Computer Aided Medical Procedures
Prof. Dr. Nassir Navab

Dissertation

Radiotracer uptake classification using deep learning for evaluation of image-derived cancer biomarkers in PET/CT

Nicolò Capobianco

Fakultät für Informatik
Technische Universität München

Technische Universität München
Fakultät für Informatik

Radiotracer uptake classification using deep learning for evaluation of image-derived cancer biomarkers in PET/CT

Nicolò Capobianco

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r):	Prof. Dr.-Ing. Pramod Bhatotia
Prüfer der Dissertation:	Prof. Dr. Nassir Navab
	Priv.-Doz. Dr. Stephan Nekolla
	Prof. Dr. Michele Piana

Die Dissertation wurde am 14.10.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 27.04.2022 angenommen.

Nicolò Capobianco
Radiotracer uptake classification
using deep learning for evaluation of
image-derived cancer biomarkers in PET/CT
Dissertation, Version 1.0

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Informatikanwendungen in der Medizin
Boltzmannstraße 3
85748 and Garching bei München

Abstract

In the care of patients with cancer, knowledge of the disease spread and overall burden is a key factor in the clinical decision-making process. Medical imaging is routinely used as noninvasive method to aid diagnosis and staging as well as treatment selection, planning, and monitoring. In addition to qualitative image inspection by physicians, standardized categorical and quantitative variables are increasingly used for accurate and reproducible image evaluation. Nevertheless, the manual determination of detailed image-derived parameters can require laborious examination and maneuvering, thus being time-consuming, error-prone, and operator dependent. Conversely, automated image analysis methods can be employed to improve robustness and repeatability. Notably, deep learning techniques have demonstrated high accuracy in automating visual tasks and are promising for enabling image-derived biomarkers in oncology.

This dissertation presents contributions to image analysis methods in positron emission tomography (PET) / computed tomography (CT) and their application in determining image-derived biomarkers in oncology. The first contribution consists in a deep learning method for automated classification of elevated tracer uptake regions, trained for multiple radiotracers. PET/CT image regions are classified as uptake suspicious or nonsuspicious for cancer. Additionally, high-uptake regions are classified with respect to their anatomical location. Results indicate that a convolutional neural network can classify high uptake sites in good agreement with the visual assessment by a physician. The second contribution consists in the application of automated uptake classification in Prostate Specific Membrane Antigen (PSMA)-ligand PET/CT to assess prostate cancer stage. Results indicate that the proposed method can be used to determine the image-based nodal and metastatic stage according to a standardized reporting framework in concordance with expert reader evaluation. The third contribution consists in the application of automated uptake classification in ^{18}F-Fluorodeoxyglucose (FDG) PET/CT images to estimate total metabolic tumor volume (TMTV) in lymphoma patients. Results show that in the examined cohort the proposed method determined an estimation of baseline TMTV significantly correlated and having comparable prognostic value with the TMTV estimation obtained semi-automatically by clinicians.

The presented methods for automated analysis of PET/CT images using deep learning show promising results for supporting physicians in evaluating cancer stage and overall burden. Moreover, the investigated methods could potentially allow clinicians to identify novel biomarkers based on the rich information extracted from whole-body tumor assessment in PET/CT images. In conclusion, deep learning algorithms may aid the evaluation of informative and actionable image-derived cancer biomarkers and, together with diligent validation in multicenter trials, drive their establishment in the clinical routine.

Zusammenfassung

Bei der Versorgung von Krebspatienten ist die Kenntnis der Krankheitsausbreitung und Gesamtbelastung ein Schlüsselfaktor im klinischen Entscheidungsprozess. Medizinische Bildgebung wird routinemäßig als nichtinvasive Methode zur Unterstützung der Diagnose und des Stagings sowie der Behandlungsauswahl, -planung und -überwachung verwendet. Neben der qualitativen Bildbetrachtung durch den Arzt werden zunehmend standardisierte kategoriale und quantitative Parameter zur genauen und reproduzierbaren Bildauswertung eingesetzt. Nichtsdestotrotz kann die manuelle Bestimmung detaillierter, bildabgeleiteter Merkmale ein mühsames, arbeitsintensives und zeitaufwendiges Verfahren sein, das fehleranfällig und bedienerabhängig ist. Umgekehrt können automatisierte Bildanalyseverfahren eingesetzt werden, um die Robustheit und Wiederholbarkeit zu verbessern. Insbesondere Deep-Learning-Techniken haben eine hohe Genauigkeit bei der Automatisierung visueller Aufgaben gezeigt und sind vielversprechend für die Ermöglichung bildbasierter Biomarker in der Onkologie.

Diese Dissertation beinhaltet Beiträge zu bildanalytischen Verfahren in der Positronen-Emissions-Tomographie (PET) / Computertomographie (CT) und deren Anwendung bei der Bestimmung bildbasierter Biomarker in der Onkologie. Der erste Beitrag besteht in einer Deep-Learning-Methode zur automatisierten Klassifizierung von Regionen mit erhöhter Traceranreicherung, trainiert für mehrere Radiotracer. PET/CT-Bildbereiche werden als krebsverdächtige oder nicht krebsverdächtige Strukturen klassifiziert. Darüber hinaus werden Regionen mit hoher Aufnahme hinsichtlich ihrer anatomischen Lage klassifiziert. Die Ergebnisse deuten darauf hin, dass ein neuronales Faltungsnetzwerk in guter Übereinstimmung mit der visuellen Beurteilung durch einen Arzt Orte mit hoher Aufnahme klassifizieren kann. Der zweite Beitrag besteht in der Anwendung der automatisierten Aufnahmeklassifikation bei PET/CT mit dem prostataspezifischen Membranantigen (PSMA)-Liganden zur Beurteilung des Prostatakrebsstadiums. Die Ergebnisse zeigen, dass die vorgeschlagene Methode verwendet werden kann, um das bildbasierte nodale und metastatische Stadium gemäß eines standardisierten Berichtsrahmens in Übereinstimmung mit einer Expertenbewertung durch Befunder zu bestimmen. Der dritte Beitrag besteht in der Anwendung der automatisierten Aufnahmeklassifikation in 18F-Fluorodesoxyglucose (FDG) PET/CT-Bildern zur Schätzung des metabolischen Gesamttumorvolumens (TMTV) bei Lymphompatienten. Die Ergebnisse zeigen, dass die vorgeschlagene Methode in der untersuchten Kohorte eine Schätzung des TMTV-Ausgangswerts erlaubt, die signifikant korreliert ist und einen vergleichbaren prognostischen Wert ergibt wie die halbautomatische, von Klinikern ermittelte TMTV-Schätzung.

