNET currently shared on request, will be public in near future and maintained as IVIM-NET evolves.

Supplementary material

Supplementary material for this paper is available at:

https://doi.org/10.1002/mrm.28671

7

Predictive value of quantitative [^{18}F]FDG-PET radiomics analysis in patients with head and neck squamous cell carcinoma

Thomas Koopman*, Roland M. Martens*, Daniel P. Noij, Elisabeth Pfaehler, Caroline Übelhör, Sughandi Sharma, Marije R. Vergeer, C. René Leemans, Otto S. Hoekstra, Maqsood Yaqub, Gerben J. Zwezerijnen, Martijn W. Heymans, Carel F. W. Peeters, Remco de Bree, Pim de Graaf, Jonas A. Castelijns and Ronald Boellaard

* These authors contributed equally to this manuscript.

Published 7 September 2020
EJNMMI Research 2020; 10: 102
DOI: 10.1186/s13550-020-00686-2

Abstract

Background: Radiomics is aimed at image-based tumour phenotyping, enabling application within clinical-decision-support-systems to improve diagnostic accuracy and allow for personalized treatment. The purpose was to identify predictive 18-fluor-fluoro-2-deoxyglucose ([^{18}F]FDG) positron-emission tomography (PET) radiomic features to predict recurrence, distant metastasis, and overall survival in patients with head and neck squamous cell carcinoma treated with chemoradiotherapy.

Methods: Between 2012 and 2018, 103 retrospectively (training cohort) and 71 consecutively included patients (validation cohort) underwent [^{18}F]FDG-PET/CT imaging. The 434 extracted radiomic features were subjected, after redundancy filtering, to a projection resulting in outcome-independent meta-features (factors). Correlations between clinical, first-order [^{18}F]FDG-PET parameters (e.g., SUVmean), and factors were assessed. Factors were combined with [^{18}F]FDG-PET and clinical parameters in a multivariable survival regression and validated. A clinically applicable risk-stratification was constructed for patients' outcome.

Results: Based on 124 retained radiomic features from 103 patients, 8 factors were constructed. Recurrence prediction was significantly most accurate by combining HPV-status, SUVmean, SUVpeak, factor 3 (histogram gradient and long-run-low-grey-level-emphasis), factor 4 (volume-difference, coarseness, and grey-level-nonuniformity), and factor 6 (histogram variation coefficient) (CI = 0.645). Distant metastasis prediction was most accurate assessing metabolic-active tumour volume (MATV) (CI = 0.627). Overall survival prediction was most accurate using HPV-status, SUVmean, SUVmax, factor 1 (least-axis-length, non-uniformity, high-dependence-of-high grey levels), and factor 5 (asphericity, major-axis-length, inversed-compactness and, inversed-flatness) (CI = 0.764).

Conclusions: Combining HPV-status, first-order [^{18}F]FDG-PET parameters, and complementary radiomic factors was most accurate for time-to-event prediction. Predictive phenotype-specific tumour characteristics and interactions might be captured and retained using radiomic factors, which allows for personalized risk stratification and optimizing personalized cancer care.

Introduction

Personalized cancer care of locally advanced head and neck squamous cell carcinoma (HNSCC) implies customization of therapy to the individual patient. This might improve the current overall 5-year survival rate of 50% (35–65%).[8] Radiotherapy with or without chemotherapy is frequently applied but fails in 50% of the cases. In the vast majority (about 90%), the locoregional failure occurs within the first 2 years after treatment.[139,140] The consequence of recurrent cancer is that surgical salvage therapy is generally the only option with curative intent, but this is associated with high morbidity.[141] More efficient pre-treatment response prediction may result in patient-tailored escalation or toxicity-reducing de-escalation (e.g., in radiosensitive HPV-positive patients) of (chemo)radiotherapy or a switch to different treatment options (e.g., surgery). Imaging is crucial in management because of its value on fast and non-invasive tumour staging, response monitoring, and prognosis prediction.[142] Exploration of quantitative imaging features might reflect underlying phenotype and response and thus may maximize the success of tailored treatments.[143]

Radiomics focuses on the methodology of extensive image-based tumour phenotyping.[144] With radiomics, it may be possible to characterize phenotypic differences providing information on the whole-lesion microenvironment and surrounding area accounting for spatial and temporal heterogeneity, such as cellular morphology, proliferative capacity, metabolism, motility, angiogenic and oxygenation status, gene expression (including expression of cell surface markers, growth factor, and hormonal receptors), proliferative, immunogenic, and metastatic potential.[142,143,145] These characteristics might be captured by radiomics-derived tumour features (i.e., intensity, shape, or texture) and might be of complementary value to other clinical parameters to predict their effect on the chemo-radiosensitivity (i.e., quantity of tumoral radiosensitive cancer stem cells, the hypoxic fraction, reoxygenation of the tumour vicinity, and/or repopulation capacity throughout the course of therapy).[144,146-148]

Radiomic features of functional imaging may provide additional information to anatomical imaging, because it provides information on pathophysiologic tumour characteristics.[149,150] Positron-emission tomography (PET)/computed tomography (CT) using 18F-fluoro-deoxy-glucose ([^{18}F]FDG) measures tumoral metabolic activity and can be quantified with [^{18}F]FDG-PET/CT by the standard uptake value (SUV). Pre-treatment [^{18}F]FDG-PET/CT was reported to be useful for detection, treatment decision support[151], planning[152,153], and the prediction and detection of recurrences and long-term outcome[139]. PET-radiomics was superior over a CT-based model (CIPET = 0.77 versus CICT = 0.72)[154] and might improve lesion characterization and patient outcome prediction compared to first-order PET parameters in daily clinical routine.[155-158]

Identified radiomic associations give insight in the biological basis of imaging appearance and could aid targeted treatment decision-making and predict prognosis non-invasively. Radiomics was mainly analysed in CT[159], or PET-CT separately[145,147], but when combined with clinical features, it resulted in higher predictive and prognostic value[154,160]. To our knowledge, a comparison of prediction models in head and neck with [^{18}F]FDG-PET radiomic factors, SUV measurements (e.g., maximum or peak SUV), and clinical parameters, associated with patient's outcome has not yet been described.

The aim of this study was to construct a model based on [^{18}F]FDG-PET radiomics features to predict locoregional recurrence, distant metastasis, and overall survival (OS) in patients with locally advanced head and neck squamous cell carcinoma treated with chemoradiotherapy.

Methods

Data selection

Between 2012 and 2014, 103 patients were included retrospectively in our training cohort. Between 2014 and 2018, 81 consecutive patients were included independently from the training cohort in a validation cohort. These training and validation single-centre cohorts were approved by the local institutional ethics committee (Amsterdam UMC Medisch Ethische Toetsing Commissie (METC), reference: 2013.191). A written informed consent was waived for the training cohort (reference: 2016.498), whereas for the validation cohort a written informed consent was obtained from all patients. Previously untreated patients with histologically proven HNSCC were included who were planned for chemoradiotherapy with curative intent (see Table 1). Exclusion criteria were nasopharyngeal tumours, age < 18 and pregnancy, previous locoregional treatment of HNSCC, or insufficient image quality. Within 5 weeks after baseline imaging, treatment was initiated consisting of a predetermined regimen of chemoradiotherapy (CRT) during a period of 7 weeks; 70 Gy in 35 fractions with concomitant cisplatin (100 mg/m^2 on days 1, 22, and 43 of radiotherapy)) or cetuximab (400 mg/m^2 loading dose followed by seven weekly infusions of 250 mg/m2). Tobacco use was defined as a smoking history of ≥ 10 pack years. Alcohol use was defined as drinking 3 or more alcoholic drinks per day.[161,162] Locoregional recurrence was defined as the location of primary tumour (PT) and/ or lymph node metastases (LN). Locoregional failure was measured from the end of CRT to the date of local or regional histological proven relapse. Metastasis was defined as a distant location from the locoregional PT and LN. Overall survival time was measured from the end of CRT until a HNSCC-related death. These patient outcomes concerned locoregional recurrence, metastasis or death within 2 years of follow-up time or a minimal follow-up time of 2 years after the end of treatment.

Table 1: Patient characteristics.

	Training cohort	Validation cohort
Patients total	103	71
No of male patients	76 (73.8%)	53 (75.7%)
Age, years (IQR)	62.3 (57.3–67.8)	63.3 (57.8–69.3)
Mean radiation dose, Gy	70	70
Chemotherapy		
Cisplatin	88 (85.4%)	57 (80.3%)
Cetuximab	15 (14.6%)	14 (19.7%)
T-stage		
2	46 (44.7%)	25 (35.2%)
3	24 (23.3%)	19 (26.8%)
4	33 (32.0%)	27 (38.0%)
N-stage		
0	14 (13.6%)	11 (15.5%)
1	13 (12.6%)	15 (21.1%)
2	75 (72.8%)	45 (63.4%)
3	1 (1.0%)	0 (0%)
HPV-status		
Positive	39 (37.9%)	26 (36.6%)
Negative	64 (62.1%)	45 (63.4%)
Tumour site		
Oropharynx	74 (71.8%)	51 (71.8%)
Hypopharynx	29 (28.2%)	20 (28.2%)
Overall alcohol history score (SD)	1.91 (1.19)	1.72 (1.24)
Smoking pack years, (IQR)	22.7 (18.2–38.9)	23.5 (19.3–41.3)
Follow-up time, months (IQR)	31.5 (20.7–44.5)	26.4 (19.8–34.1)
Recurrence	27 (26.2%)	19 (27.1%)
Metastasis	10 (9.7%)	18 (25.7%)
Death	37 (35.9%)	22 (31.4%)

IQR: Interquartile range

[^{18}F]FDG-PET/CT acquisition

[^{18}F]FDG-PET/low-dose-CT was performed according to the EANM guidelines 1.0 and since 2015 using version 2.0 on a Gemini-TF or Ingenuity TF PET/CT (Philips Medical Systems, Best, The Netherlands) with EARL accreditation.[163] The examination was performed after a 6-h fasting period and adequate hydration. Scans with arms down were acquired; from mid-thigh to skull vertex, 60 min after intravenous administration of 2.5 MBq/ kg [^{18}F]FDG (3 min per bed position). The [^{18}F]FDG-PET/CT images were reconstructed using time of flight iterative ordered subsets expectation maximization (3 iterations and 21 subsets) with photon attenuation correction using a low dose CT.[164] Reconstructed images of both PET scanners were acquired with similar settings and had an image matrix size of 144 × 144, voxel size of 4 × 4 × 4 mm, FWHM of 6.75 mm. Low-dose-CT was collected using a beam current of 50 mAs at 120 kV for anatomical correlation of [^{18}F]FDG uptake and attenuation correction. CT-scans were

reconstructed using an image matrix size of 512 × 512 resulting in pixel sizes of 1.17 × 1.17 mm and a slice thickness of 5 mm.

Whole-lesion delineation

Whole-lesion delineation was performed, as previously described[165], by an experienced nuclear medicine physician with 5 years of experience (BZ) supervised by another nuclear medicine physician with 30 years of experience (OH) in head and neck nuclear medicine, respectively, with knowledge of the HNSCC diagnosis, TNM-stage (7th edition)[166], and primary tumour location for delineation of proven malignant lesions. Delineation of primary tumours (PT) was performed semi-automatically on [18F]FDG-PET/CT using a 50% isocontour of the SUVpeak of the tumour volume adapted for the local background, providing low variability, low number of outliers, and high repeatability.[167,168] SUV was normalized to body weight. Within the volume of interest (VOI), the maximum and mean SUV were defined (SUVmax and SUVmean). SUVpeak was defined as the uptake in a 1-mL spherical VOI with the highest value across all tumour voxel locations. Partial volume effects were minimized by taking lesion only with a minimum volume of 4.2 mL into account (i.e., 3 times the PET system's spatial resolution of 6.75 mm FWHM).[169]

Feature extraction

Radiomic features were extracted from the [18F]FDG-PET images using the in-house built Accurate tool (for making vois) in combination with the RadCat tool for feature calculation (Supplement 10), as described previously.[170-172] It provides 3D implementation of feature extraction methods for four types of features: shape, intensity, texture based on co-occurrence, and run-length matrices (description of tumour voxels with homogeneous/heterogeneous high or low grey-levels) according to the International biomarker standardization initiative (IBSI) standard.[173] For each patient, 434 [18F]FDG-PET radiomics features were extracted. For the texture analysis, PET images were discretized to a fixed bin size of 0.25 SUV.[171] The radiomic features were not normalized and only raw values were used that were directly computed from the DICOM images. The radiomic data processing consisted of dimension reduction to arrive at a limited number of latent features that retain most of the information contained in the original feature-space (see the next subsection and Supplement 1).

Radiomic data processing

Redundancy filtering

First, the marginal associations between the retained radiomic features of the patient in the retrospective training cohort were assessed in a heat map. As radiomic data are inherently multicollinear, some redundancy was expected: that is, there were pairs of features whose marginal correlation neared (negative) unity. Hence, redundancy

filtering was performed, using a custom redundancy-filtering algorithm.[174] This algorithm removes the minimal number of features under a marginal correlation threshold, which we set at 0.95.

Correlation matrix regularization

The correlation matrix between the remaining features after redundancy filtering was ill-conditioned.[175] The remaining correlation matrix was subjected to ridge regularization.[175] The optimal value of the penalty parameter was determined by 5-fold cross-validation of the log-likelihood. We considered the scaled features (centered around 0 and variance 1) to avoid a situation where the features with the largest scale dominate the analysis.

Factor analytic data compression

Then, we performed a maximum likelihood factor analysis on the regularized feature-correlation matrix.[175] The goal was to reduce the dimension of the data without losing (much) information. When the features naturally clustered into latent factors (meta-features), it was desirable to extract these factors, as it allowed us to build a parsimonious model that retained (as much as possible) the information of the full feature set. A latent radiomic meta-feature represents a projection of the shared information in a collection of observed features. It represents a latent domain underlying a cluster of observables. The dimension of the latent space was determined by Guttman bounds.[176] The factor-solution was rotated to a simple (i.e., sparse) orthogonal structure.

Obtaining factor scores

After projection of the original variable-space onto the lower-dimensional factor-space, we desired factor scores: the score each individual obtains on each of the latent factors. These were obtained by regressing the latent features on the observed data by way of the obtained factor solution. The resulting factor scores of the retrospective training set were used as predictors in further modelling.

Validation

Previously described four steps were then performed separately in the prospective validation cohort in order to validate similar radiomic factors in the prediction analysis.

Statistical analysis

The correlation between clinical parameters, standard [^{18}F]FDG-PET/CT parameters (SUVmax, SUVmean, SUVpeak), and radiomic factors was determined in the training and validation set with Spearman's correlation coefficient. Corresponding p values were multiplicity corrected using Bonferroni's method. The difference in outcome was assessed between patients who received cisplatin and cetuximab (log rank test). The

difference in outcome was assessed for patients with an oropharyngeal and hypopharyngeal tumour location between HPV-positive and HPV-negative status (log rank test).

The prognostic performance of clinical parameters, [^{18}F]FDG-PET/CT parameters, and radiomic factors was firstly assessed in the training set separately for the patient outcomes (locoregional recurrence, distant metastases, and death) by performing a Cox regression analysis. Thereafter, significant clinical, [^{18}F]FDG-PET/CT parameters, and radiomic factors were combined in a multivariable analysis. Multivariable regression analysis was performed according to the TRIPOD-statement (Supplement 9), accepting p values up to 0.157 to enhance the model applicability to other patient groups.[177,178] Predictive performance of the models was assessed by a 5-fold cross-validation[179] and by using the incident area under the receiver operating curves (ROC) and concordance index (CI).

The predictive accuracy of the constructed prediction models in the training set was validated in a separate validation set. The prognostic performance was assessed by the incident area under the receiver operating curves (ROC) and concordance index (CI). Finally, the prediction models were compared in the validation set using the loglikelihood chi-square test and area under the curve (AUC).