Die vorgestellten Methoden zur automatisierten Analyse von PET/CT-Bildern mittels Deep Learning zeigen vielversprechende Ergebnisse, um Ärzte bei der Beurteilung des Krebsstadiums und der Tumorgesamtbelastung zu unterstützen. Darüber hinaus könnten die untersuchten Methoden es Klinikern möglicherweise ermöglichen, neue Biomarker basierend auf den

umfangreichen Informationen zu identifizieren, die mit Hilfe der Ganzkörper-PET/CT gewonnen werden kann. Zusammenfassend lässt sich sagen, dass Deep-Learning-Algorithmen die Bewertung informativer und umsetzbarer bildbasierter Krebsbiomarker unterstützen und zusammen mit einer sorgfältigen Validierung in multizentrischen Studien ihre Etablierung in der klinischen Routine vorantreiben können.

Acknowledgments

I would like to sincerely thank my supervisors Nassir Navab and Stephan Nekolla for their advice and support which greatly enabled my work. My strong gratitude goes to all the clinical and research team at the Nuklearmedizinische Klinik of the Klinikum rechts der Isar. I would like to thank Esteban Lucas Solari and Andrei Gafita for the stimulating discussions and cooperation as well as Matthias Eiber, Wolfgang Weber and colleague physicians for the valuable collaboration.

I am extremely grateful to all the participants of the HYBRID network for envisioning and realizing an outstanding framework for fostering scientific collaboration and training of early-stage researchers. I would like to thank the hosts and collaborators of my secondments, including Irène Buvat, for her invaluable guidance and support, Michel Meignan, Thomas Beyer, and Ivo Rausch. The HYBRID Innovative Training Network, including the project presented in this dissertation, has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska–Curie grant agreement (no. 764458). I am grateful to the organizers of this program and the European Union.

I am deeply thankful to all the people at Siemens who supported my work and the HYBRID network, including Christine Lorenz, for shaping an exciting project and welcoming me with superb kindness, Hartwig Newiger, for the exceptional support and guidance, Sven Zuehlsdorff, for the precious supervision and encouragement, Bruce Spottiswoode, for the valuable mentorship and enthusiasm, Ludovic Sibille, for the great advice and collaboration, Vijay Shah, for the stimulating conversations and teamwork, Guenther Platsch, for sharing his knowledge and insights, and many others.

My heartfelt thank you goes to Annalisa for being by my side and encouraging me, filling my days with laughter and joy. I would like to deeply thank Gian Franco for the inestimable friendship and support, his friendliness and wisdom are immensely inspiring. My gratitude goes to Flavio and Marco for sharing many enjoyable moments. Finally, I would like to sincerely thank my family for their precious support and cheerfulness.

Contents

Part I

Introduction

1 Introduction

Cancer is one of the leading causes of death worldwide [80]. Information on cancer genesis, molecular profile, metabolism, progression, and response to external stimuli is fundamental for developing effective treatments as well as tailoring and implementing them for each patient. The analysis of tissue samples obtained via biopsy or surgery allows ex-vivo evaluation at cellular level and is often considered as gold standard examination. Nevertheless, this sampling method can only provide a limited coverage of organs and tissues, it is not systematically applicable in all cases and can cause side-effects. Medical imaging, on the other hand, can be used to noninvasively assess defined tissue properties in vivo with a larger sampling space up to the entire body. Technological advancements in the imaging field have enabled to approach or surpass millimetric resolution, increase sensitivity and in some applications to record multiple tissue properties and their evolution over time within a single examination, creating a wealth of information. In addition to qualitative visual image inspection by physicians, categorical and quantitative parameters are increasingly being used to characterize disease as image-derived biomarkers. Therefore, accurate and reproducible image analysis methods are necessary to enable the identification and clinical application of robust and repeatable biomarkers. Notably, machine learning methods have recently demonstrated high accuracy in automating visual tasks and their application in medical image analysis is a promising enabler of informative and actionable image-derived biomarkers.

1.1 Contributions

This dissertation focuses on methods for the analysis of positron emission tomography (PET) / computed tomography (CT) images aimed at identifying the presence of sites suspicious for cancer and determining their anatomical localization, with applications in assessing the disease spread and overall burden. Contributions are related to algorithm development and evaluation of the presented methods for clinical applications. First, a convolutional neural network is described for the classification of foci with elevated radiotracer uptake as nonsuspicious or suspicious for cancer and the classification of their anatomical location. Methods to leverage PET/CT data obtained with different radiotracers for training purposes are presented, including the use of transfer learning and a dedicated network architecture for concurrent training. Second, the application of the presented algorithm is assessed for the automated identification of prostate cancer stage according to a standardized reporting framework, showing fair concordance with physicians. Third, the application of uptake classification is assessed for the automated estimation of baseline total metabolic tumor volume in lymphoma patients, showing comparable prognostic value with estimates obtained manually by physicians.

1.2 Outline

In this dissertation essential background information is first introduced.

- Section 1.3 introduces the medical imaging modalities discussed in the thesis. Fundamental aspects of underlying principles, data acquisition and quantification are described.

- Section 1.4 provides background information on imaging biomarkers in PET/CT for applications in oncology. Radiotracers commonly used in clinical examinations are concisely described. Essential aspects of staging based on PET/CT imaging are introduced. Established and emerging semiquantitative parameters are presented.

- Section 1.5 summarizes the main elements of image classification related to the contributions presented in the dissertation. Convolutional neural networks, transfer learning techniques, and evaluation metrics are described.

Contributions to images analysis methods and their use for clinical applications are then presented and discussed.

- Chapter 2 presents contributions in image analysis methods for classification of foci with elevated tracer uptake. Strategies to train a convolutional neural network by leveraging PET/CT data obtained with different radiotracers are described and shown to increase performance when the availability of image data and expert-annotated ground truth is limited.

- Chapter 3 describes the use of uptake classification for automated assessment of prostate cancer nodal and metastatic stage according to a standardized reporting framework. Results are compared to the ones indicated by physicians showing fair concordance.

- Chapter 4 presents the use of uptake classification for automated estimation of total metabolic tumor volume. The method is evaluated with a retrospective analysis in a cohort of lymphoma patients enrolled in a multi-center trial. Baseline metabolic tumor volume estimates obtained automatically with the proposed method and semi-automatically by clinicians are shown to be significantly correlated and have comparable prognostic value for progression free survival and overall survival.

Significant parts of the results presented in the dissertation have been published. The corresponding publications are indicated at the beginning of the respective chapters and listed in the Appendix. While the realization of the contributions described was conducted by the author of this thesis, the first-person plural is occasionally used to reflect a collective team effort. Finally, chapter 5 concludes the dissertation by discussing the results presented in the thesis and directions for future research.

1.3 Medical imaging with PET/CT

This section introduces fundamental aspects of the medical imaging modalities discussed in the dissertation. The purpose is to provide an essential understanding to readers unacquainted with positron emission tomography / computed tomography, while references to more detailed descriptions are indicated. Physical principles of image acquisition as well as quantification characteristics are concisely described. Medical images are typically obtained by measuring the energy emerging from the human body as result of either natural processes or the interaction with artificial stimuli. In several modalities employed in clinical routine, images are formed by acquiring an electromagnetic or mechanical radiation signal generated by an artificial source, external or introduced into the subject, and modulated by the human body. Medical imaging allows the unique possibility to measure defined properties of organs and tissues in multiple spatial dimensions and detect as well as monitor alterations. Such capability has resulted in numerous clinical applications from diagnosis to staging as well as treatment selection, planning, and monitoring.

1.3.1 X-ray Computed Tomography

Computed Tomography (CT) is a type of transmission imaging in which X-ray electromagnetic radiation, typically within an energy window of $50\,\mathrm{keV} \leq h\nu \leq 140\,\mathrm{keV}$ for diagnostic imaging [10], is generated by an external source and directed to the target object. By rotating the source around the target object and measuring the emerging radiation with an array of detectors, the spatial distribution of the object attenuation coefficient is reconstructed, through the solution of an inverse problem, and can be visualized as a CT image. A schematic illustration of a conventional CT scanner is displayed in Fig. 1.1.

The X-ray source consists in a vacuum tube in which electrons are accelerated with an electric field between a cathode filament, where they are liberated via thermionic emission, and an anode disk, where they decelerate, resulting in the emission of X-ray radiation named bremsstrahlung with a continuous spectrum, and X-ray emission with sharp spectral lines characteristic of the anode material. The intensity of the generated radiation is controlled by the electron current, while the spectrum is related to the acceleration voltage and anode material. A metal filter is used to reduce low energy components and lower the patient dose.

As the X-ray photons traverse the target object, they interact with it through different absorption and scattering mechanisms, resulting in an attenuation of the radiation intensity. For a monochromatic X-ray and a homogenous medium, this can be modeled by an exponential reduction of the intensity as a function of the material thickness, with a fixed linear attenuation coefficient related to the material. For a monochromatic X-ray and an inhomogeneous medium, the attenuation is obtained by a negative exponential of the summed attenuation contributions of the volume elements traversed by the radiation. Each contribution consists in the linear attenuation coefficient of a volume element multiplied by its thickness. Thus, each radiation ray emerging from the target object carries defined information on the attenuation properties of the material it intersects, namely the integral of the linear attenuation coefficient along the ray line.

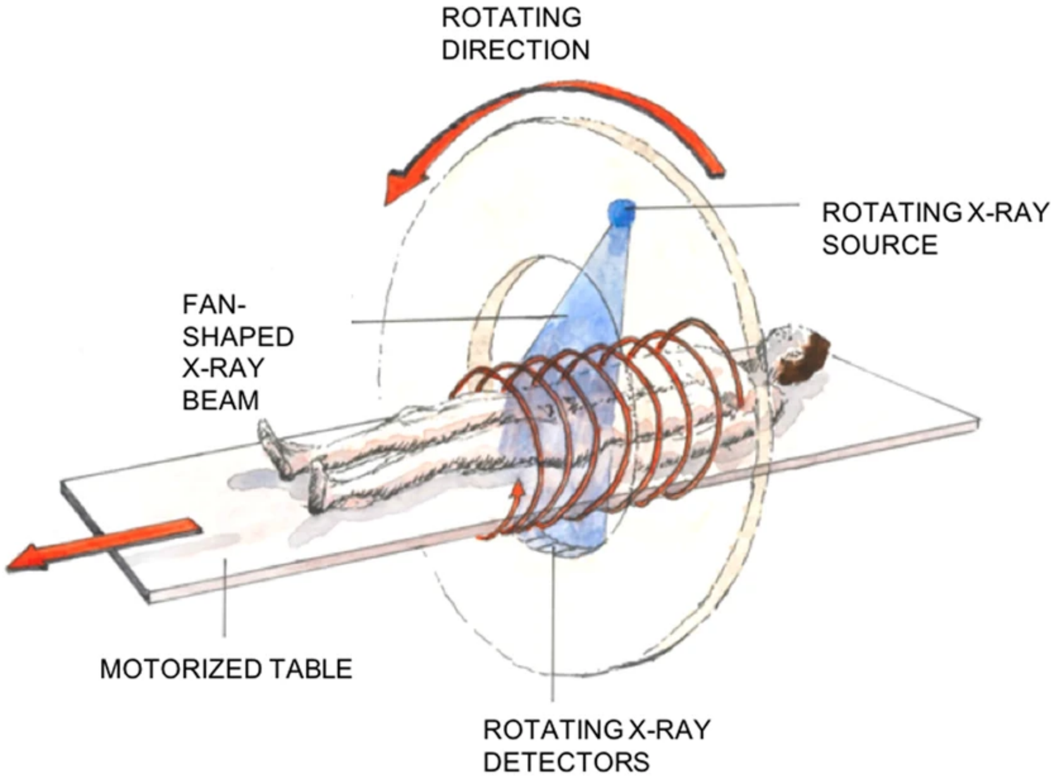


Fig. 1.1. Schematic illustration of an X-Ray Computed Tomography scanner. Adapted from [74].