A risk calculator for all outcomes was constructed, based on the normalized standard hazard and the coefficient of each parameter or radiomic factor of the predictive model. This risk stratification was divided into a high (≥ 66%), medium (≥ 33–66%), and low risk (< 33%) for a patient outcome using the most accurate prediction model. The correlation assessment was performed on IBM SPSS Statistics for Windows. Analyses regarding the factor-analytical data-compression and prognostic modelling were performed with R.

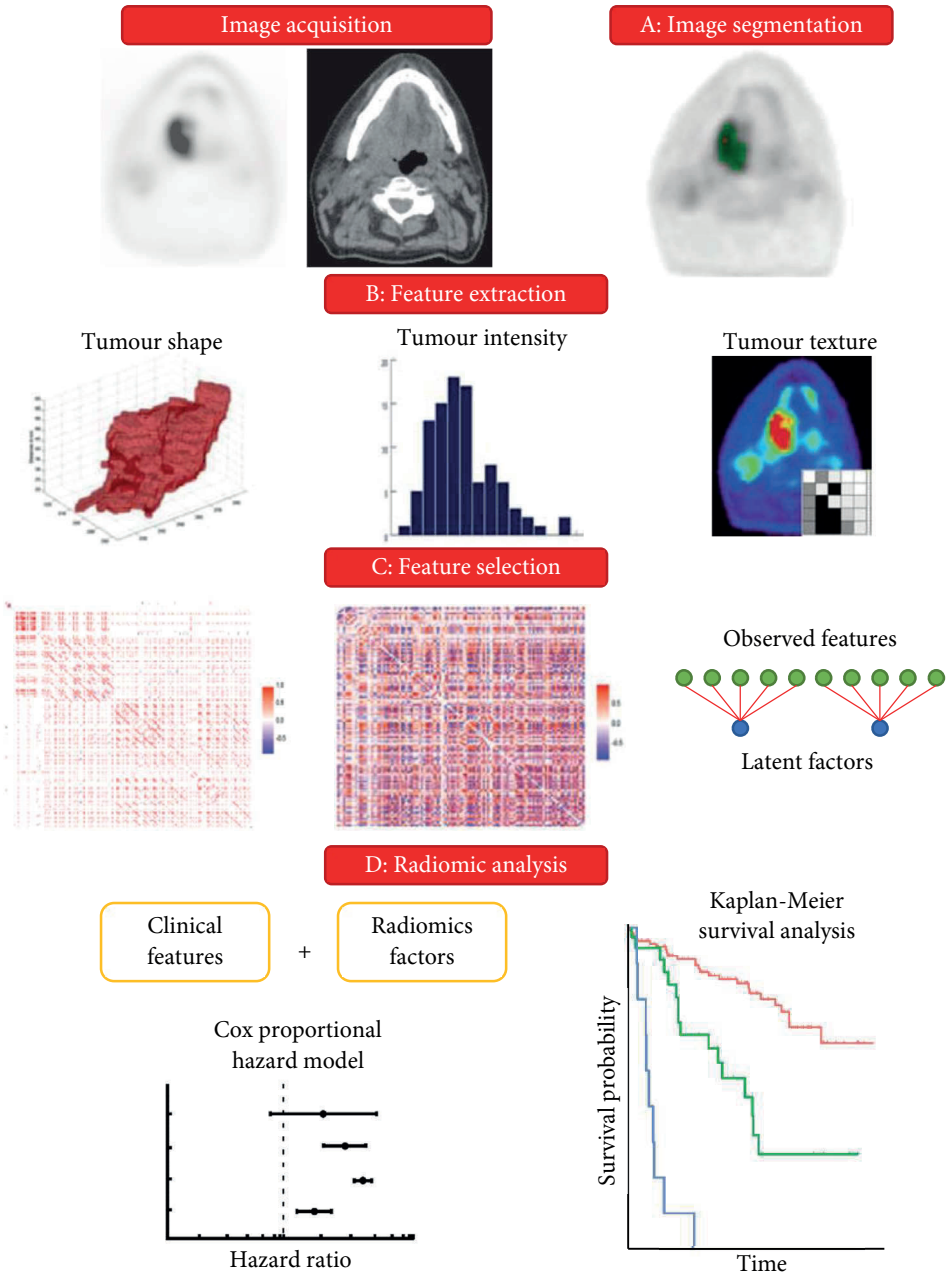


Figure 1: An overview of the radiomics workflow. A, delineation; B, extraction of intensity, texture, morphologic, and shape radiomics features; C, removal of redundancy of highly correlated features (Pearson r > 0.95) and the construction of factors; D, construction of prediction models with clinical, first-order PET-features, and/or radiomic factors and the risk-stratification into a high/medium/low risk for developing an event based on the constructed prediction models.

Results

Patient characteristics

Overall, 184 patients were included, of which 103 retrospectively (training set) and 71 consecutive independent patients (validation set) (see Table 1 for patient characteristics). The mean age of the training cohort was 62.3 years (inter-quartile range (IQR): 57.3–67.8). The mean age of the validation cohort was 63.3 (IQR 57.8–69.3). Treatment of all included patients consisted of pre-determined regimens: in 88 patients, radiotherapy was combined with a cisplatin dose, 15 patients received radiotherapy with cetuximab. The mean follow-up time in the training set was 31.5 months (IQR: 20.7-44.5) and in the validation set 26.4 months (IQR 19.8–34.1). In the training cohort, 27 recurrences, 10 metastases, and 37 deaths occurred. In the validation cohort, 19 recurrences, 18 metastases, and 22 deaths occurred. The outcome was not significantly different between patients who received cisplatin and those who received cetuximab in the training set and test set; for recurrence ($p = 0.071$, $p = 0.877$, respectively), metastasis ($p = 0.60$, $p = 0.295$, respectively), and OS ($p = 0.053$, $p = 0.276$, respectively). The median OS in the training set for patients with cisplatin 32.1 months and for cetuximab 27.6 months and in the validation set for cisplatin 23.2 months and for cetuximab 18.1 months. A significantly better OS was found for HPV-positive cancers with both oropharyngeal and hypopharyngeal primary tumour location (both $p < 0.05$).

Radiomic factors

Redundancy filtering showed many strong (absolute) associations, which was echoed in the heatmap on the thresholded correlation matrix (Figure 1C), including all correlations whose absolute value equals or exceeds 0.95. After redundancy thresholding, 124 radiomic features were retained (Figure 1D). The remaining correlation matrix was subjected to ridge-regularization with the optimal regularization parameter value determined by 5- fold cross-validation of the log-likelihood. The resulting regularized matrix was well-conditioned.

The factor analytic data compression of the regularized correlation matrix resulted in eight latent meta features (factors). These retained 80% of the covariation between the original 124 features. Hence, the factor solution was deemed to sufficiently represent the original feature-space (Supplement 1).

Representation of original features in the radiomic factors

Factor 1 consisted mainly of (I) least axis length (morphology) and (II) non-uniformity (GLRLM; grey-level-run-length matrix and GLDZM; grey-level-distance zone-matrix (counts the number of groups of linked voxels, which share a specific discretized grey-level and possess the same distance to ROI edge), and (III) high dependence of high grey levels (NGLDM; neighbourhood grey-level difference matrix, which aims to capture the coarseness of the overall texture[172]).

Factor 2 consisted mainly of (I) histogram range (intensity), (II) (A) contrast, dissimilarity, cluster prominence (GLCM; grey-level-co-occurrence matrix), (B) zone size non-uniformity (GLSZM; grey-level-size-zone matrix) (C) complexity, contrast, and strength (NTGDM; neighbourhood-grey-tone-difference matrices), and (D) small distance high grey level emphasis (GLDZM).

Factor 3 consisted mainly of (I) maximum histogram gradient and inversed minimum histogram gradient (Intensity), (II) (A) long run low grey-level emphasis and run-length variance (GLRLM), (B) zone size variance (GLSZM) (C) busyness (NGTDM), and (D) high dependence emphasis and dependence count variance (NGLDM).

Factor 4 consisted mainly of (I) volume difference (intensity), (II) (A) inversed 3D coarseness, grey-level nonuniformity, large distance low grey-level (NGTDM), and (B) inversed low grey-level count and energy count (NGLDM).

Factor 5 consisted mainly of (I) asphericity, major axis length, inversed compactness, and flatness (morphology).

Factor 6 consisted mainly of (I) histogram coefficient of variation (intensity) (II) second measure of information correlation (GLCM) and (III) Morans I (Morphology).

Factor 7 consisted mainly of (I) inversed small zone low grey-level emphasis (GLSZM).

Factor 8 consisted mainly of inversed difference features (GLCM), but scored lower than the overlapping factor 1 features.

Associations between clinical and [^{18}F]FDG-PET parameters with radiomic factors

The significant associations after Bonferroni's correction of each of the 8 factors with T-stage, N-stage, HPV-status, and smoking in the training set (Table 2) showed that factor 1 had a significant positive correlation with T-stage (r = 0.454), SUVmax (r = 0.440), SUVpeak (r = 0.521), SUVmean (r = 0.468), TLG (r = 0.807), and MATV (r = 0.947). Factor 2 correlated significantly with SUVmax, SUVpeak, and SUVmean (r = 0.704–0.740). Furthermore, T-stage correlated significantly with SUVmax (r = 0.412),

SUVpeak (r = 0.438), SUVmean (r = 0.422), and MATV (r = 0.405). HPV-status correlated negatively with SUVmean (r = – 0.338). In the validation set, associations between factor 1 and TLG and MATV (r = 0.812, 0.887), factor 2 and SUVmax, SUVpeak and TLG (r = 0.838–0.876), and factor 3 and TLG and MATV (r = 0.494, 0.815, respectively) remained significant (Supplement 2). Low association was found between factors (Supplement 3).

Table 2: Correlations of radiomic factors with clinical parameters and FDG-PET parameters in the training set.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	SUV-max	SUV-peak	SUV-mean	TLG	MATV
T-stage	**0.45**	0.23	-0.10	0.11	0.11	-0.10	0.12	0.04	**0.41**	**0.44**	**0.42**	0.32	**0.41**
p-value	0.00	0.02	0.31	0.27	0.26	0.31	0.21	0.73	0.00	0.00	0.00	0.00	0.00
N-stage	0.08	-0.08	0.09	0.14	0.07	0.04	-0.05	0.05	0.06	0.05	0.04	0.06	0.07
p-value	0.45	0.41	0.39	0.15	0.49	0.67	0.60	0.64	0.57	0.65	0.68	0.54	0.52
HPV	-0.26	-0.27	0.13	0.01	-0.15	0.06	0.01	0.04	-0.33	-0.33	**-0.34**	-0.20	-0.25
p-value	0.01	0.01	0.21	0.95	0.13	0.58	0.93	0.72	0.00	0.00	0.00	0.05	0.01
Smoking	0.02	0.05	-0.12	0.08	0.27	0.01	-0.04	0.08	0.14	0.10	0.10	-0.03	0.03
p-value	0.84	0.61	0.23	0.44	0.01	0.89	0.70	0.44	0.16	0.34	0.32	0.75	0.77
SUV-max	**0.44**	**0.72**	-0.09	**0.31**	0.08	0.04	0.17	0.17					
p-value	0.00	0.00	0.35	0.00	0.43	0.66	0.10	0.09					
SUV-peak	**0.52**	**0.70**	-0.03	0.28	0.04	0.02	0.15	0.18					
p-value	0.00	0.00	0.73	0.00	0.68	0.84	0.14	0.07					
SUV-mean	**0.47**	**0.74**	-0.07	0.29	0.01	-0.02	0.15	0.16					
p-value	0.00	0.00	0.46	0.00	0.92	0.87	0.13	0.11					
TLG	**0.81**	0.17	**0.40**	0.08	0.04	0.01	0.07	-0.11					
p-value	0.00	0.08	0.00	0.43	0.69	0.92	0.48	0.25					
MATV	**0.95**	0.02	0.03	0.04	0.10	0.00	0.02	-0.23					
p-value	0.00	0.82	0.73	0.66	0.30	1.00	0.82	0.02					

Bold numbers were significantly correlated ($p < 0.001$), after the Bonferroni multiple testing correction.

Prognostic value of clinical, [18F]FDG-PET parameters, and radiomic factors in the training set

The significant predictors of recurrence were in the training set per clinical, PET parameter of radiomic factors separately; HPV-status; MATV; and factors 1 and 4 (Supplement 4).

The combination of clinical and [18F]FDG-PET parameters resulted in N-stage, HPV-status; and SUVmean as significant predictors (Supplement 5). The combination of clinical and radiomics parameters resulted in HPV-status; and factors 1, 4, 5 as significant predictors. The combination of clinical, [18F]FDG-PET, and radiomics parameters resulted in HPV-status, SUVmean, SUVpeak, factor 3, 4, and 6 as significant predictors (Supplement 4) and was significantly (p = 0.041; Supplement 5) most accurate to predict recurrences (CI = 0.796, SE = 0.045) as compared with other combinations (Table 3).

The significant predictors for distant metastasis were in the training set per clinical, PET parameter of radiomic factors separately; only MATV (Supplement 3).

The combination of clinical and [^{18}F]FDG-PET parameters resulted in N-stage and SUVmean as significant predictors (Supplement 4). The combination of clinical parameters, [^{18}F]FDG-PET parameters, and radiomics resulted in only MATV as significant predictor (Supplement 4).

The significant predictors for overall survival were in the training set per clinical, PET parameter of radiomic factors separately; T-stage, HPV-status; MATV; factors 1 and 5 (Supplement 4).

The combination of clinical and [^{18}F]FDG-PET parameters resulted in HPV-status and MATV as significant predictors (Supplement 4). The combination of clinical parameters and radiomics resulted in factors 1 and 5 as significant predictors.

The combination of clinical parameters, [^{18}F]FDG-PET parameters, and radiomics resulted in HPV-status, SUVmax, SUVmean, factors 1 and 5 as significant predictors (Supplement 5) and was non-significantly ($p > 0.05$; Supplement 6) most predictive (CI = 0.750, SE = 0.046) as compared with other combinations (Table 3).

Table 3: Predictive accuracy of clinical parameters, PET-parameters, and radiomics factors separately and combined for the prediction of locoregional recurrence, metastasis, and death.

n=103	Recurrence (27) CI (SE)	Metastasis (10) CI (SE)	Death (37) CI (SE)
Clinical parameters T-stage, N-stage, HPV-status, Smoking	0.70 (0.05)	0.69 (0.10)	0.69 (0.04)
PET parameters SUVmax, SUVmean, SUVpeak, TLG, MATV	0.62 (0.07)	0.76 (0.06)	0.71 (0.04)
Radiomics parameters Factor 1 to 8	0.72 (0.06)	0.75 (0.08)	0.71 (0.05)
Combined clinical + PET	0.76 (0.05)	0.82 (0.05)	0.74 (0.04)
Combined clinical + radiomics	0.77 (0.04)	0.83 (0.07)	0.75 (0.05)
Combined clinical + PET + radiomics	0.80 (0.05)	0.95 (0.03)	0.75 (0.05)

CI: concordance index. SE: standard error.

Validation of the prognostic models

In the validation set, the prognostic accuracy of each trained model predicting the risk for recurrence, metastasis, and overall survival was validated (Table 4). This resulted in a validated CI = 0.645 (SE = 0.071) for recurrence, CI = 0.627 (SE = 0.094) for metastasis, and CI = 0.764 (SE = 0.062) for overall survival (Table 4 and Figure 3).

The risk stratification into a high, medium, and low risk for adverse outcome was constructed; for recurrence (p = 7E−5), metastasis (p = 0.002) and overall survival (p = 4E−7) (Figure 2, Supplement 7 and 8). A clinical applicable patient-specific risk calculator was constructed for a single patient to predict recurrence, metastasis, or death (Table 5).

Table 4: The accuracy of the prediction models for recurrence, metastasis, and overall survival in the training set and validated in the validation set. For the recurrence prediction, the combination of HPV, SUVmean, SUVpeak, factors 3, 4, and 6 was most accurate. For the metastasis prediction, the use of only MATV was most accurate. For overall survival prediction, the combination of HPV, SUVmax, SUVmean, factors 1 and 5 was most accurate.