The intensity of the radiation emerging from the object is measured by an array of detectors. Different detector designs can be employed in CT systems, including gaseous ionization detectors, scintillator detectors, and direct semiconductor detectors. In gas detectors, X-ray photons ionize a gas producing ions and free electrons in a chamber with an applied electric field, creating a measurable electric current. Scintillator detectors consist in a scintillator medium, in which X-ray photons are converted to lower energy photons, coupled with photodiodes, which convert the lower energy radiation intensity signal to an electric current signal. In direct semiconductor detectors, X-ray photons interact with a semiconductor material under an applied electric field producing hole-electron pairs generating an electric current. Direct semiconductor detectors have been recently employed in research on photon-counting computed tomography, aimed at exploiting the detection and energy measurement of individual X-ray photons to improve image quality. In all the detector types described, the measurement of the generated electric current allows to store and process the X-ray intensity information digitally.

In modern CT scanners the X-ray tube and detector block are mounted on a slip ring structure on which they can rotate and receive electrical power. Traditionally, the X-ray source generates a fan beam, and the resulting radiation is measured by a diametrically opposed detector arc. By rotating the source and detector arc, multiple one-dimensional intensity profiles are recorded at different source positions, of the radiation emerging from a slice of the scanned object. These intensity profiles can be interpolated and visualized in a two-dimensional image named sinogram (Fig. 1.2a), whose spatial dimensions correspond to the X ray angles and offsets in the slice plane. The sinogram values represent the integral of the linear attenuation coefficient for each line in the slice plane, an operation known as Radon transform. It is proven analytically that from a sinogram the spatial distribution of the attenuation coefficient can be reconstructed [93]. A common reconstruction method equivalent to the inverse Radon transform is Filtered Back Projection (FBP), which consists in applying a ramp filter to the measured one-dimensional attenuation profiles and summing the resulting values projected back along the corresponding X-ray lines for all angles. Different kernels can be used for FPB other than the ramp filter, with a different trade-off between image noise and spatial resolution. A separate class of reconstruction methods comprises iterative reconstruction (IR) algorithms. In IR methods, images are reconstructed by iterative optimization of an objective function guaranteeing data fidelity. Physical factors such as photon statistics, X-ray beam spectrum, and detector geometry can be incorporated in IR methods. For detailed descriptions of reconstruction algorithms we refer to Computed Tomography reviews [10][67].

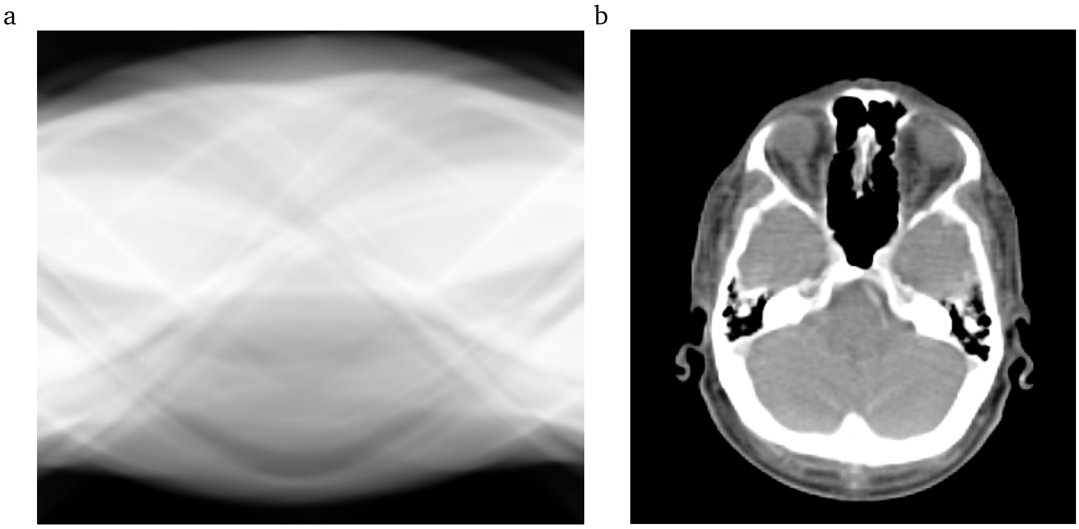


Fig. 1.2. Examples of (a) computed tomography sinogram and (b) reconstructed CT image. CT image courtesy of Klinikum rechts der Isar, Technical University of Munich, Munich, Germany.

As result of the reconstruction, a discretized two-dimensional linear attenuation coefficient map of an object slice is obtained, which can be visualized as a CT image (Figure 1.2b). Typically, multiple object slices are scanned resulting in a three-dimensional acquisition. The linear attenuation coefficient depends in general on the X-ray energy spectrum used. In medical imaging applications, CT images are expressed in units termed Hounsfield units (HU) by linearly rescaling the attenuation coefficient measured, such that the attenuation of water corresponds to $0\,\mathrm{HU}$ and zero attenuation corresponds to $-1000\,\mathrm{HU}$. Lung tissue and fat correspond to negative Hounsfield units, while muscles, connective tissue and most soft tissue organs correspond to positive Hounsfield units, and bone is related to higher values typically up to $2000\,\mathrm{HU}$. Medical CT scanners typically provide integer values between $-1024\,\mathrm{HU}$ and $3071\,\mathrm{HU}$. CT is routinely employed in a broad variety of clinical contexts including imaging of stroke, trauma, as well as oncologic, cardiovascular, abdominal and lung diseases with applications in diagnosis, treatment guidance and monitoring.

1.3.2 Positron Emission Tomography

Positron Emission Tomography (PET) is a type of emission imaging in which a radiopharmaceutical containing a radionuclide is injected in a scanned subject where it distributes, and the radionuclide decays by positron emission, generating pairs of gamma photons resulting from electron–positron annihilation. By measuring the number and coincidence time of gamma photons emerging from the scanned object with a ring of detectors, the spatial distribution of the radioactivity concentration within the subject is reconstructed, trough the solution of an inverse problem, and can be visualized as a PET image. A schematic illustration of a PET scanner is displayed in Fig. 1.3.

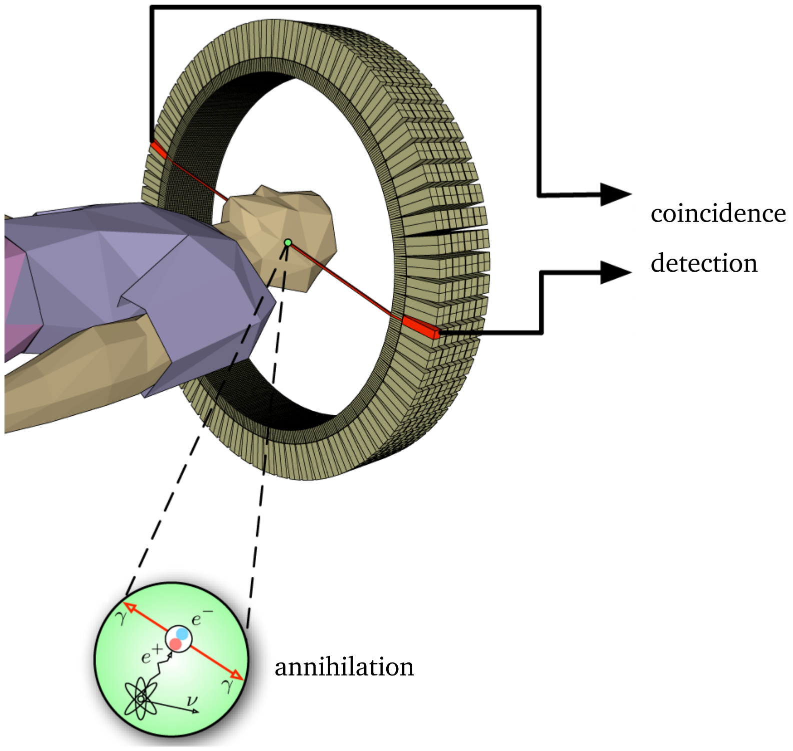


Fig. 1.3. Schematic illustration of a Positron Emission Tomography scanner. Adapted from [75].

PET radiopharmaceuticals are chemical substances containing a positron emitting radionuclide. In PET imaging, a radionuclide is typically bound to a ligand, forming a radiopharmaceutical compound which interacts with physiological and pathological processes once injected in a subject. By measuring the spatial distribution of radioactivity concentration within the subject through PET imaging, functional information can be assessed on a biological process of interest marked by the concentration of radiopharmaceutical. Radionuclides used in PET include ^{18}F, ^{11}C, ^{13}N, ^{15}O, ^{68}Ga, ^{82}Rb, ^{62}Cu among others and are produced through a cyclotron, reactor, or generator either directly or indirectly via production of a parent radionuclide. Radionuclide characteristics useful for PET imaging include a low fraction of non-positron decays, a short positron range, a half-life allowing to image the biologic process of interest without excessive radiation dose and with a viable radiotracer supply chain. Radiopharmaceuticals can be designed to act as analogues of biological molecules or bind to specific targets, allowing to assess a biological process of interest. Radiopharmaceutical properties useful for PET imaging include high affinity for the intended target as well as metabolic, kinetic, and excretion characteristics allowing to image the target of interest without interference from other processes and without excessive radiation dose. PET radiopharmaceuticals for clinical use in oncology are described in section 1.4.1.

As the radiotracer is distributed within the subject, the proton-rich radionuclides undergo one or multiple decay processes including positron decay, by which a proton is converted to a neutron with emission of a positron and a neutrino. The residual transition energy is converted in kinetic energy of the positron and the neutrino. The positron traverses the medium within a range which depends on its energy and the density of the medium, typically in the order of a few millimeters for PET radionuclides in water [103]. Subsequently, the positron interacts with an electron of the medium and both particles are annihilated resulting in the emission of two 511-keV photons in opposite directions, with an angle close to 180 degrees and depending on the residual positron momentum at annihilation. As the gamma photons traverse the object, they undergo different absorption and scattering processes causing attenuation of the radiation intensity, which can be accounted for by a total attenuation coefficient that depends on the photon energy and the traversed medium. Photon pairs which emerge from the subject without interaction and in close coincidence therefore carry defined information on the location of the originating annihilation and radiotracer decay.

The number of gamma photons emerging from the subject is measured by a ring of detectors. Modern PET scanners employ scintillation crystals in which an incident gamma photon is converted in lower energy photons. Different materials have been used as scintillation crystals in PET scanners, including bismuth germinate (BGO), lutetium oxyorthosilicate (LSO), and lutetium yttrium oxyorthosilicate (LYSO), each with different physical properties in terms of attenuation coefficient, photon yield, and scintillation decay time. Lower energy photons produced by the scintillation crystals are converted to electric signal and amplified either via a photomultiplier tube or, in more recent scanners, via a silicon photomultiplier. The conversion to lower energy photons and then electric signal allows to measure the count, energy, and timestamp of gamma photons detected as well as record and process this information digitally.

In modern PET scanners, detectors are arranged in an array of rings, with the scanned subject positioned at the center. By detecting gamma photon pairs in opposing detectors within a short time window named coincidence window, a line of response is defined connecting the two detectors, where an annihilation event may have occurred. In addition to cases where an annihilation actually occurred along the defined line of response, named true coincidences, other events can be detected along the same line of response due to random coincidence of unrelated annihilations, or due to the scattering of gamma photons. Different methods can be used to correct for random coincidences, including the subtraction of the count rate measured with a delayed coincidence window, which approximates the random coincidence rate. Dedicated methods can be used to correct for scatter events, including a Monte Carlo simulation of the scatter distribution from an initial estimate of the reconstructed activity distribution.

The true coincidence events measured for all the lines of response in a plane can be represented in a sinogram image (Figure 1.4a). In this image, each sinogram value represents the total number of annihilations occurred generating photon pairs along a line of response minus undetected annihilations due to the attenuation of gamma photons. After correcting for attenuation, the spatial distribution of annihilation events can be reconstructed through the solution of an inverse problem with analytic or iterative reconstruction methods. Moreover, the time of flight technique is employed in recent PET scanners, through which the measurement of the time difference in arrival of photon pairs at the detectors allows to estimate the distance they traveled and in which segment of the line of response the originating annihilation may have occurred. Including the time of flight information in the reconstruction allows to improve sensitivity and image quality, particularly in heavy subjects in which the amount of attenuation and scattering is higher. Further details of attenuation correction in PET/CT will be provided in section 1.3.3. For a detailed description of methods to correct for attenuation, random coincidences, and scatter, as well as image reconstruction algorithms we refer to Positron Emission Tomography reviews [103][2].