Final prediction models	Training set (n=103)		Validation set (n=71)	
	Events	CI (SE)	Events	CI (SE)
Recurrence prediction HPV, SUVmean, SUVpeak, factor 3, factor 4, factor 6	27	0.78 (0.05)	19	0.65 (0.07)
Metastasis prediction MATV	10	0.66 (0.09)	18	0.63 (0.09)
Overall survival prediction HPV, SUVmax, SUVmean, factor 1, factor 5	37	0.75 (0.05)	22	0.76 (0.05)

Events: number of recurrences in the recurrence prediction model; number of distant metastases in the metastasis prediction model; number of deaths in the overall survival prediction model.
CI: concordance index. SE: standard error.

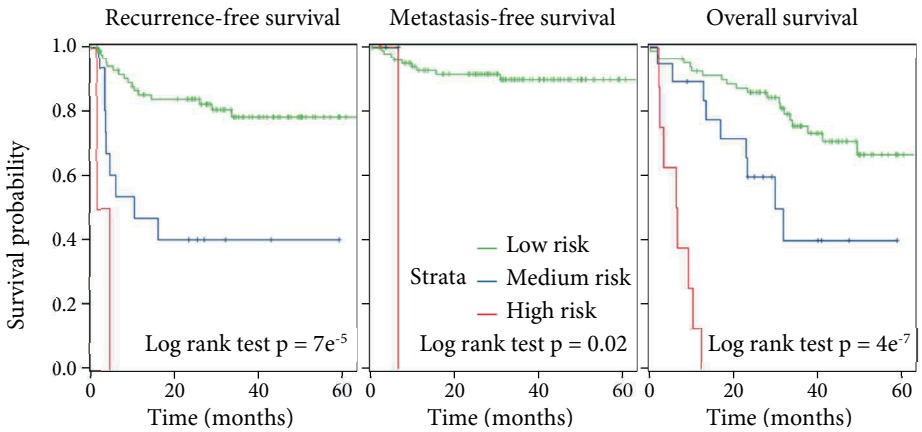


Figure 2: Accuracy of the combined prediction of locoregional recurrence (left), metastasis (middle), and overall survival (right) in the validation cohort. The curve of the relatively small medium risk group for metastasis is short; this is due to the short follow-up time until the metastasis occurred. A significant predictive risk stratification ($p < 0.05$) was shown, divided in low (0–33%), medium (33–66%), and high (66–100%) risk for an unfavourable prognosis.

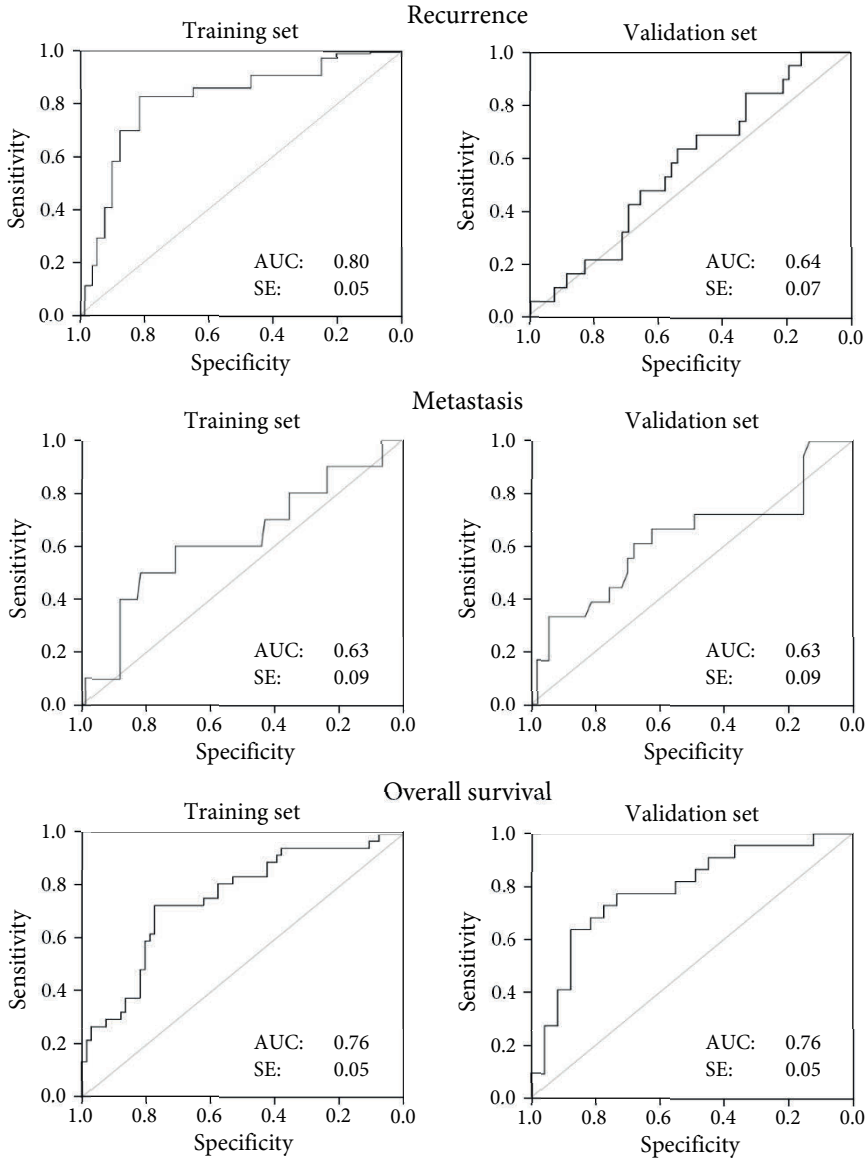


Figure 3: ROC curves in the training and validation set per patient outcome prediction. AUC: area under the incident receiver operating characteristic curve (ROC) for each final model in the training set as well as in the validation set for the prediction of recurrence, metastasis, and death within 2 years of follow-up after end of treatment. SE: standard error.

Table 5: Risk calculators

Recurrence				Metastasis				Death			
Predictor			Result	Predictor			Result	Predictor			Result
HPV*	0	× -1.45 =	0.00	MATV	51.7	× .064 =	3.33	HPV*	0	× -0.98 =	0.00
SUVmean	6.74	× -2.48 =	-16.76					SUVmax	8.38	× -0.58 =	-4.89
SUVpeak	8.55	× 1.89 =	16.15					SUVmean	4.85	× 0.95 =	3.64
Factor 3	-1.29	× -1.27 =	1.65					Factor 1	-0.81	× 0.52 =	-0.43
Factor 4	-0.26	× -0.36 =	0.09					Factor 5	1.23	× 0.54 =	0.67
Factor 6	-0.55	× -0.60 =	0.33								
		Sum	1.46			Sum	3.33			Sum	-1.01
		$1 - e^{-0.13e^{Sum+1.39}}$ =	0.90			$1 - e^{-0.10e^{Sum-0.66}}$ =	0.75			$1 - e^{-0.21e^{Sum+0.26}}$ =	0.09
		Risk	**90.1%**			**Risk**	**75.2%**			**Risk**	**9.5%**

The risk calculators can be used in clinical practice to calculate the risk per specific patient for locoregional recurrence, metastasis or death during the follow-up time of 2 years. The yellow boxes could be filled-in with the single patient data in order to calculate the risks.
* HPV-status: 0 = negative, 1 = positive

Discussion

In this study, the examination of the prognostic value of pre-treatment [18F]FDG-PET radiomics in locally advanced HNSCC showed that the discriminatory performance of the combination of latent radiomics factors of [18F]FDG-PET was of additional value in predicting recurrence, metastasis, and overall survival and that the combination of clinical, PET, and radiomics parameters was most predictive.

Radiomics process

The primary goal of radiomics is to build clinical models using machine learning techniques[180] in order to predict patient outcome, thereby allowing for better personalized treatment management. These multivariable prediction models might be unintelligible for clinicians, because they combine a large number of high-order multimodality image features.[181,182] However, they may outperform visual analysis in terms of accuracy.

Aerts et al. selected only the single best predictive features on CT from each of their four main feature categories (statistical features (e.g., mean, maximum, peak, mode), shape, grey-level-non-uniformity, and wavelet grey-level-non-uniformity HLH (i.e., describing intratumoral heterogeneity after decomposing the image in mid-frequencies).[159] Bogowicz et al. reported that performing PET, the combination of principle component analysis (PCA; a statistical procedure that converts a large set of observations of possibly correlated variables into a smaller projection of the most informative linearly uncorrelated variables) and univariate feature selection using the Cox regression with backward selection, resulted in the least complicated model with best discriminative power.[154] However, their final PET model consisted of only 2 single radiomic features,

and no clinical variables were considered. Vallières et al. trained predictive models for each radiomic feature combined with clinical variables and patient outcome by performing random forests and made adjustments to model imbalance.[145] Finally, only one PET-radiomics (GLNGLSZM) and two CT-radiomics features were included in the model. These methods manually excluded all other possible prognostic features.

In this study, a dimension reduction was performed of the feature space by removing redundant features (retaining 124 features). Based on these features, a factor analysis was performed, which consisted of a feature subset (i.e., factor) and contains a part of the predictive feature spectrum on a scale of importance. This allowed the preservation of the multiple predictive features and assess possible interactions or associations. This might provide insight in the underlying concepts of the heterogeneous whole-lesion PET data, as a basis for identification and targeting tumoral subvolumes which are predictive for adverse outcome.[183] Moreover, this factor analysis was done separately from the patient outcome, which might allow for the improvement of the tumour specific classification, as basis for prognosis prediction. However, in other studies which selected single features, this inter-correlation of feature was lost.[154,159] Thirdly, it overcomes the risk of data overfitting, which arises when the number of features is large and the number of training data is comparatively small.[184]

Tumour characteristics by radiomic factors

The spectrum of known predictive clinical and first order PET parameters might be extended with noncorrelated PET-radiomic features we found in this study, capturing complementary characteristics of the complex heterogeneous tumoral microenvironment.

Low values of factor 3, 4, and 6 were predictive of recurrence, complementary to negative HPV-status, low SUVmean, and high SUVpeak. Factor 3 correlated in the validation set with MATV and measured mainly maximum histogram gradient and long low grey-level lengths with a variance of lengths and zones, and high busyness, which might indicate tumoral intensity heterogeneity in tumoral zones of varying size, with long rows of low grey-level voxels (i.e., low [^{18}F]FDG uptake). These features might capture the presence of necrotic regions within the core of tumours. Previously, this correlation between heterogeneity and volume in PET-data was reported by Hatt et al.[157] Also Cheng et al. found that besides TLG, uniformity (local scale texture parameter) and zone-size non-uniformity (ZSNU) were usable as prognostic stratifiers.[185] This was confirmed by Vallières et al., who also reported that GLSZMGLN (grey-level size zone matrix with grey-level non-uniformity) was predictive for locoregional recurrence.[145] Also Bogowicz et al. found that GLSZMZSLGE (grey-level size zone matrix; with zone size low grey-level emphasis) was predictive for favorable prognosis (CI 0.71).[154] However, in their study, different scanners were used between

training and validation cohorts, which reduced data quality. Factor 4 measured slightly different characteristics such as intensity differences with high grey-level counts (inversed low grey-level count) and grey-level non-uniformity (inversed coarseness). This factor might capture the heterogeneity of tumoral sub-areas with a mainly high [18F]FDG-tracer uptake. Factor 6 measured the histogram variety of intensity and quantifies the complexity of the texture (second measure of information correlation), which might capture the tumoral range of [18F]FDG-uptake and differences of uptake between sub-areas. These radiomics features, bundled in factors, were not previously described in literature and might provide insights in the extent of tumoral clonal heterogeneity and interactions, which might help us to control tumours.[143]

For *distant metastasis* prediction, we found in this study the use of MATV only was most accurate and outperformed all other clinical and radiomic parameters. This was partly confirmed by Vallières et al., who also found tumoral volume, as well as age, tumour type, and N-stage as well as CT-radiomic heterogeneity features as predictive parameter.[145] The large metabolic active tumour volume might enable large numbers of cell divisions, tumour progression into genetic instability, which might lead to metastatic ability.[143]

High values of factors 1 and 5 were most predictive of adverse overall survival, complementary to negative HPV-status, SUVmax, and SUVmean. Factor 1 correlated significantly with T-stage and all PET parameters, with the highest correlation of those which were volume-related. This was in line with Vallières et al. [8], who found that volume outperformed each radiomic models. However, factor 1 consisted also of mainly morphologic and non-uniformity texture features and was dependent on high intensity, which might correlate with large heterogeneous tumoral entities. This factor might capture the voluminous extent of the tumour, combined with areas of high [18F]FDG-tracer uptake. El Naqa et al. also reported that intensity histogram and shape features were predictive of survival.[160] Factor 5 measured also morphological tumour characteristics, such as asphericity, major axis length, and inversed compactness and inversed flatness. This was found complementary to the volume-related features in factor 1, and in line with Bogowicz et al., who reported that besides GLSZMZSLGE, sphericity was most predictive for favourable prognosis (CI = 0.71).[154] Also, Aerts et al. reported similar results in CT-data, showing that patients with more compact/ spherical tumours had better survival probability.[159] Factors 1 and 2 both correlated with PET parameters and reflected particular heterogeneous distribution of [18F]FDG-PET uptake. Factor 1 correlated with volume-related TLG and MATV in the validation set. Factor 2 measured the histogram range, contrast, and small high grey emphasis, and correlated with SUVmax, SUVpeak, and SUVmean, and did not remain predictive.

Discriminative power of prediction models

In order to improve predictive accuracy, patient-specific tumoral characteristics were captured by radiomics features and such as low grey-level zone sizes, heterogeneous busyness and morphologic tumour volume, and bundled by factors. Prediction models including these factors are hypothesized to be more patient-specific, because of more unique characteristics, than models which do not investigate underlying feature correlations and include only the single most predictive feature. Vallières et al. combined clinical parameters, without HPV-status, with only one PET- and CT-radiomic feature; however, the prediction accuracy was similar for locoregional recurrences (AUC = 0.69) and overall survival (CI = 0.74).[145] Aerts et al. used the top 4 performing CT-features of each radiomics feature category, where inclusion of TNM-stage improved performance and showed a survival prediction of CI = 0.69.[159] Bogowicz et al. reported a CI of 0.71 using PET-radiomics; however, data was influenced by artifacts, scanner, and protocol heterogeneity.[154] Also, current study showed that for metastasis prediction, the use of only MATV was most accurate. The accuracy of the prediction model combining all clinical (T-stage), first-order PET (SUVmean), and radiomic factors was found to be higher than the final model, consisting of only MATV. This might be due to the fact that the other features still hold some predictive power. Although this might provide insights in metastatic tumour characteristics, it should be validated in future studies. This was partly in line with Vallières et al., who also found volume-parameter was most predictive, but they found additional value for CT-radiomics features.[145]

Clinical applicability

The efficacy of a treatment plan, nowadays based on information from clinical examination (under anaesthesia), visual interpretation of imaging, and invasive biopsies, could be optimized by taking the patient-specific pathophysiologic phenotype into account[186] using quantitative imaging assessment. The underlying tumour biology could be heterogeneous with different sub-clonal populations, continuously changing and associated with resistance to treatment, recurrence, and overall survival.[145,159] Many studies[145,154,159,160] constructed predictive models based on the selection of a few radiomic features excluding clinical parameters (e.g., HPV status) and interactions with radiomic features, in order to reduce the risk for overfitting.[145,154,159]

In this study, we showed an advanced factor analysis using three-dimensional whole-lesion radiomic features as well as retaining feature interactions captured in radiomic factors. These complementary factors improved predictive accuracy to the basis of clinical factors, including HPV-status and first-order PET parameters, and remained accurate after validation. Although we found a correlation between MATV and T-stage (mainly based on tumour volume), volume-related parameters were more predictive.