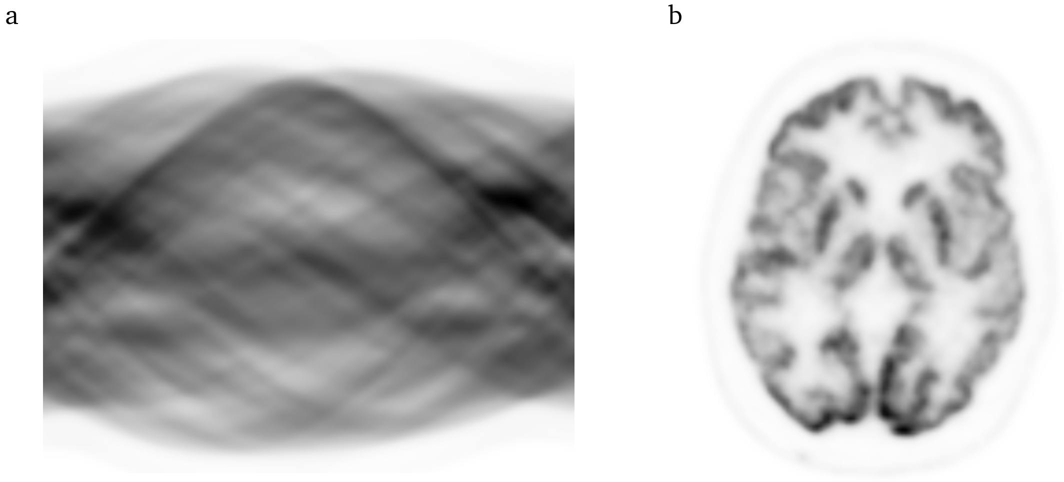


Fig. 1.4. Examples of (a) positron emission tomography sinogram and (b) reconstructed PET image. PET image courtesy of Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland.

Following image reconstruction, a discretized spatial map of positron emission events occurred during the acquisition is obtained, which can be visualized as a PET image (Fig. 1.4b). Typically, events occurring in multiple planes are measured by an array of detector rings, resulting in a three-dimensional acquisition. The spatial map of decay events within the subject carries defined information on the radiotracer distribution and interaction with physiological and pathological processes during the acquisition time. In static PET scans, image acquisition is performed at a fixed time delay after injection for which the highest image quality and reproducibility is obtained for the radiotracer and biologic process of interest. Assuming the radiotracer distribution is constant during the acquisition time, PET images can be expressed in units of radioactivity concentration, such as $\mathrm{Bq/mL}$ or equivalent units, calculated at the time of scan start by correcting for decay during the acquisition time. The measured radioactivity concentration is proportional to the concentration of radiopharmaceutical and its metabolites. However, since the administered dose, distribution volume, and kinetic behavior of the radiotracer are different for each PET scan, the measured units of radioactivity concentration are not a quantitative assessment of the biochemical process of interest and cannot be compared between scans. To partially correct for factors of inter-scan variability and describe the biological process of interest under simplistic assumptions, static PET images can be expressed in semiquantitative units, which are described in section 1.4.3. In dynamic PET scans, temporal variations in the radiotracer distribution are measured through an extended acquisition time. By describing the tracer kinetic behavior with compartmental models, kinetic parameters can be derived from the measured dynamic PET data, which are quantitative indicators of the biochemical process of interest. For a detailed description of dynamic PET techniques and tracer kinetic modeling methods we refer to dedicated reviews [2][94]. PET imaging is routinely employed for several clinical applications including cancer staging, response assessment in oncology, evaluation of cardiac viability and perfusion as well as examination of Parkinson's disease, Alzheimer's disease, and epilepsy.

1.3.3 Hybrid PET/CT

PET images are often visualized in combination with high-resolution CT or MR images, to facilitate anatomical localization of findings and support image interpretation. Integrated PET/CT scanners have been developed for combined acquisition with both imaging modalities in close spatial alignment, allowing to assess functional and structural information with a single machine and imaging procedure. Moreover, in PET/CT scanners the acquired CT data can be used for attenuation correction of PET images, more accurately and with a shorter scan time compared to other techniques used for attenuation correction.

In PET/CT scanners, both imaging units are mounted on the same support with the CT unit in front adjacent to the PET unit in the back. The centers of both imaging units are aligned such that the respective fields of view are separated only by a fixed axial displacement. Typically, in a PET/CT protocol the CT scan is first acquired in less than one minute, during which the patient is asked to perform breath hold or, if uncapable, shallow breathing, to minimize motion artifacts. Following the CT scan, the patient bed is translated, and the PET acquisition is performed. This allows to acquire co-registered PET and CT images in a single session, which can be visualized side by side or fused to support image interpretation.

The acquired CT data can be effectively used for attenuation correction of PET images. Typically, a single blank CT scan, acquired without a subject in the scanner, is used in combination with a CT scan of each patient to obtain the corresponding map of attenuation correction factors. Since the attenuation correction factor depends on the photon energy, a scaling is applied from the lower energy of CT X-ray radiation to the higher energy of PET gamma photons. The scaling factor is defined for a given tissue as ratio of the mass attenuation coefficient of 511-keV photons to the mass attenuation coefficient of X-ray photons at a single effective energy representing the CT spectrum, typically in the range $50–80\,\text{keV}$ [71]. A fixed set of scaling factors is applied, each one for a different tissue type and applied to all volume elements of the corresponding tissue type segmented from the CT image. The resulting map of attenuation coefficients for 511-keV photons is used in modern PET/CT scanners for PET attenuation correction as well as scatter correction via Monte Carlo simulation. Compared to other PET attenuation correction techniques which require a transmission scan using an external long-lived source of gamma photons and last several minutes to acquire sufficient counts, CT-based attenuation correction allows a shorter scan time. Depending on clinical requirements for the PET/CT examination, solely a low dose CT with reduced image quality may be acquired for the purpose of attenuation correction, or a higher dose CT with superior image quality may be acquired instead or in addition, to serve diagnostic purposes.

Factors affecting the CT image quality and mismatches between PET and CT data can cause artifacts in PET/CT images, for instance due to patient motion, truncation of the CT field of view, CT contrast agents and metal objects. In presence of PET/CT artifacts, the visualization of a PET image reconstructed without correcting for attenuation and scatter may be required to support correct image interpretation. Dedicated reconstruction algorithms and acquisition protocols have been developed to eliminate or limit the occurrence of artifacts in PET/CT [103][2].

1.4 PET/CT imaging biomarkers in oncology

This section introduces biomarkers used in PET/CT imaging for clinical applications in oncology. A biomarker is described as a defined characteristic measured as indicator of physiological processes, pathological processes, or biological response to an intervention or exposure [49]. As a plethora of management options is available in oncology, spanning from watchful waiting to localized and systemic treatments, accurate knowledge of cancer spread and progression is crucial for clinical decision making. Whole-body PET/CT imaging allows to obtain meaningful disease information, combining a functional assessment, with PET radiopharmaceuticals designed to interact with pathologic processes, and a structural assessment, with CT images reflecting variations in morphology and density. In this section PET/CT imaging biomarkers are presented, either established or emerging, for clinical use in oncology, including both different radiopharmaceuticals and image-based parameters used to evaluate the disease status and inform clinical decisions.

1.4.1 PET radiotracers in clinical oncology

Numerous radiopharmaceuticals have been developed and investigated to assess different biological processes related to cancer through PET imaging. However, a limited set of radiotracers is widely employed in clinical routine, in part due to challenges and costs associated with establishing regulatory approval, reimbursement or payment for healthcare providers, sustainable tracer production and supply. PET radiotracers which are broadly employed for clinical applications in oncology are here concisely described.

The most common PET radiotracer used clinically in oncology is ^{18}F-fluoro-2-deoxy-glucose (^{18}F-FDG) (Fig. 1.5a). ^{18}F-FDG is a glucose analogue which enters cells via glucose transporters, where it is either phosphorylated and does not undergo further glycolytic reactions or is transported back outside the cells. By measuring the accumulation of ^{18}F-FDG in tissue, PET imaging can therefore be used to evaluate glucose consumption noninvasively. Notably, tumors commonly have increased glucose consumption, overexpressing glucose transporters and glycolytic enzymes. Thus, ^{18}F-FDG PET images are used to identify tumors in a broad variety of cancer types. Nevertheless, not all cancer cells have increased glycolytic activity as to accumulate ^{18}F-FDG: neuroendocrine and prostate tumors for instance do not have significantly upregulated ^{18}F-FDG uptake. Moreover, several biological processes unrelated to cancer are associated with increased glucose consumption, including muscle activation, brown fat activation, inflammation, infection, and bone regeneration. While the numerous processes of glucose absorption unrelated to cancer may be assessed with ^{18}F-FDG PET for alternative clinical and research purposes, in clinical ^{18}F-FDG PET oncology scans such processes may constitute a confounding factor, limiting the specificity in identifying tumor sites and requiring careful image inspection for appropriate interpretation [112].

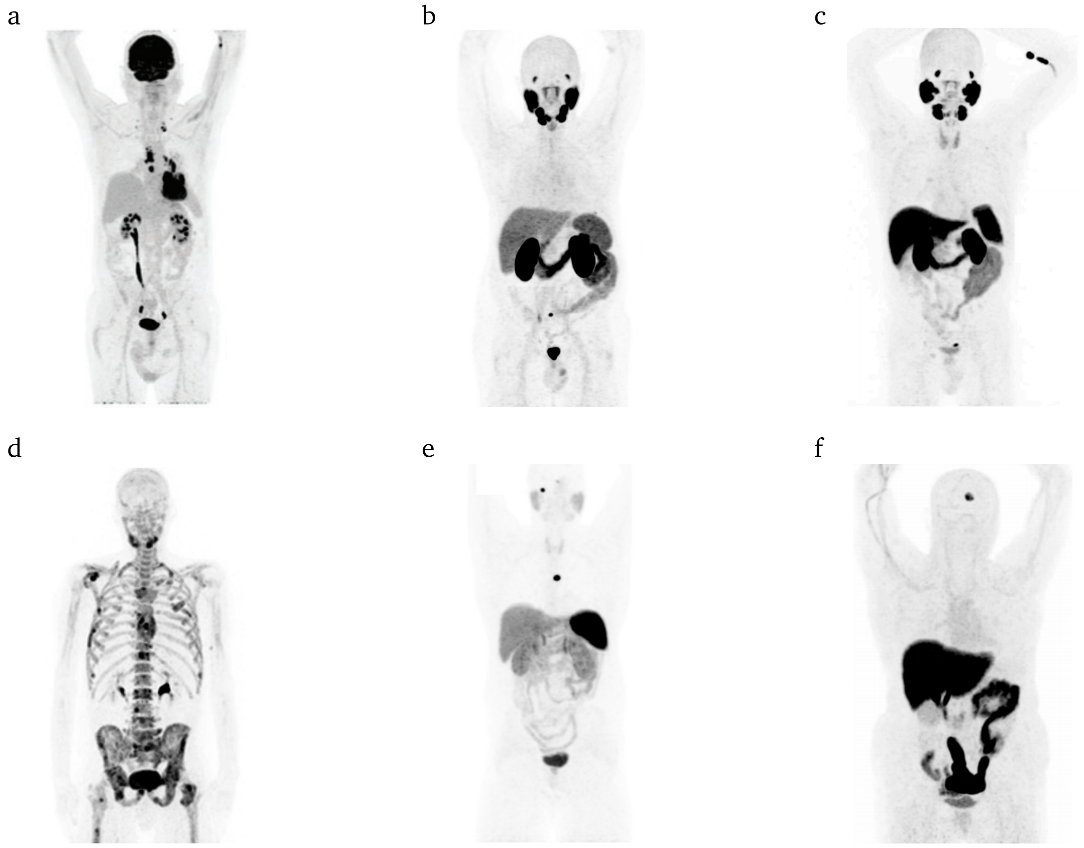


Fig. 1.5. Maximum Intensity Projections of PET scans obtained with different radiotracers used in oncology: (a) ^{18}F-FDG, (b) ^{68}Ga-PSMA-11, (c) ^{18}F-PSMA-1007, (d) ^{18}F-NaF PET, (e) ^{68}Ga-DOTATE, (f) ^{18}F-FES. Adapted from [114][41][45][53][19][43].