Furthermore, we presented a patient-specific clinical-applicable risk stratification for patients with head and neck cancer treated with (chemo)radiotherapy. Low-risk patients could be candidates for treatment de-escalation studies[187,188], whereas high-risk patients could benefit from treatment escalation[189], immunotherapy[190], or surgical treatment. This optimization of treatment efficacy might also result in a beneficial reduction of costs. Identification and validation of optimal machine-learning methods for radiomic applications using standardized EANM guidelines[163] is crucial towards reproducible biomarkers in clinical practice, complementary to the clinical and first-order PET parameters.

Limitations

At the assessment of multiple clinical, first-order, and radiomic features, there is a risk for overfitting bias. In the current study, we used a relatively large patient sample size and performed a multicollinearity filtering to exclude highly correlated features. Moreover, the factor analysis projects the large and collinear radiomic feature-space onto an orthogonal latent-feature-space of smaller dimension (8 factors) while retaining the bulk of the information contained in the full data. This projection is thus geared towards the avoidance of overfitting. Finally, a limited amount of clinical, first-order PET and PET-radiomic factors was combined in a multivariable model. However, it is still possible that the number of events was not enough to construct a statistically robust prediction model. In this study, validation was performed internally by 5-fold cross-validation of the prognostic models. Moreover, we used an independent validation cohort of similar institute to estimate the performance of a prediction model. In Table 4 and Figure 3, we present the results obtained for the training set as well as the independent validation set. We can see that for the recurrence prediction model, the concordance index for the independent validation set is somewhat lower, while for the other 2 models, a similar performance was found between the training and (independent) validation dataset. However, in future studies, validation in a larger cohort from an external institute is still needed.

The prognostic model performance might be optimized by a stricter redundancy filtering to retain only complementary factors; however, in this study, we saved the inclusion of possible predictive underlying relationships of features. This model should be constructed using a limited number of factors separate from patient's outcome, in order to solely include predictive tumoral processes and to minimize cohort-dependent prognostic influences. Another improvement of the prognostic model performance might be the implementation of complementary predictive CT-radiomic features[159,191,192], which would require similar acquisition parameters, artifacts reduction techniques, and a larger patient population to overcome the risk of overfitting and should be evaluated in future studies.

This study was hypothesis generating and the feasibility was tested. However, in the next step to clinical translation, more extensive validation and refinement on larger and external datasets as well as evaluation of the clinical applicable calculators, is needed. Moreover, it is of interest to perform further technical validation, such as by the use of voxel randomization.[193,194] Our study suggests that adding radiomics to the [^{18}F]FDG-PET image analysis can improve prognostication as a step towards personalized treatment of HNSCC patients.

Conclusion

The combination of HPV-status, first-order [^{18}F]FDG-PET parameters, and complementary radiomic phenotype specific factors improved time-to-event prediction most accurately. Predictive tumour-specific characteristics and interactions might be captured and retained using radiomic factors, which allows for personalized risk stratification and optimizing personalized cancer care.

Supplementary material

Supplementary material for this paper is available at:

doi.org/10. 1186/s13550-020-00686-2

8

Discussion

Biomarkers are the measures used to perform clinical assessment and therefore need to be validated before they can be used to make well-founded medical decisions. The studies included in this thesis have mainly addressed some technical aspects of imaging biomarker validation and investigated simplifications that would facilitate clinical use. This chapter provides a summarizing discussion of the implications and describes the future perspectives.

Technical validation and method simplification are two important elements on the biomarker roadmap towards clinical translation.[195] Technical validation looks into the accuracy and precision of measurement of a potential biomarker. Parallel tracks on the roadmap look into biological validation (whether the biomarker is actually related to underlying biology) and clinical validation (whether the biomarker is clinically useful in, for example, diagnosis, prognosis or treatment response prediction). Method simplification is an important aspect for translation into clinical practice, because complicated methods that can be studied in research may not be feasible in clinical practice. This simplification must always be balanced against validity—in our case, technical validity—of the biomarker.

The study in chapter 2 studied the possibility of omitting the invasive arterial input function measurement to simplify the clinical protocol for estimating brain perfusion with $[^{15}O]H_2O$ PET. In earlier research it has already been indicated that absolute quantification of CBF using $[^{15}O]H_2O$ PET without an arterial input function is unlikely to be accurate. This proved to be especially true for the short bolus imaging protocol that we used in our study. Nevertheless, one of the methods (from Treyer et al.) estimated perfusion with reasonable precision, which could allow for assessment of short-term longitudinal changes. Obtaining accurate *relative* perfusion estimates—relative to the global brain perfusion—is possible, as the study confirmed. The double integration method proved to be best method for measuring relative perfusion in terms of both accuracy and repeatability. Our study also included several simplified parameters that are derived directly from the time activity curve. These simplified parameters are popular in perfusion research using other techniques, for example in MR perfusion imaging. The problem with these simplified parameters is that they are often derived from a small part of the time activity curve. Therefore, they are sensitive to noise. This problem occurs especially in PET, where the level of noise for a signal from a single voxel is quite high. However, as the study's simulations show, even without noise, many of these parameters from the time-activity curve tend to be biased or nonlinearly related to the parameter of interest: perfusion. Perhaps unsurprisingly, the area under the curve is one of the better parameters that can still cross the first translational gap and can be used as a reliable biomarker in clinical research, such as in perfusion ultrasound.[196] As a result of the integration, the parameter is relatively insensitive to noise. However, the parameter is still

nonlinearly related to perfusion. The double integration method removes this nonlinearity and is still equally insensitive to noise. The method shows that it is possible to develop biomarkers that are easily obtained and still provide robust and biologically relevant estimations.

Tracer kinetic modelling is a fundamental step in validation of PET tracer biomarkers. The chapters 3 and 4 address the validation of simplified methods through kinetic modelling of the tracer FET in glioma. First, in chapter 3, the optimal plasma input kinetic model was identified (the reversible two-tissue compartment model with fitted blood volume fraction). Although the size of our dataset is small, the data indicates that the model preference is independent of tumour grade. Based on [^{15}O]H_2O PET CBF measurements we concluded that extraction fractions were highly variable between patients, which could be caused by differences in transporter expression and/or blood brain barrier breakdown. This disabled the investigation of perfusion dependence of the simplified methods. The model was therefore only used to assess agreement and correlation between full kinetic parameters and simplified methods. The results show that a possible downside of early static imaging might be that variability in uptake time will lead to variability in SUV. Better correlation was found at later uptake time intervals (60-90 min post injection). Reference tissue input compartment models did not correspond well with plasma input derived distribution volume ratios (DVR). This may be possibly explained by the violation of the reference tissue model assumptions, i.e., the measured signals are not represented by a single-tissue compartment model and the ratio K_1 over k_2 is not constant between the reference and target tissue. Consequently, use of reference tissue input models may not be valid for dynamic [^{18}F]FET brain studies. The SUV-ratio, however, showed slightly better results. The results of the study suggest that for both SUV and SUV-ratio later time point imaging (60-90 min post injection) outperforms currently recommended uptake time (20-40 min post injection) in terms of correspondence with the kinetic model.

In chapter 4, we evaluated the performance of several methods of parametric mapping for the analysis of dynamic brain ^{18}F-FET PET studies. Parametric maps provide parameter values for each voxel, and can therefore be used to assess the location of tumour boundaries or assessing tracer uptake distribution within the tumour. The results indicate that Logan graphical analysis is best suited for deriving the volume of distribution (V_T) and that $SUVr^{60-90}$ (tumour-to-normal maps at interval 60–90 min) is a good substitute for the distribution volume ratio (DVR). In line with chapter 3, SUV-ratio images at a later interval provided better quantitative performance; however, their use to estimate the tumour extent may prove problematic in some glioblastoma patients: the contrast between tumour and surrounding tissue decreases over time, which can make it harder to see and locate the boundaries of the tumour.

In our study in chapter 5, we investigated the measurement of AIFs in DCE-MR images from head and neck cancer patients and found that we were not able to accurately estimate arterial concentrations from the images. Concentration was measured in several arteries in the neck and the concentration differences between the arteries, and the difference with concentrations from literature, all point to this conclusion. This inaccuracy is likely due to blood flow and partial volume and T_2^* effects. Additionally, the accuracy is further decreased by the nonlinearity of the MR signal to concentration. This nonlinearity is more pronounced at higher concentrations, which makes it hard to estimate high concentrations. The high concentration peak is an important part of the arterial input function. The poor accuracy of AIF measurement is probably the reason why the repeatability of estimated kinetic parameters does not improve when an image-derived AIF is used instead of a population averaged AIF. Therefore, the use of the latter is recommended. However, by definition, the population averaged AIF does not correct for variability in cardiac output between patients. The image-derived AIF does not provide an improvement in this regard and can therefore be omitted.

The intravoxel incoherent motion (IVIM) model estimates diffusion and perfusion related parameters from diffusion-weighted imaging. These biomarkers can be useful for disease management in head and neck cancer. Chapter 6 describes the head-to-head comparison of IVIM fitting methods in terms of test-retest repeatability. The investigated methods were nonlinear least-squares regression, Bayesian probability estimation, and two implementations of a neural network. Use of a neural network is appealing for clinical practice because its application is very quick (calculation time is in the order of seconds). Clinical translation of the neural network approach is, however, still hampered by the lack of consistency because even a repeated training on the same dataset already yields different results. If the model is retrained, for example by another medical centre, the precision of the new model is not the same as the old model. A challenging task for further research is therefore to develop a way to make model training more robust.

In chapter 7 we investigated the clinical value of [^{18}F]FDG PET radiomics for patients with locally advanced head and neck squamous cell carcinoma. These patients undergo pre-treatment [^{18}F]FDG PET imaging which can be useful for detection, treatment decision support[151], planning[152,153], and the prediction and detection of recurrences and long-term outcome[139]. Lesions can be characterized further using radiomics features. Our study shows that analysis of these features in addition to clinical and first-order [^{18}F]FDG PET parameters improves the prediction of recurrence, metastasis, and overall survival.

A distinct step in our methodology, compared to other radiomics studies, is the dimension reduction, which consists of first removing the redundant (highly correlated) features and then combining the remaining features into a smaller number of factors. The factors are still subjected to parameter selection in the final model definition, further

reducing the number of parameters to avoid overfitting the data. The idea is that combining the features into factors better retains the information than when parameter selection is performed on the feature level directly. Our results indicate that this approach is indeed feasible and provides additional value to clinical information and first order image parameters. However, a downside to this factor conversion is that the factors themselves are hard to interpret. Because they form a combination of radiomic features it is difficult to intuitively relate them to the disease. This would make it hard for a clinician to apply in routine care, because there is no clear understanding of why a certain outcome is predicted.

Future perspectives

Development of quantification methods for imaging biomarkers is a scientific process of technical, biological and clinical evaluation. The continuous development is an indication of our confidence: we can improve those methods and, with them, we can make better clinical decisions for patients. Here we discuss the potential improvements.

PET pharmacokinetic modelling

Part of this thesis has worked on the evaluation of methods which do not rely on the arterial input function. Elimination of its use has been a widely used strategy towards simplification of quantitative PET biomarkers. The main reason for this is that continuous arterial sampling used to acquire the input function is technically challenging and invasive and burdensome for the patient. With the introduction of new PET scanners with a long axial field of view (total body PET), continuous arterial sampling may no longer be required. Most scanners today have an axial length of up to 30 cm. Total body scanners on the other hand have an axial field of view of more than 70 cm. This enables the use of an image derived input function in both brain and lower body scanning, where the aorta would otherwise be outside of the field of view. For most tracers, it is likely that one or more blood samples will still be required to account for the fraction of radioactive metabolites, which is a crucial step to acquire reliable biomarker estimates from pharmacokinetic modelling. Potentially, whole body distributions of the tracer and uptake in the liver can be used to replace these samples, making the imaging process completely non-invasive.

In our studies regarding the pharmacokinetic modelling of FET in glioma (chapters 3 and 4), we evaluated simplified parametric methods that can for example be used for delineation of the tumour extent. In our data we could see that when an image is created in the interval of 20-40 minutes post injection of FET, the contrast between tumour and healthy tissue is larger than when the image is acquired at 60-90 minutes post injection. Hence, it would seem obvious that the image with better contrast is the better interval for measurement. This early interval is also currently recommended in RANO guidelines. However, from our technical validation study we conclude that images taken at the later interval are more accurate when compared to the results of kinetic modelling. The standardized uptake value ratio provides a composite measurement and it seems that this composition is more variable during early interval imaging. From a technical investigation we cannot estimate the potential clinical implications: changing the imaging interval could be of negligible importance to the clinical endpoint. For instance, it cannot be concluded that later imaging will lead to better tumour delineation. However, results from a subsequent clinical study on the accuracy of FET to detect glioma infiltration (using the same data) showed that the later interval estimates achieved higher accuracy.[197]

The potential for improvement justifies further investigation and clinical comparison of a simplified parameter from static images measured at different intervals is feasible.

MR quantification

MR based quantification is mainly challenging because of the inherent complexity of the imaging technique. For example, in the case of DCE MRI, the concentration estimation requires a number of measurements, each associated with its own uncertainty or error. The measured signal intensity is not only related to the concentration, but also to other tissue parameters, and instrument settings. This is why not only the signal enhancement needs to be measured, but also the pre-contrast relaxation time T1, and B1 inhomogeneity. Moreover, the T1 value in blood cannot be easily measured and is usually assumed. Hence, future studies may improve DCE quantification—and MRI quantification in general—by optimizing and standardizing the measurement and estimation techniques (such as improved T1 and B1 mapping). Sequence standardization and harmonization will also accelerate biomarker validation. Efforts in this area are ongoing, for example by the Quantitative Imaging Biomarkers Alliance (QIBA), which collects biomarker evaluation studies to create guidelines and standards for future research. Using these standards in future studies will decrease the variability of acquisition parameters, and thus make it easier to compare results. This is important because the variability in methods extends to clinical studies. For example, literature shows a large variability in results regarding the clinical value of MR diffusion parameters, e.g. to predict local recurrences or the overall survival of patients with head and neck squamous cell carcinoma. The differences between clinical reports are attributed to various factors, such as study design, statistical methods and image acquisition. Thus, standardisation of image acquisition will reduce disparities between clinical studies.

Technological advancement of MR systems may improve the measurements accuracy itself (for example with improved B1 homogeneity). Note, however, that technological improvements do not necessarily lead to improvement in quantitative performance. The imaging industry is generally driven by image quality measures, such as image resolution and signal-to-noise ratio, which do not always go hand in hand with quantification accuracy and precision. MR based imaging biomarker development will therefore benefit greatly from quality control intended specifically for quantification.

The complexity of the MR imaging technique makes (vendor) differences larger and the challenge of acquisition and analysis standardisation harder. Moreover, this complexity also naturally leads to differences in terminology: identical techniques can have different names. A common lexicon can help in recognising true differences between vendors and will also clarify technique descriptions in literature. However, even identical models and sequences can exhibit systematic differences, which is why reported system models and acquisition parameters ultimately have limited value. It may be worthwhile to develop

formats to summarize relevant performance metrics of the system and to include that information in publications. Most importantly, however, there is a need for more test-retest studies on MR quantification methods.

Deep learning

A general concern with the application of deep or machine learning in biomarker estimation is the interpretability of these models and the explainability of parameter estimates, such as we see with the factor analysis of radiomic features. Advanced and complex models may be able to very reliably predict for example treatment response; however, if it is unclear what the prediction is based on, any physician will be hesitant to use this information. Hence, future studies could either focus on models that provide explainable outcomes, or work to provide insight into why a models output has been generated.

With technological advancements the amount of imaging data is growing, for example due to increased spatial or temporal resolution. Processing of these data for parameter estimation can be challenging due to the time it takes to process the images. As demonstrated by the neural network for IVIM biomarker estimation, deep learning methods can provide quick calculation of imaging biomarkers. It is therefore likely that the number of deep learning applications will increase, especially for the processing of dynamic imaging data that are often slower to process with more conventional parameter estimation techniques. Conventional techniques for image analysis may be useful in the development of deep learning models. Such techniques could provide supervision during training and validation. This would not produce new or improved imaging biomarkers—after all, in this case the model will have learned to copy the conventional technique—however, it could considerably reduce the processing time and/or required resources in the clinic, which is where deep learning methods can provide the most benefit.

Addendum

References

1. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics* 2001; 69: 89–95.
2. Strimbu K, Tavel JA. What are biomarkers? *Current Opinion in HIV and AIDS* 2010; 5: 463–466.
3. O'Connor JP, Jackson A, Asselin M-C, et al. Quantitative imaging biomarkers in the clinical development of targeted therapeutics: current and future perspectives. *The Lancet Oncology* 2008; 9: 766–776.
4. Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. *Stat Methods Med Res* 2015; 24: 9–26.
5. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* 2018; 68: 394–424.
6. Marur S, Forastiere AA. Head and Neck Squamous Cell Carcinoma: Update on Epidemiology, Diagnosis, and Treatment. *Mayo Clinic Proceedings* 2016; 91: 386–396.
7. Huang SH, O'Sullivan B. Overview of the 8th Edition TNM Classification for Head and Neck Cancer. *Curr Treat Options in Oncol* 2017; 18: 40.
8. Pulte D, Brenner H. Changes in Survival in Head and Neck Cancers in the Late 20th and Early 21st Century: A Period Analysis. *The Oncologist* 2010; 15: 994–1001.
9. Shannon CE. A mathematical theory of communication. *The Bell System Technical Journal* 1948; 27: 623–656.
10. Innis RB, Cunningham VJ, Delforge J, et al. Consensus Nomenclature for in vivo Imaging of Reversibly Binding Radioligands. *J Cereb Blood Flow Metab* 2007; 27: 1533–1539.
11. Wintermark M, Sesay M, Barbier E, et al. Comparative Overview of Brain Perfusion Imaging Techniques. *Stroke* 2005; 36: e83–e99.
12. Frackowiak RS, Lenzi GL, Jones T, et al. Quantitative measurement of regional cerebral blood flow and oxygen metabolism in man using 15O and positron emission tomography: theory, procedure, and normal values. *J Comput Assist Tomogr* 1980; 4: 727–736.
13. Herscovitch P, Markham J, Raichle ME. Brain blood flow measured with intravenous H2(15)O. I. Theory and error analysis. *J Nucl Med* 1983; 24: 782–789.
14. Raichle ME, Martin WR, Herscovitch P, et al. Brain blood flow measured with intravenous H2(15)O. II. Implementation and validation. *J Nucl Med* 1983; 24: 790–798.
15. Huang SC, Carson RE, Hoffman EJ, et al. Quantitative measurement of local cerebral blood flow in humans by positron computed tomography and 15O-water. *J Cereb Blood Flow Metab* 1983; 3: 141–153.
16. Kanno I, Lammertsma AA, Heather JD, et al. Measurement of cerebral blood flow using bolus inhalation of C15O2 and positron emission tomography: description of the method and its comparison with the C15O2 continuous inhalation method. *Journal of Cerebral Blood Flow & Metabolism* 1984; 4: 224–234.
17. Meyer E. Simultaneous Correction for Tracer Arrival Delay and Dispersion in CBF Measurements by the H215O Autoradiographic Method and Dynamic PET. *J Nucl Med* 1989; 30: 1069–1078.
18. Lammertsma AA, Cunningham VJ, Deiber MP, et al. Combination of dynamic and integral methods for generating reproducible functional CBF images. *Journal of Cerebral Blood Flow & Metabolism* 1990; 10: 675–686.
19. Mejia MA, Itoh M, Watabe H, et al. Simplified nonlinearity correction of oxygen-15-water regional cerebral blood flow images without blood sampling. *Journal of Nuclear Medicine* 1994; 35: 1870–1877.

20. Watabe H, Itoh M, Cunningham V, et al. Noninvasive Quantification of rCBF Using Positron Emission Tomography. *J Cereb Blood Flow Metab* 1996; 16: 311–319.
21. Iida H, Law I, Pakkenberg B, et al. Quantitation of Regional Cerebral Blood Flow Corrected for Partial Volume Effect Using O-15 Water and PET: I. Theory, Error Analysis, and Stereologic Comparison. *Journal of Cerebral Blood Flow & Metabolism* 2000; 20: 1237–1251.
22. Treyer V, Jobin M, Burger C, et al. Quantitative cerebral H2 15O perfusion PET without arterial blood sampling, a method based on washout rate. *European Journal of Nuclear Medicine and Molecular Imaging* 2003; 30: 572–580.
23. Boellaard R, Knaapen P, Rijbroek A, et al. Evaluation of Basis Function and Linear Least Squares Methods for Generating Parametric Blood Flow Images Using 15O-Water and Positron Emission Tomography. *Molecular Imaging and Biology* 2005; 7: 273–285.
24. Kety SS, Schmidt CF. The Nitrous Oxide Method for the Quantitative Determination of Cerebral Blood Flow in Man: Theory, Procedure and Normal Values. *J Clin Invest* 1948; 27: 476–483.
25. Boellaard R, van Lingen A, van Balen SCM, et al. Characteristics of a new fully programmable blood sampling device for monitoring blood radioactivity during PET. *European Journal of Nuclear Medicine* 2001; 28: 81–89.
26. Zanotti-Fregonara P, Chen K, Liow J-S, et al. Image-derived input function for brain PET studies: many challenges and few opportunities. *Journal of Cerebral Blood Flow & Metabolism* 2011; 31: 1986–1998.
27. Lammertsma AA. Noninvasive estimation of cerebral blood flow. *Journal of Nuclear Medicine* 1994; 35: 1878–1879.
28. Heijtel DFR, Mutsaerts HJMM, Bakker E, et al. Accuracy and precision of pseudo-continuous arterial spin labeling perfusion during baseline and hypercapnia: A head-to-head comparison with 15O H2O positron emission tomography. *NeuroImage* 2014; 92: 182–192.
29. Svarer C, Madsen K, Hasselbalch SG, et al. MR-based automatic delineation of volumes of interest in human brain PET images using probability maps. *NeuroImage* 2005; 24: 969–979.
30. Hammers A, Allom R, Koepp MJ, et al. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human Brain Mapping* 2003; 19: 224–247.
31. Bol A, Vanmelckenbeke P, Michel C, et al. Measurement of cerebral blood flow with a bolus of oxygen-15-labelled water: comparison of dynamic and integral methods. *European Journal of Nuclear Medicine and Molecular Imaging* 1990; 17: 234–241.
32. Lammertsma AA, Martin AJ, Friston KJ, et al. In vivo measurement of the volume of distribution of water in cerebral grey matter: effects on the calculation of regional cerebral blood flow. *Journal of Cerebral Blood Flow & Metabolism* 1992; 12: 291–295.
33. Alpert NM, Eriksson L, Chang JY, et al. Strategy for the measurement of regional cerebral blood flow using short-lived tracers and emission tomography. *Journal of Cerebral Blood Flow & Metabolism* 1984; 4: 28–34.
34. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307–310.
35. Yaqub M, Boellaard R, Kropholler MA, et al. Optimization algorithms and weighting factors for analysis of dynamic PET studies. *Phys Med Biol* 2006; 51: 4217.
36. Wester HJ, Herz M, Weber W, et al. Synthesis and radiopharmacology of O-(2-[18F]fluoroethyl)-L-tyrosine for tumor imaging. *J Nucl Med* 1999; 40: 205–212.
37. Langen K-J, Stoffels G, Filss C, et al. Imaging of amino acid transport in brain tumours: Positron emission tomography with O-(2-[18 F]fluoroethyl)- L -tyrosine (FET). *Methods* 2017; 130: 124–134.
38. Heiss P, Mayer S, Herz M, et al. Investigation of transport mechanism and uptake kinetics of O-(2-[18F]fluoroethyl)-L-tyrosine in vitro and in vivo. *J Nucl Med* 1999; 40: 1367–1373.

39. Albert NL, Weller M, Suchorska B, et al. Response Assessment in Neuro-Oncology working group and European Association for Neuro-Oncology recommendations for the clinical use of PET imaging in gliomas. *Neuro-Oncology* 2016; 18: 1199–1208.

40. Lammertsma AA. Tracer Kinetic Modelling. In: Dierckx RAJO, Otte A, de Vries EFJ, et al. (eds) *PET and SPECT in Neurology*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 59–73.

41. Vander Borght T, Asenbaum S, Bartenstein P, et al. EANM procedure guidelines for brain tumour imaging using labelled amino acid analogues. *European Journal of Nuclear Medicine and Molecular Imaging* 2006; 33: 1374–1380.

42. Langen K-J, Bartenstein P, Boecker H, et al. German guidelines for brain tumour imaging by PET and SPECT using labelled amino acids: *Nuklearmedizin* 2011; 50: 167–173.

43. Bolcaen J, Lybaert K, Moerman L, et al. Kinetic Modeling and Graphical Analysis of 18F-Fluoromethylcholine (FCho), 18F-Fluoroethyltyrosine (FET) and 18F-Fluorodeoxyglucose (FDG) PET for the Fiscrimination between High-Grade Glioma and Radiation Necrosis in Rats. *PLOS ONE* 2016; 11: e0161845.

44. Richard MA, Fouquet JP, Lebel R, et al. Determination of an Optimal Pharmacokinetic Model of ^{18}F-FET for Quantitative Applications in Rat Brain Tumors. *Journal of Nuclear Medicine* 2017; 58: 1278–1284.

45. Kratochwil C, Combs SE, Leotta K, et al. Intra-individual comparison of 18F-FET and 18F-DOPA in PET imaging of recurrent brain tumors. *Neuro-Oncology* 2014; 16: 434–440.

46. Thiele F, Ehmer J, Piroth MD, et al. The quantification of dynamic FET PET imaging and correlation with the clinical outcome in patients with glioblastoma. *Phys Med Biol* 2009; 54: 5525–5539.

47. Loeb R, Navab N, Ziegler SI. Direct Parametric Reconstruction Using Anatomical Regularization for Simultaneous PET/MRI Data. *IEEE Transactions on Medical Imaging* 2015; 34: 2233–2247.

48. Pauleit D, Floeth F, Herzog H, et al. Whole-body distribution and dosimetry of O-(2-[18F]fluoroethyl)-l-tyrosine. *European Journal of Nuclear Medicine and Molecular Imaging* 2003; 30: 519–524.

49. Verburg N, Pouwels PJW, Boellaard R, et al. Accurate Delineation of Glioma Infiltration by Advanced PET/MR Neuro-Imaging (FRONTIER Study): A Diagnostic Study Protocol. *Neurosurgery* 2016; 79: 535–540.

50. Louis DN, Ohgaki H, Wiestler OD, et al. (eds). *WHO classification of tumours of the central nervous system*. Revised 4th edition. Lyon: International Agency For Research On Cancer, 2016.

51. Zuhayra M, Alfteimi A, Forstner CV, et al. New approach for the synthesis of [18F]fluoroethyltyrosine for cancer imaging: Simple, fast, and high yielding automated synthesis. *Bioorganic & Medicinal Chemistry* 2009; 17: 7441–7448.

52. Gunn RN, Sargent PA, Bench CJ, et al. Tracer Kinetic Modeling of the 5-HT1AReceptor Ligand [carbonyl-11C]WAY-100635 for PET. *NeuroImage* 1998; 8: 426–440.

53. Gunn RN, Gunn SR, Cunningham VJ. Positron emission tomography compartmental models. *Journal of Cerebral Blood Flow & Metabolism* 2001; 21: 635–652.

54. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; 19: 716–723.

55. Blomqvist G, Pauli S, Farde L, et al. Maps of receptor binding parameters in the human brain ? a kinetic analysis of PET measurements. *European Journal of Nuclear Medicine* 1990; 16: 257–265.

56. Cunningham VJ, Hume SP, Price GR, et al. Compartmental analysis of diprenorphine binding to opiate receptors in the rat in vivo and its comparison with equilibrium data in vitro. *Journal of Cerebral Blood Flow & Metabolism* 1991; 11: 1–9.

57. Lammertsma AA, Hume SP. Simplified reference tissue model for PET receptor studies. *Neuroimage* 1996; 4: 153–158.

58. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical methods in medical research* 1999; 8: 135–160.

59. Salinas CA, Searle GE, Gunn RN. The Simplified Reference Tissue Model: Model Assumption Violations and Their Impact on Binding Potential. *J Cereb Blood Flow Metab* 2015; 35: 304–311.

60. Koopman T, Verburg N, Schuit RC, et al. Quantification of O-(2-[18F]fluoroethyl)-L-tyrosine kinetics in glioma. *EJNMMI Research*; 8. Epub ahead of print December 2018. DOI: 10.1186/s13550-018-0418-0.

61. Logan J, Fowler JS, Volkow ND, et al. Graphical Analysis of Reversible Radioligand Binding from Time—Activity Measurements Applied to [N-11C-Methyl]-(-)-Cocaine PET Studies in Human Subjects. *Journal of Cerebral Blood Flow & Metabolism* 1990; 10: 740–747.

62. Cunningham VJ, Jones T. Spectral analysis of dynamic PET studies. *Journal of Cerebral Blood Flow & Metabolism* 1993; 13: 15–23.

63. Logan J, Fowler JS, Volkow ND, et al. Distribution volume ratios without blood sampling from graphical analysis of PET data. *Journal of Cerebral Blood Flow & Metabolism* 1996; 16: 834–840.

64. Gunn RN, Lammertsma AA, Hume SP, et al. Parametric Imaging of Ligand-Receptor Binding in PET Using a Simplified Reference Region Model. *NeuroImage* 1997; 6: 279–287.

65. Wu Y, Carson RE. Noise Reduction in the Simplified Reference Tissue Model for Neuroreceptor Functional Imaging. *Journal of Cerebral Blood Flow & Metabolism* 2002; 22: 1440–1452.

66. Ichise M, Ballinger JR, Golan H, et al. Noninvasive quantification of dopamine D2 receptors with iodine-123-IBF SPECT. *J Nucl Med* 1996; 37: 513–520.

67. Ichise M, Liow J-S, Lu J-Q, et al. Linearized Reference Tissue Parametric Imaging Methods: Application to [11C]DASB Positron Emission Tomography Studies of the Serotonin Transporter in Human Brain: *Journal of Cerebral Blood Flow & Metabolism* 2003; 1096–1112.

68. Yaqub M, Tolboom N, Boellaard R, et al. Simplified parametric methods for [11C]PIB studies. *NeuroImage* 2008; 42: 76–86.

69. Ashburner J, Friston KJ. Unified segmentation. *NeuroImage* 2005; 26: 839–851.

70. Slifstein M, Laruelle M. Effects of Statistical Noise on Graphic Analysis of PET Neuroreceptor Studies. *J Nucl Med* 2000; 41: 2083–2088.

71. Ichise M, Cohen RM, Carson RE. Noninvasive Estimation of Normalized Distribution Volume: Application to the Muscarinic-2 Ligand [^{18} F]FP-TZTP. *Journal of Cerebral Blood Flow & Metabolism* 2008; 28: 420–430.

72. Yaqub M, Tolboom N, van Berckel BNM, et al. Simplified parametric methods for [18F]FDDNP studies. *NeuroImage* 2010; 49: 433–441.

73. Tofts PS, Brix G, Buckley DL, et al. Estimating kinetic parameters from dynamic contrast-enhanced t1-weighted MRI of a diffusable tracer: Standardized quantities and symbols. *Journal of Magnetic Resonance Imaging* 1999; 10: 223–232.

74. Bernstein JM, Homer JJ, West CM. Dynamic contrast-enhanced magnetic resonance imaging biomarkers in head and neck cancer: Potential to guide treatment? A systematic review. *Oral Oncology* 2014; 50: 963–970.

75. Shukla-Dave A, Obuchowski NA, Chenevert TL, et al. Quantitative imaging biomarkers alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. *Journal of Magnetic Resonance Imaging* 2019; 49: e101–e121.

76. Lavini C. Simulating the effect of input errors on the accuracy of Tofts' pharmacokinetic model parameters. *Magnetic Resonance Imaging* 2015; 33: 222–235.

77. Peled S, Vangel M, Kikinis R, et al. Selection of Fitting Model and Arterial Input Function for Repeatability in Dynamic Contrast-Enhanced Prostate MRI. *Academic Radiology* 2019; 26: e241–e251.

78. Huang W, Chen Y, Fedorov A, et al. The Impact of Arterial Input Function Determination Variations on Prostate Dynamic Contrast-Enhanced Magnetic Resonance Imaging Pharmacokinetic Modeling: A Multicenter Data Analysis Challenge. *Tomography* 2016; 2: 56–66.

79. Keil VC, Mädler B, Gieseke J, et al. Effects of arterial input function selection on kinetic parameters in brain dynamic contrast-enhanced MRI. *Magnetic Resonance Imaging* 2017; 40: 83–90.

80. Azahaf M, Haberley M, Betrouni N, et al. Impact of arterial input function selection on the accuracy of dynamic contrast-enhanced MRI quantitative analysis for the diagnosis of clinically significant prostate cancer: Impact of AIF on K^{trans} in PCa. *J Magn Reson Imaging* 2016; 43: 737–749.

81. Parker GJM, Roberts C, Macdonald A, et al. Experimentally-derived functional form for a population-averaged high-temporal-resolution arterial input function for dynamic contrast-enhanced MRI. *Magnetic Resonance in Medicine* 2006; 56: 993–1000.

82. Rata M, Collins DJ, Darcy J, et al. Assessment of repeatability and treatment response in early phase clinical trials using DCE-MRI: comparison of parametric analysis using MR- and CT-derived arterial input functions. *European Radiology* 2016; 26: 1991–1998.

83. Rijpkema M, Kaanders JHAM, Joosten FBM, et al. Method for quantitative mapping of dynamic MRI contrast agent uptake in human tumors. *Journal of Magnetic Resonance Imaging* 2001; 14: 457–463.

84. Ashton E, Raunig D, Ng C, et al. Scan-rescan variability in perfusion assessment of tumors in MRI using both model and data-derived arterial input functions. *Journal of Magnetic Resonance Imaging* 2008; 28: 791–796.

85. Lavini C, Verhoeff JJC. Reproducibility of the gadolinium concentration measurements and of the fitting parameters of the vascular input function in the superior sagittal sinus in a patient population. *Magnetic Resonance Imaging* 2010; 28: 1420–1430.

86. Mendichovszky IA, Cutajar M, Gordon I. Reproducibility of the aortic input function (AIF) derived from dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) of the kidneys in a volunteer study. *European Journal of Radiology* 2009; 71: 576–581.

87. Willemsen ACH, Hoeben A, Lalisang RI, et al. Disease-induced and treatment-induced alterations in body composition in locally advanced head and neck squamous cell carcinoma. *Journal of Cachexia, Sarcopenia and Muscle*. Epub ahead of print 19 September 2019. DOI: 10.1002/jcsm.12487.

88. Yarnykh VL. Actual flip-angle imaging in the pulsed steady state: A method for rapid three-dimensional mapping of the transmitted radiofrequency field. *Magnetic Resonance in Medicine* 2007; 57: 192–200.

89. Gupta RK. A new look at the method of variable nutation angle for the measurement of spin-lattice relaxation times using fourier transform NMR. *Journal of Magnetic Resonance (1969)* 1977; 25: 231–235.

90. Wehrli FW. Fast-Scan Magnetic Resonance - Principles and Applications. New York: Raven Press, 1991, p. 12.

91. Heilmann M, Kiessling F, Enderlin M, et al. Determination of Pharmacokinetic Parameters in DCE MRI: Consequence of Nonlinearity Between Contrast Agent Concentration and Signal Intensity. *Investigative Radiology* 2006; 41: 536–543.

92. Schabel MC, Parker DL. Uncertainty and bias in contrast concentration measurements using spoiled gradient echo pulse sequences. *Physics in Medicine and Biology* 2008; 53: 2345–2373.

93. Stanisz GJ, Odrobina EE, Pun J, et al. T1, T2 relaxation and magnetization transfer in tissue at 3T. *Magnetic Resonance in Medicine* 2005; 54: 507–512.

94. Klawer EME, van Houdt PJ, Simonis FFJ, et al. Improved repeatability of dynamic contrast-enhanced MRI using the complex MRI signal to derive arterial input functions: a test-retest study in prostate cancer patients. *Magnetic Resonance in Medicine* 2019; 81: 3358–3369.

95. Yaqub M, Boellaard R, Kropholler MA, et al. Optimization algorithms and weighting factors for analysis of dynamic PET studies. *Physics in Medicine and Biology* 2006; 51: 4217–4232.

96. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 1999; 8: 135–160.

97. Bland M. How should I calculate a within-subject coefficient of variation?, https://www-users.york.ac.uk/~mb55/meas/cv.htm (2006, accessed 12 August 2019).

98. Simonis FFJ, Sbrizzi A, Beld E, et al. Improving the arterial input function in dynamic contrast enhanced MRI by fitting the signal in the complex plane: Improving AIF in DCE-MRI by Fitting the Complex Signal. *Magnetic Resonance in Medicine* 2016; 76: 1236–1245.

99. Garpebring A, Wirestam R, Östlund N, et al. Effects of inflow and radiofrequency spoiling on the arterial input function in dynamic contrast-enhanced MRI: A combined phantom and simulation study. *Magnetic Resonance in Medicine* 2011; 65: 1670–1679.

100. Roberts C, Little R, Watson Y, et al. The effect of blood inflow and B_1-field inhomogeneity on measurement of the arterial input function in axial 3D spoiled gradient echo dynamic contrast-enhanced MRI: AIF Errors in DCE-MRI. *Magnetic Resonance in Medicine* 2011; 65: 108–119.

101. van Schie JJN, Lavini C, van Vliet LJ, et al. Estimating the arterial input function from dynamic contrast-enhanced MRI data with compensation for flow enhancement (I): Theory, method, and phantom experiments: Flow AIF 1. *Journal of Magnetic Resonance Imaging* 2018; 47: 1190–1196.

102. Yazıcı B, Erdoğmuş B, Tugay A. Cerebral blood flow measurements of the extracranial carotid and vertebral arteries with Doppler ultrasonography in healthy adults. *Diagn Interv Radiol* 2005; 11: 195–198.

103. Makkat S, Luypaert R, Sourbron S, et al. Assessment of tumor blood flow in breast tumors with T1-dynamic contrast-enhanced MR Imaging: Impact of dose reduction and the use of a prebolus technique on diagnostic efficacy. *Journal of Magnetic Resonance Imaging* 2010; 31: 556–561.

104. Risse F, Semmler W, Kauczor H-U, et al. Dual-bolus approach to quantitative measurement of pulmonary perfusion by contrast-enhanced MRI. *Journal of Magnetic Resonance Imaging* 2006; 24: 1284–1290.

105. Christian TF, Aletras AH, Arai AE. Estimation of absolute myocardial blood flow during first-pass MR perfusion imaging using a dual-bolus injection technique: Comparison to single-bolus injection method. *Journal of Magnetic Resonance Imaging* 2008; 27: 1271–1277.

106. Köstler H, Ritter C, Lipp M, et al. Prebolus quantitative MR heart perfusion imaging: Prebolus Quantitative MR Heart Perfusion. *Magnetic Resonance in Medicine* 2004; 52: 296–299.

107. Oechsner M, Mühlhäusler M, Ritter CO, et al. Quantitative contrast-enhanced perfusion measurements of the human lung using the prebolus approach. *Journal of Magnetic Resonance Imaging* 2009; 30: 104–111.

108. Onxley JD, Yoo DS, Muradyan N, et al. Comprehensive Population-Averaged Arterial Input Function for Dynamic Contrast–Enhanced vMagnetic Resonance Imaging of Head and Neck Cancer. *International Journal of Radiation Oncology*Biology*Physics* 2014; 89: 658–665.

109. Galbraith SM, Lodge MA, Taylor NJ, et al. Reproducibility of dynamic contrast-enhanced MRI in human muscle and tumours: comparison of quantitative and semi-quantitative analysis. *NMR in Biomedicine* 2002; 15: 132–142.

110. Wong KH, Panek R, Welsh L, et al. The Predictive Value of Early Assessment After 1 Cycle of Induction Chemotherapy with 18F-FDG PET/CT and Diffusion-Weighted MRI for Response to Radical Chemoradiotherapy in Head and Neck Squamous Cell Carcinoma. *Journal of Nuclear Medicine* 2016; 57: 1843–1850.

111. Wong KH, Panek R, Dunlop A, et al. Changes in multimodality functional imaging parameters early during chemoradiation predict treatment response in patients with locally advanced head and neck cancer. *European Journal of Nuclear Medicine and Molecular Imaging* 2018; 45: 759–767.

112. Hauser T, Essig M, Jensen A, et al. Prediction of treatment response in head and neck carcinomas using IVIM-DWI: Evaluation of lymph node metastasis. *European Journal of Radiology* 2014; 83: 783–787.

113. Noij DP, Martens RM, Marcus JT, et al. Intravoxel incoherent motion magnetic resonance imaging in head and neck cancer: A systematic review of the diagnostic and prognostic value. *Oral Oncology* 2017; 68: 81–91.

114. Le Bihan D, Breton E, Lallemand D, et al. MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology* 1986; 161: 401–407.

115. Le Bihan D, Breton E, Lallemand D, et al. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. *Radiology* 1988; 168: 497–505.

116. Vandecaveye V, Dirix P, De Keyzer F, et al. Diffusion-Weighted Magnetic Resonance Imaging Early After Chemoradiotherapy to Monitor Treatment Response in Head-and-Neck Squamous Cell Carcinoma. *International Journal of Radiation Oncology*Biology*Physics* 2012; 82: 1098–1107.

117. Gurney-Champion OJ, Froeling M, Klaassen R, et al. Minimizing the Acquisition Time for Intravoxel Incoherent Motion Magnetic Resonance Imaging Acquisitions in the Liver and Pancreas: *Investigative Radiology* 2016; 51: 211–220.

118. Barbieri S, Donati OF, Froehlich JM, et al. Impact of the calculation algorithm on biexponential fitting of diffusion-weighted MRI in upper abdominal organs: Impact of the Calculation Algorithm on IVIM Parameters in Upper Abdominal Organs. *Magnetic Resonance in Medicine* 2016; 75: 2175–2184.

119. Gurney-Champion OJ, Klaassen R, Froeling M, et al. Comparison of six fit algorithms for the intra-voxel incoherent motion model of diffusion-weighted magnetic resonance imaging data of pancreatic cancer patients. *PLOS ONE* 2018; 13: e0194590.

120. Orton MR, Collins DJ, Koh D-M, et al. Improved intravoxel incoherent motion analysis of diffusion weighted imaging by data driven Bayesian modeling: Improved IVIM Analysis with Bayesian Modelling. *Magn Reson Med* 2014; 71: 411–420.

121. Barbieri S, Gurney-Champion OJ, Klaassen R, et al. Deep learning how to fit an intravoxel incoherent motion model to diffusion-weighted MRI. *Magnetic Resonance in Medicine*. Epub ahead of print 7 August 2019. DOI: 10.1002/mrm.27910.

122. Wetscherek A, Stieltjes B, Laun FB. Flow-compensated intravoxel incoherent motion diffusion imaging: Flow-Compensated IVIM Diffusion Imaging. *Magn Reson Med* 2015; 74: 410–419.

123. Wáng YXJ. Living tissue intravoxel incoherent motion (IVIM) diffusion MR analysis without b=0 image: an example for liver fibrosis evaluation. *Quant Imaging Med Surg* 2019; 9: 127–133.

124. Xiao B-H, Huang H, Wang L-F, et al. Diffusion MRI Derived per Area Vessel Density as a Surrogate Biomarker for Detecting Viral Hepatitis B-Induced Liver Fibrosis: A Proof-of-Concept Study. *SLAS TECHNOLOGY: Translating Life Sciences Innovation* 2020; 25: 474–483.

125. van der Bel R, Gurney-Champion OJ, Froeling M, et al. A tri-exponential model for intravoxel incoherent motion analysis of the human kidney: In silico and during pharmacological renal perfusion modulation. *European Journal of Radiology* 2017; 91: 168–174.

126. Neal RM. Slice sampling. *Ann Statist* 2003; 31: 705–767.

127. Bretthorst GL, Hutton WC, Garbow JR, et al. Exponential parameter estimation (in NMR) using Bayesian probability theory. *Concepts in Magnetic Resonance Part A* 2005; 27A: 55–63.

128. Gustafsson O, Montelius M, Starck G, et al. Impact of prior distributions and central tendency measures on Bayesian intravoxel incoherent motion model fitting: Impact of Prior and Central Tendency Measure on Bayesian IVIM Model Fitting. *Magnetic Resonance in Medicine* 2018; 79: 1674–1683.

129. Barbieri S. GitHub Repository Deep IVIM. *GitHub*, https://github.com/sebbarb/deep_ivim/commit/4aa19e77 (2019, accessed 30 September 2019).

130. LeNail A. NN-SVG: Publication-Ready Neural Network Architecture Schematics. *JOSS* 2019; 4: 747.

131. Kang KM, Choi SH, Kim DE, et al. Application of Cardiac Gating to Improve the Reproducibility of Intravoxel Incoherent Motion Measurements in the Head and Neck. *Magnetic Resonance in Medical Sciences* 2017; 16: 190–202.

132. Hoang JK, Choudhury KR, Chang J, et al. Diffusion-Weighted Imaging for Head and Neck Squamous Cell Carcinoma: Quantifying Repeatability to Understand Early Treatment-Induced Change. *American Journal of Roentgenology* 2014; 203: 1104–1108.

133. Paudyal R, Konar AS, Obuchowski NA, et al. Repeatability of Quantitative Diffusion-Weighted Imaging Metrics in Phantoms, Head-and-Neck and Thyroid Cancers: Preliminary Findings. *Tomography* 2019; 5: 15–25.

134. Jensen JH, Helpern JA, Ramani A, et al. Diffusional kurtosis imaging: The quantification of non-gaussian water diffusion by means of magnetic resonance imaging. *Magnetic Resonance in Medicine* 2005; 53: 1432–1440.

135. Lu Y, Jansen JFA, Mazaheri Y, et al. Extension of the intravoxel incoherent motion model to non-gaussian diffusion in head and neck cancer. *Journal of Magnetic Resonance Imaging* 2012; 36: 1088–1096.

136. Bennett KM, Schmainda KM, Bennett (Tong) R, et al. Characterization of continuously distributed cortical water diffusion rates with a stretched-exponential model. *Magnetic Resonance in Medicine* 2003; 50: 727–734.

137. Perucho JAU, Chang HCC, Vardhanabhuti V, et al. B-Value Optimization in the Estimation of Intravoxel Incoherent Motion Parameters in Patients with Cervical Cancer. *Korean J Radiol* 2020; 21: 218.

138. Sijtsema ND, Petit SF, Poot DHJ, et al. An optimal acquisition and post-processing pipeline for hybrid IVIM-DKI in head and neck. *Magn Reson Med* 2020; mrm.28461.

139. Bonomo P, Merlotti A, Olmetto E, et al. What is the prognostic impact of FDG PET in locally advanced head and neck squamous cell carcinoma treated with concomitant chemo-radiotherapy? A systematic review and meta-analysis. *Eur J Nucl Med Mol Imaging* 2018; 45: 2122–2138.

140. Brockstein B, Haraf DJ, Rademaker AW, et al. Patterns of failure, prognostic factors and survival in locoregionally advanced head and neck cancer treated with concomitant chemoradiotherapy: a 9-year, 337-patient, multi-institutional experience. *Annals of Oncology* 2004; 15: 1179–1186.

141. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012: Globocan 2012. *Int J Cancer* 2015; 136: E359–E386.

142. Wong AJ, Kanwar A, Mohamed AS, et al. Radiomics in head and neck cancer: from exploration to application. *Transl Cancer Res* 2016; 5: 371–382.

143. Marusyk A, Polyak K. Tumor heterogeneity: Causes and consequences. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 2010; 1805: 105–117.

144. 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.

145. Vallières M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep* 2017; 7: 10117.

146. Pickering CR, Shah K, Ahmed S, et al. CT Imaging Correlates of Genomic Expression for Oral Cavity Squamous Cell Carcinoma. *AJNR Am J Neuroradiol* 2013; 34: 1818–1822.

147. Dang M, Lysack JT, Wu T, et al. MRI Texture Analysis Predicts p53 Status in Head and Neck Squamous Cell Carcinoma. *American Journal of Neuroradiology* 2015; 36: 166–170.

148. Yaromina A, Krause M, Baumann M. Individualization of cancer treatment from radiotherapy perspective. *Molecular Oncology* 2012; 6: 211–221.

149. Quon H, Brizel DM. Predictive and Prognostic Role of Functional Imaging of Head and Neck Squamous Cell Carcinomas. *Seminars in Radiation Oncology* 2012; 22: 220–232.

150. King AD, Thoeny HC. Functional MRI for the prediction of treatment response in head and neck squamous cell carcinoma: potential and limitations. *Cancer Imaging* 2016; 16: 23.

151. Lambin P, Roelofs E, Reymen B, et al. 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy'. *Radiotherapy and Oncology* 2013; 109: 159–164.

152. Troost EGC, Schinagl DAX, Bussink J, et al. Innovations in Radiotherapy Planning of Head and Neck Cancers: Role of PET. *Journal of Nuclear Medicine* 2010; 51: 66–76.

153. Heron DE, Andrade RS, Beriwal S, et al. PET-CT in Radiation Oncology: The Impact on Diagnosis, Treatment Planning, and Assessment of Treatment Response. *American Journal of Clinical Oncology* 2008; 31: 352–362.

154. Bogowicz M, Riesterer O, Stark LS, et al. Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma. *Acta Oncologica* 2017; 56: 1531–1536.

155. Buvat I, Orlhac F, Soussan M. Tumor Texture Analysis in PET: Where Do We Stand? *Journal of Nuclear Medicine* 2015; 56: 1642–1644.

156. Sollini M, Cozzi L, Antunovic L, et al. PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology. *Sci Rep* 2017; 7: 358.

157. Hatt M, Majdoub M, Vallieres M, et al. 18F-FDG PET Uptake Characterization Through Texture Analysis: Investigating the Complementary Nature of Heterogeneity and Functional Tumor Volume in a Multi-Cancer Site Patient Cohort. *Journal of Nuclear Medicine* 2015; 56: 38–44.

158. Cheng N-M, Dean Fang Y-H, Tung-Chieh Chang J, et al. Textural Features of Pretreatment 18F-FDG PET/CT Images: Prognostic Significance in Patients with Advanced T-Stage Oropharyngeal Squamous Cell Carcinoma. *Journal of Nuclear Medicine* 2013; 54: 1703–1709.

159. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications* 2014; 5: 4006.

160. El Naqa I, Grigsby PW, Apte A, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognition* 2009; 42: 1162–1171.

161. Hashibe M, Brennan P, Benhamou S, et al. Alcohol Drinking in Never Users of Tobacco, Cigarette Smoking in Never Drinkers, and the Risk of Head and Neck Cancer: Pooled Analysis in the International Head and Neck Cancer Epidemiology Consortium. *JNCI Journal of the National Cancer Institute* 2007; 99: 777–789.

162. Freedman ND, Schatzkin A, Leitzmann MF, et al. Alcohol and head and neck cancer risk in a prospective study. *Br J Cancer* 2007; 96: 1469–1474.

163. 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.

164. Surti S, Kuhn A, Werner ME, et al. Performance of Philips Gemini TF PET/CT Scanner with Special Consideration for Its Time-of-Flight Imaging Capabilities. *J Nucl Med* 2007; 48: 471–480.

165. Martens RM, Noij DP, Koopman T, et al. Predictive value of quantitative diffusion-weighted imaging and 18-F-FDG-PET in head and neck squamous cell carcinoma treated by (chemo)radiotherapy. *European Journal of Radiology* 2019; 113: 39–50.

166. Sobin LH, Gospodarowicz MK, Wittekind C (eds). *TNM Classification of Malignant Tumours*. 7th edition. Chichester, West Sussex, UK: Wiley-Blackwell, 2009.

167. Frings V, de Langen AJ, Smit EF, et al. Repeatability of Metabolically Active Volume Measurements with 18F-FDG and 18F-FLT PET in Non-Small Cell Lung Cancer. *Journal of Nuclear Medicine* 2010; 51: 1870–1877.

168. Cheebsumon P, van Velden FHP, Yaqub M, et al. Effects of Image Characteristics on Performance of Tumor Delineation Methods: A Test-Retest Assessment. *Journal of Nuclear Medicine* 2011; 52: 1550–1558.

169. Cysouw MCF, Kramer GM, Schoonmade LJ, et al. Impact of partial-volume correction in oncological PET studies: a systematic review and meta-analysis. *Eur J Nucl Med Mol Imaging* 2017; 44: 2105–2116.

170. Pfaehler E, Zwanenburg A, Jong JR de, et al. RaCaT: An open source and easy to use radiomics calculator tool. *PLOS ONE* 2019; 14: e0212223.

171. Pfaehler E, Sluis J van, Merema BBJ, et al. Experimental Multicenter and Multivendor Evaluation of the Performance of PET Radiomic Features Using 3-Dimensionally Printed Phantom Inserts. *J Nucl Med* 2020; 61: 469–476.

172. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* 2020; 295: 328–338.

173. Zwanenburg A, Leger S, Vallières M, et al. Image biomarker standardisation initiative. *arXiv:161207003*. Epub ahead of print 17 December 2019. DOI: 10.1148/radiol.2020191145.

174. Peeters CFW, Übelhör C, Mes SW, et al. Stable prediction with radiomics data. *arXiv:190311696*, http://arxiv.org/abs/1903.11696 (2019, accessed 30 November 2020).

175. Peeters CFW, van de Wiel MA, van Wieringen WN. The spectral condition number plot for regularization parameter evaluation. *Comput Stat* 2020; 35: 629–646.

176. Guttman L. Some necessary conditions for common-factor analysis. *Psychometrika* 1954; 19: 149–161.

177. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015; 13: 1.

178. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med* 2015; 162: W1.

179. Browne MW. Cross-Validation Methods. *Journal of Mathematical Psychology* 2000; 44: 108–132.

180. Parmar C, Grossmann P, Bussink J, et al. Machine Learning methods for Quantitative Radiomic Biomarkers. *Scientific Reports* 2015; 5: 13087.

181. Desseroit M-C, Visvikis D, Tixier F, et al. Development of a nomogram combining clinical staging with 18F-FDG PET/CT image features in non-small-cell lung cancer stage I–III. *Eur J Nucl Med Mol Imaging* 2016; 43: 1477–1485.

182. Hatt M, Tixier F, Visvikis D, et al. Radiomics in PET/CT: More Than Meets the Eye? *J Nucl Med* 2017; 58: 365–366.

183. Chow LQM. Head and Neck Cancer. *N Engl J Med* 2020; 382: 60–72.

184. Ressom HW, Varghese RS, Zhang Z, et al. Classification algorithms for phenotype prediction in genomics and proteomics. *Front Biosci* 2008; 13: 691–708.

185. Cheng N-M, Fang Y-HD, Lee L, et al. Zone-size nonuniformity of 18F-FDG PET regional textural features predicts survival in patients with oropharyngeal cancer. *Eur J Nucl Med Mol Imaging* 2015; 42: 419–428.

186. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. *Cell* 2011; 144: 646–674.

187. Mirghani H, Blanchard P. Treatment de-escalation for HPV-driven oropharyngeal cancer: Where do we stand? *Clinical and Translational Radiation Oncology* 2018; 8: 4–11.

188. van den Bosch S, Dijkema T, Kunze-Busch MC, et al. Uniform FDG-PET guided GRAdient Dose prEscription to reduce late Radiation Toxicity (UPGRADE-RT): study protocol for a randomized clinical trial with dose reduction to the elective neck in head and neck squamous cell carcinoma. *BMC Cancer* 2017; 17: 208.

189. van den Bosch S, Dijkema T, Verhoef LCG, et al. Patterns of Recurrence in Electively Irradiated Lymph Node Regions After Definitive Accelerated Intensity Modulated Radiation Therapy for Head and Neck Squamous Cell Carcinoma. *Int J Radiat Oncol Biol Phys* 2016; 94: 766–774.

190. Ling DC, Bakkenist CJ, Ferris RL, et al. Role of Immunotherapy in Head and Neck Cancer. *Seminars in Radiation Oncology* 2018; 28: 12–16.

191. Parmar C, Grossmann P, Rietveld D, et al. Radiomic Machine-Learning Classifiers for Prognostic Biomarkers of Head and Neck Cancer. *Front Oncol* 2015; 5: 272.

192. Leijenaar RTH, Carvalho S, Hoebers FJP, et al. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncologica* 2015; 54: 1423–1429.

193. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiotherapy and Oncology* 2019; 130: 2–9.

194. Hatt M, Rest CCL, Tixier F, et al. Radiomics: Data Are Also Images. *J Nucl Med* 2019; 60: 38S-44S.

195. O'Connor JPB, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol* 2017; 14: 169–186.

196. Lassau N, Bonastre J, Kind M, et al. Validation of Dynamic Contrast-Enhanced Ultrasound in Predicting Outcomes of Antiangiogenic Therapy for Solid Tumors: The French Multicenter Support for Innovative and Expensive Techniques Study. *Investigative Radiology* 2014; 49: 794–800.

197. Verburg N, Koopman T, Yaqub M, et al. Direct comparison of [11C] choline and [18F] FET PET to detect glioma infiltration: a diagnostic accuracy study in eight patients. *EJNMMI Res* 2019; 9: 57.

Summary

This thesis describes various technical aspects of validation of quantitative imaging biomarkers derived from different imaging studies (PET, PET-CT, MRI) for various targets (perfusion, diffusion, cell metabolism). For the quantification of each of these biomarkers a combination of system modelling, signal processing and parameter estimation is required. Several aspects hereof (such as model selection, fitting routines and signal extraction methods) have been investigated in this PhD project.

A recurring theme in these studies is the accuracy and precision of a quantitative imaging biomarker. Precision or repeatability can be assessed using test-retest studies. With poor precision, the clinical value of a biomarker is limited. Moreover, clinical applicability of a method is important for implementation in routine care. The research described in this thesis therefore focused on optimization as well as simplification of quantification methods.

The first quantification method under study concerns tracer kinetic modelling: the application of the radiolabelled water tracer $[^{15}O]H_2O$ to measure tissue perfusion. $[^{15}O]H_2O$ is an ideal perfusion tracer, because extraction of water (from the blood compartment) is 100%. The kinetic model to quantify perfusion consists of only one tissue compartment and the arterial blood compartment.

While the model is relatively simple, the downside of the method is the need for measurement of the arterial input function, i.e., the tracer concentration in arterial blood. Deriving this input function directly from the image is only possible when a large arterial vessel, such as the aorta, is within the field of view of the PET system. Without such an alternative, the tracer concentration in blood must be measured by (continuous) arterial blood sampling, which is invasive and burdensome for the patient.

The study in **chapter two** investigated the possibility of simplifying the clinical protocol in tracer kinetic modelling of brain perfusion with $[^{15}O]H_2O$ PET. In brain PET measurements, the aorta is not in the field of view of the scanner and PET resolution is not high enough to accurately measure concentration in the small arteries of the neck.[1] The study therefore looked into the possibility of calculating perfusion without knowing the arterial input function. A number of simplified methods for estimating (absolute and relative) cerebral perfusion, independent of measurement of the AIF, were compared in healthy volunteers. Their performance was investigated against a reference kinetic

[1] Due to partial volume effects, accurate quantification is not possible, although there are partial volume correction techniques to overcome this. However, a great challenge for these techniques is accurate segmentation of the arteries.

method, which did use the arterial input function. Moreover, the study included assessment of repeatability performance.

The study confirmed observations from earlier work that without the AIF relative quantification is still possible. The double integration method turned out to be best method for measuring relative CBF in terms of both agreement with the reference method and in terms of repeatability. Although none of the methods was able to provide an accurate estimate of absolute perfusion, the method proposed by Treyer et al. did provide a reasonable precision and could therefore be used to study changes in absolute CBF within the same subject over time.

The aim of the study in **chapter three** was, firstly, to derive the optimal plasma input kinetic model for dynamic O-(2-[^{18}F]fluoroethyl)-L-tyrosine ([^{18}F]FET) PET studies and, secondly, to use that model as a reference to evaluate simplified methods. [^{18}F]FET is a PET tracer used in cancer imaging, in particular for brain tumours. The tracer kinetics reflect uptake of amino acids necessary for cell membrane synthesis and thus indirectly reflect tissue growth. Although the transport mechanisms of the tracer are still not entirely understood, the tracer has been shown to be very sensitive in detecting neoplasia in the brain and is therefore useful in determining the tumour extent in glioma.

In seven patients with diffuse glioma, three well-known metabolite corrected plasma input models were evaluated and the optimal model was determined to be the reversible two-tissue compartment model. Agreement with the optimal model was assessed for various simplified methods, including approaches already often used in [^{18}F]FET PET studies in glioma: the standardized uptake value (SUV) and SUV ratio. An important finding in the study is that, in terms of this agreement, later time point imaging (60-90 min post injection) outperforms currently recommended uptake time (20-40 min post injection).

In **chapter four**, the data from chapter three was processed at the voxel level (as opposed to region level). This process is called parametric mapping and is useful for evaluation of tracer uptake distribution within a tumour or to be able to delineate the tumour extent. In the study, several parametric methods and SUV ratio maps were compared in terms of accuracy (when compared to results from chapter 3) and map noise level. Both plasma input methods and reference tissue methods were included. (Non-invasive) Logan graphical analysis provided volume of distribution (ratio) maps with the lowest level of noise, but poor accuracy, while the basis function implementations provided the best accuracy, but also high noise levels. SUV ratio maps provided better results if later interval times were used, i.e., 60–90 min instead of 20–40 min, leading to lower bias (2.9% vs. 10.8%, respectively) and less noise in the map (12.8% vs. 14.4%).

Chapter five investigated the precision of image-derived arterial input functions obtained with dynamic contrast enhanced MRI in head and neck cancer patients. The arterial input function (AIF) is necessary to estimate pharmacokinetic parameters with dynamic contrast enhanced MRI. The AIF can be measured within the image (image-derived), or a population averaged AIF can be used. An *accurate* patient-specific measurement is preferred over the population average AIF, because it can account for the variability in cardiac output between patients.

However, the results show that accurate measurement of an image-derived AIF is unlikely in the head and neck region. AIFs obtained from different arteries in the head and neck region in the same patient differ in both magnitude and shape. This in turn affects the estimation of pharmacokinetic parameters, which differed significantly from those estimated using of a population averaged AIF. Usage of the population averaged AIF is therefore recommended.

The intravoxel incoherent motion (IVIM) model for diffusion-weighted imaging may provide useful biomarkers for disease management in head and neck cancer. Using data from healthy volunteers, **chapter six** compared the repeatability of three IVIM fitting methods to the conventional nonlinear least-squares regression: Bayesian probability estimation, a recently introduced neural network approach, and a modified version of the neural network.

The Bayesian and neural network approaches substantially outperformed conventional nonlinear regression in terms of test-retest repeatability. The processing speed of the neural network makes it viable for use in clinical practice. However, repeated training of networks on the same imaging data gives inconsistent results. Our presented modifications improve the neural network approach in this regard; however, the approach needs to be further improved to identify neural networks that are both consistent and precise.

Finally**, Chapter seven** shows that [^{18}F]FDG PET radiomics provides additional prognostic value when combined with clinical information and first-order [^{18}F]FDG PET imaging biomarkers. After extraction, 434 radiomics features were filtered for redundancy and combined into 8 latent factors. The results show that these factors can improve prediction of recurrence, distant metastasis and overall survival. Moreover, the study shows how this information can be used for personalized risk-stratification of patients' outcome. Better prognosis prediction in locally advanced head and neck squamous cell carcinoma can thus optimize personalized cancer care.

Acknowledgements

Thanks to all patients and their families for taking part in the Prediction and Frontier studies. For their hard work during the studies, thanks to the medical technicians, lab technicians and planning officers at the VUmc radiology and nuclear medicine department.

For their advice and assistance, thanks to coauthors Oliver Gurney-Champion, Dennis Heijtel, Aart Nederveen and Daniel Noij; members of the Frontier group: Petra Pouwels, Robert Schuit, Niels Verburg, Pieter Wesseling, Bert Windhorst and Philip de Witt Hamer; members of the Prediction group: Richard Ayres, Pim de Graaf, Otto Hoekstra, Adriaan Lammertsma, Cristina Lavini, Marije Vergeer, Johannes Rijken and Georgy Shakirin.

For our collaboration during the RNC course, thanks to Jan Boesten, Gerard Visser, Tjaard Weijer and Gertrüd Warmerdam.

For the great work during the organization of the iQC, thanks to Marcel van Schie, Lena Vaclavu, Martin Visser and Jing Zuo.

For all the good times, writing goals and great food, thanks to all friends, colleagues, J- and F-wing dwellers, Agraphia club members and Food club eaters: Hannan Ababri, Sophie Adriaanse, Margot Bleeker, Max Blokker, Marlies van den Born, Coreline Burggraaff, José Castillo, Kevin Cheng, Matthijs Cysouw, Bart de Vries, Corinne Eertink, Sandeep Golla, Hans Harms, Fiona Heeman, Dennis Heijtel, Nikie Hoetjes, Paul Horstman, Lieke Hoyng, Marc Huisman, Ramsha Iqbal, Emerson Itikawa, Bernard Jansen, Yvonne Jauw, Cemille Karga, Han Keijzers, Gem Kramer, Arthur van Lingen, Nikos Makris, Syahir Mansor, Adinda Mieras, Lino Miltenburg, Emma Mulder, Fasco van Ommen, Simone Pieplenbosch, Merlijn van der Plas, Remi Schmeits, Mette Stam, Hayel Tunçel, Sanne van Velzen, Floris van Velden, Eline Verwer, Kars van der Weijden, Sanne Wiegers and Leah Wilk.

For guiding me towards pursuing a PhD, thanks to Theo Faes and Jan de Munck.

Special thanks to my clinical counterpart Roland Martens. It was a pleasure working together.

Special thanks to my promotors and co-promotors; Ronald Boellaard, who kept everything going; Jonas Castelijns, who taught me the clinical picture; Maqsood Yaqub, who solved all the problems; and Tim Marcus, who remembered all the details.

List of publications

Thomas Koopman, Maqsood Yaqub, Dennis F.R. Heijtel, Aart J. Nederveen, Bart N.M. van Berckel, Adriaan A. Lammertsma, and Ronald Boellaard.
Semi-quantitative cerebral blood flow parameters derived from non-invasive [^{15}O]H_2O PET Studies.
2017. *Journal of Cerebral Blood Flow & Metabolism* 39(1):163–72.
doi: 10.1177/0271678X17730654.

Thomas Koopman, Niels Verburg, Robert C. Schuit, Petra J.W. Pouwels, Pieter Wesseling, Albert D. Windhorst, Otto S. Hoekstra, Philip C. de Witt Hamer, Adriaan A. Lammertsma, Ronald Boellaard, and Maqsood Yaqub.
Quantification of O-(2-[^{18}F]Fluoroethyl)-L-Tyrosine kinetics in glioma.
2018. *EJNMMI Research* 8(1):72. doi: 10.1186/s13550-018-0418-0.

Thomas Koopman, Niels Verburg, Petra J.W. Pouwels, Pieter Wesseling, Otto S. Hoekstra, Philip C. de Witt Hamer, Adriaan A. Lammertsma, Maqsood Yaqub, and Ronald Boellaard.
Quantitative parametric maps of O-(2-[^{18}F]Fluoroethyl)-L-Tyrosine kinetics in diffuse glioma.
2020. *Journal of Cerebral Blood Flow & Metabolism* 40(4):895–903.
doi: 10.1177/0271678X19851878.

Thomas Koopman, Roland M. Martens, Cristina Lavini, Maqsood Yaqub, Jonas A. Castelijns, Ronald Boellaard, and J. Tim Marcus.
Repeatability of arterial input functions and kinetic parameters in muscle obtained by dynamic contrast enhanced MR imaging of the head and neck.
2020. *Magnetic Resonance Imaging* 68:1–8. doi: 10.1016/j.mri.2020.01.010.

Thomas Koopman, Roland Martens, Oliver J. Gurney-Champion, Maqsood Yaqub, Cristina Lavini, Pim de Graaf, Jonas A. Castelijns, Ronald Boellaard, and J. Tim Marcus.
Repeatability of IVIM biomarkers from diffusion-weighted MRI in head and neck: bayesian probability versus neural network.
2021. *Magnetic Resonance in Medicine* 85(6):3394–3402. doi: 10.1002/mrm.28671.

Roland M. Martens, Daniel P. Noij, Meedie Ali, Thomas Koopman, J. Tim Marcus, Marije R. Vergeer, Henrica de Vet, Marcus C. de Jong, C. René Leemans, Otto S. Hoekstra, Remco de Bree, Pim de Graaf, Ronald Boellaard, and Jonas A. Castelijns.
Functional imaging early during (chemo)radiotherapy for response prediction in head and neck squamous cell carcinoma; a systematic review.
2019. *Oral Oncology* 88:75–83. doi: 10.1016/j.oraloncology.2018.11.005.

Roland M. Martens, Daniel P. Noij, Thomas Koopman, Gerben J. Zwezerijnen, Martijn W. Heymans, Marcus C. de Jong, Otto S. Hoekstra, Marije R. Vergeer, Remco de Bree, C. René Leemans, Pim de Graaf, Ronald Boellaard, and Jonas A. Castelijns.
Predictive value of quantitative diffusion-weighted imaging and ^{18}F-FDG-PET in head and neck squamous cell carcinoma treated by (chemo)radiotherapy.
2019. *European Journal of Radiology* 113:39–50. doi: 10.1016/j.ejrad.2019.01.031.

Roland M. Martens, Thomas Koopman, Daniel P. Noij, Remco de Bree, Marije R. Vergeer, Gerben J. Zwezerijnen, C. René Leemans, Pim de Graaf, Ronald Boellaard, and Jonas A. Castelijns.
Adherence to pretreatment and intratreatment imaging of head and neck squamous cell carcinoma patients undergoing (chemo)radiotherapy in a research setting.
2021. *Clinical Imaging* 69:82–90. doi: 10.1016/j.clinimag.2020.06.047.

Roland M. Martens, Thomas Koopman, Cristina Lavini, Meedie Ali, Carel F.W. Peeters, Daniel P. Noij, Gerben J. Zwezerijnen, J. Tim Marcus, Marije R. Vergeer, C. René Leemans, Remco de Bree, Pim de Graaf, Ronald Boellaard, and Jonas A. Castelijns.
Multiparametric functional MRI and ^{18}F-FDG-PET for survival prediction in patients with head and neck squamous cell carcinoma treated with (chemo)radiation.
2021. *European Radiology* 31(2):616–28. doi: 10.1007/s00330-020-07163-3.

Roland M. Martens, Thomas Koopman, Daniel P. Noij, Elisabeth Pfaehler, Caroline Übelhör, Sughandi Sharma, Marije R. Vergeer, C. René Leemans, Otto S. Hoekstra, Maqsood Yaqub, Gerben J. Zwezerijnen, Martijn W. Heymans, Carel F.W. Peeters, Remco de Bree, Pim de Graaf, Jonas A. Castelijns, and Ronald Boellaard.
Predictive value of quantitative ^{18}F-FDG-PET radiomics analysis in patients with head and neck squamous cell carcinoma.
2020. *EJNMMI Research* 10(1):102. doi: 10.1186/s13550-020-00686-2.

Roland M. Martens, Ruud van der Stappen, Thomas Koopman, Daniel P. Noij, Emile F. Comans, Gerben J. Zwezerijnen, Marije R. Vergeer, C. René Leemans, Remco de Bree, Ronald Boellaard, Jonas A. Castelijns, and Pim de Graaf.
The additional value of ultrafast DCE-MRI to DWI-MRI and ^{18}F-FDG-PET to detect occult primary head and neck squamous cell carcinoma.
2020. *Cancers* 12(10):2826. doi: 10.3390/cancers12102826.

Roland M. Martens, Thomas Koopman, Cristina Lavini, Tim van de Brug, Gerben J. Zwezerijnen, J. Tim Marcus, Marije R. Vergeer, C. René Leemans, Remco de Bree, Pim de Graaf, Ronald Boellaard, and Jonas A. Castelijns.
Early response prediction of multiparametric functional MRI and ^{18}F-FDG-PET in patients with head and neck squamous cell carcinoma treated with (chemo)radiation.
2022. *Cancers* 14(1):216. doi: 10.3390/cancers14010216.

Daniel P. Noij, Roland M. Martens, Gerben J. Zwezerijnen, Thomas Koopman, Remco de Bree, Otto S. Hoekstra, Pim de Graaf, and Jonas A. Castelijns.
Diagnostic value of diffusion-weighted imaging and ^{18}F-FDG-PET/CT for the detection of unknown primary head and neck cancer in patients presenting with cervical metastasis.
2018. *European Journal of Radiology* 107:20–25. doi: 10.1016/j.ejrad.2018.08.009.

Daniel P. Noij, Roland M. Martens, Thomas Koopman, Otto S. Hoekstra, Emile F. Comans, Gerben J. Zwezerijnen, Remco de Bree, Pim de Graaf, and Jonas A. Castelijns.
Use of diffusion-weighted imaging and ^{18}F-Fluorodeoxyglucose positron emission tomography combined with computed tomography in the response assessment for (chemo)radiotherapy in head and neck squamous cell carcinoma.
2018. *Clinical Oncology* 30(12):780–92. doi: 10.1016/j.clon.2018.09.007.

Niels Verburg, Thomas Koopman, Maqsood Yaqub, Otto S. Hoekstra, Adriaan A. Lammertsma, Lothar A. Schwarte, Frederik Barkhof, Petra J.W. Pouwels, Jan J. Heimans, Jaap C. Reijneveld, Annemieke J.M. Rozemuller, William P. Vandertop, Pieter Wesseling, Ronald Boellaard, and Philip C. de Witt Hamer.
Direct comparison of [^{11}C] Choline and [^{18}F] FET PET to detect glioma infiltration: a diagnostic accuracy study in eight patients.
2019. *EJNMMI Research* 9(1):57. doi: 10.1186/s13550-019-0523-8.

Niels Verburg, Thomas Koopman, Maqsood Yaqub, Otto S. Hoekstra,
Adriaan A. Lammertsma, Frederik Barkhof, Petra J.W. Pouwels, Jaap C. Reijneveld,
Jan J. Heimans, Annemarie J.M. Rozemuller, Anne M.E. Bruynzeel, Frank Lagerwaard,
William P. Vandertop, Ronald Boellaard, Pieter Wesseling,
and Philip C. de Witt Hamer.
Improved detection of diffuse glioma infiltration with imaging combinations: a diagnostic accuracy study.
2020. *Neuro-Oncology* 22(3):412–22. doi: 10.1093/neuonc/noz180.

Niels Verburg, Floris P. Barthel, Kevin J. Anderson, Kevin C. Johnson,
Thomas Koopman, Maqsood Yaqub, Otto S. Hoekstra, Adriaan A. Lammertsma,
Frederik Barkhof, Petra J.W. Pouwels, Jaap C. Reijneveld, Annemieke J.M. Rozemuller,
Jeroen A.M. Beliën, Ronald Boellaard, Michael D. Taylor, Sunit Das, Joseph F. Costello,
William P. Vandertop, Pieter Wesseling, Philip C. de Witt Hamer,
and Roel G.W. Verhaak.
Spatial concordance of DNA methylation classification in diffuse glioma.
2021. *Neuro-Oncology* 23(12):2054–65. doi: 10.1093/neuonc/noab134.

Portfolio of education

	Date achieved	ECTS
Courses		
Radiation Safety level 5B	13-11-2015	2.00
PET tracer pharmacokinetics and data analysis procedures	18-11-2015	2.00
Research Integrity	02-12-2015	2.00
institute QuantiVision Winter school 2016 – Quantitative Analysis of Medical Images	26-02-2016	1.50
Machine Learning	01-04-2016	6.00
CCA neXt – 2nd tri-annual meeting	21-06-2016	0.05
Imaging Technology Summer Workshop – TOPIM-TECH 2016 – Multiscale & Multiparametric Imaging	15-07-2016	2.00
Basiscursus oncologie – Introductie tot de Klinische en Fundamentele Oncologie	24-03-2017	2.00
PET Pharmacokinetics Course 2017	31-03-2017	2.00
Organisation of institute QuantiVision Conference 2017	21-08-2017	2.00
institute QuantiVision Winter school 2018 – Machine Learning Applied to Quantitative Analysis of Medical Images	09-03-2018	1.50
Research group meetings	01-01-2019	2.50
Writing scientific articles under supervision of a senior scientist	01-01-2019	2.00
Meetings department Imaging methodology group	01-01-2019	2.50
Statistics exemption	07-02-2019	0.00
Research related		
European Molecular Imaging Meeting	10-03-2016	0.85
Hands-On MRI course on Head & Neck Imaging	19-03-2016	0.45
EANM'16	19-10-2016	2.00
institute QuantiVision Conference 2017	24-02-2017	0.30
Brain & Brain PET 2017	04-04-2017	2.00
Medical Imaging Symposium for PhD students 2017	19-05-2017	0.30
Medical Imaging Symposium for PhD Students 2018	24-05-2018	0.30
Teaching/Student supervision		
Radiation safety level 5B	15-11-2017	3.25
Advanced Medical Technology – Tracer Kinetic Modelling	25-05-2018	0.25
Master thesis Daniëlle de Jong: Optimisation of Arterial Input Function extraction in Dynamic Contrast Enhanced MRI	16-08-2018	1.50
		41.25

About the author

Thomas Koopman was born on 8 March 1992 in Huizen, the Netherlands. After completing secondary school in 2010, he started studying medical natural sciences at the Vrije Universiteit Amsterdam. During his study, he was an active member of the study association Mens. He completed the master track Medical Physics after working on Monte Carlo simulations for absorbed dose calculation of radionuclides at the department of radiology and nuclear medicine of the Vrije Universiteit Medical Center and evaluating supervised machine-learning algorithms for the segmentation of brain tissue in magnetic resonance images at the Image Sciences Institute of the University Medical Center Utrecht.

In 2015, he started his research doctorate under supervision of prof.dr. Boellaard and prof.dr. Castelijns and the guidance of dr.ir. Yaqub and dr. Marcus. He worked together with Roland Martens in the collection and analysis of the multimodal and multiparametric images of the Prediction study and worked on the analysis of PET images in the Frontier study. In 2017, he helped organize the first institute Quantivision Conference. During his time at the VUmc, he discovered his enthusiasm for programming.

In 2020, Thomas continued with medical image processing and started working as a researcher and developer at the Research and Trials department of Quest Medical Imaging, part of Olympus, in Middenmeer. Here, he helps to build multispectral imaging solutions for, for instance, fluorescence guided surgery.