Several radiotracers are used for prostate cancer imaging. Carbon-11 choline is used to assess tumor lipid metabolism, as choline is a precursor for the synthesis of phospholipids. Prostate tumors have an altered lipid metabolism which results in increased uptake of ^{11}C-choline [91]. Fluorine-18 fluciclovine is a synthetic leucine amino acid analogue used to assess tumor amino acid metabolism. Prostate tumors have increased expression of amino acid transporters and can be identified in ^{18}F-fluciclovine PET imaging [105]. Radiotracers designed to bind with prostate specific membrane antigen (PSMA) have been increasingly employed for prostate cancer imaging (Fig. 1.5b, Fig. 1.5c). PSMA is a transmembrane protein expressed in prostate tissue and highly overexpressed in a large majority of prostate cancer cells [33]. PSMA-ligand PET can be used to identify prostate cancer sites for primary staging of high-risk patients, after biochemical recurrence, during systemic treatment, and prior to PSMA-targeted radioligand therapy [38]. Several PSMA-targeted radiotracers have been developed, including ^{68}Ga-labeled and ^{18}F-labeled compounds, with multiple ones being evaluated in clinical trials (Tab. 1.1) [134]. Notably, in up to 5% of prostate cancer patients PSMA-ligand PET can produce false negatives [38], as a fraction of prostate cancer cells do not overexpress PSMA. Other than in prostate tissue, physiologic PSMA expression is present in salivary glands, kidneys, intestine, and parasympathetic ganglia. Moreover, PSMA is overexpressed in the neovasculature of non-prostate cancer tumors. PSMA-targeted tracer uptake has been described in several malignant tumor types and nonmalignant conditions including increased osteoblastic activity, hemangiomas, and inflammation [59].

Isotope	Tracer	Excretion
^{68}Ga	^{68}Ga-PSMA-11	Kidney-dominant
^{18}F	^{18}F-DCFPyL	Kidney-dominant
	^{18}F-PSMA-1007	Liver-dominant
	^{18}F-rhPSMA-7.3	Liver-dominant

Tab. 1.1. Exemplar ^{68}Ga-labeled and ^{18}F-labeled radiotracers targeting PSMA with completed or ongoing clinical trials

Fluorine-18 sodium fluoride (^{18}F-NaF) (Fig. 1.5d) is a radiotracer used to assess bone remodeling. Fluoride-18 ions ([^{18}F]F-) bind to the bone matrix where they are in part incorporated forming fluorapatite. [^{18}F]F- uptake is increased in regions with higher perfusion and bone remodeling rates [29]. In oncology applications, ^{18}F-NaF PET is used to identify bone metastases, in which bone remodeling is altered resulting in higher tracer uptake. Several physiological processes and nonmalignant diseases related to bone can result in increased [^{18}F]F- uptake and may constitute a confounding factor for identifying tumor sites [4].

Somatostatin receptor-targeted radiotracers are used for imaging neuroendocrine tumors. Tumors originating from neuroendocrine cells typically overexpress somatostatin receptors, and multiple radiotracers with different affinity to somatostatin receptor subtypes have been developed to identify neuroendocrine tumors via PET imaging. The tracers ^{64}Cu-DOTATATE and ^{68}Ga-DOTATATE (Fig. 1.5e) have the highest affinity to SSTR2, the most frequent receptor subtype. The tracer ^{68}Ga-DOTATOC has affinity to both SSTR2 and SSTR5 receptor subtypes [66]. PET imaging with somatostatin receptor-targeted radiotracers can be used to aid diagnosis, staging and follow up of patients with neuroendocrine tumors as well as to select patients for peptide receptor radionuclide therapy [121], which was shown to be beneficial for well-differentiated metastatic disease [119].

Fluorine-18 fluoroestradiol (^{18}F-FES) (Fig. 1.5f) is an estrogen analogue used to assess the expression of estrogen receptor in tumor cells. Brest cancer cells express the estrogen receptor (ER) in around 75% of the cases at diagnosis [127]. Patients with ER-positive breast cancer are more likely to respond to antihormonal therapy. The ER-status of metastatic lesions can differ from the primary tumor and be heterogeneous within a single tumor or between lesions within a single patient. ER-status evaluation via biopsy is only possible for a limited number of patients and metastatic sites. ^{18}F-FES PET can be used to identify ER-positive breast cancer noninvasively and support the selection of patients who are more likely to benefit from antihormone therapy.

1.4.2 Cancer stage

The anatomical extent of cancer spread is an informative biomarker which can be evaluated in PET/CT imaging and is used to estimate prognosis, plan treatment, and monitor response to therapy. Several standardized classification systems for different cancer types have been proposed to characterize the disease spread, facilitating reproducible clinical assessment and communication of findings. Moreover, cancer spread, in some cases combined with other prognostic markers, determines cancer stage, which is an indicator of cancer progression highly relevant for patient management.

The TNM classification of malignant tumors is a standardized system used to classify the anatomical extent of cancer spread in numerous types of malignancies. The TNM system defines specific classes of cancer spread related to the primary tumor (T), regional lymph nodes (N) and distant metastases (M). The exact anatomical boundaries of each class are defined differently for each cancer type depending on the organ or tissue of origin. The TNM classification system is maintained in collaboration between the Union for International Cancer Control (UICC) and the American Joint Committee on Cancer (AJCC). Both UICC and AJCC publish separate staging manuals [8][1], while they work jointly to regularly update their staging systems based on recent scientific evidence and changes in cancer management.

The TNM classification is an assessment of cancer stage solely based and anatomic spread. Recent AJCC staging systems define prognostic stage groups (from stage I to stage IV), to stratify patients with respect to outcome, on the basis of the TNM classification and additional prognostic indicators, such as tumor grade or serum tumor markers, where these are highly relevant for prognosis. For instance, the AJCC prognostic stage groups for prostate cancer are determined based on TNM classification, tumor grade, and blood level of prostate specific antigen.

Dedicated staging systems have been developed for specific cancer types, including blood cancer subtypes for which the TNM classification is not used. The Ann Arbor staging system [17] and the Lugano staging classification [20] are used to classify the extent of anatomical spread of lymphoma, based on the degree of involvement of lymph nodes and other organs. While the TNM classification is globally recognized, several cancer type-specific classification systems have been devised by dedicated working groups, including broadly established ones such as the International Federation of Gynecology and Obstetrics (FIGO) [89] and the International Association for the Study of Lung Cancer (IASLC) [47]. The TNM classification is not used for brain and spinal cord tumors, in which the primary tumor size is significantly less relevant than its histology and location, while extraneural metastases are rare [1].

PET/CT imaging, combining functional and structural information, allows to identify tumor sites and their anatomical location in the whole body in several cancer types and is an established clinical tool for staging [92].

1.4.3 Semiquantitative PET parameters

In PET imaging, the radioactivity concentration at a given spatial location and time point is proportional to the concentration of radiotracer and its metabolites. The radioactivity concentration measured therefore depends on administered dose, distribution volume, and kinetic behavior of the radiotracer. Thus, PET images expressed in units of radioactivity concentration are not a quantitative measure of the biochemical process of interest. In static PET scans, semiquantitative units have been proposed to partially correct for factors of inter-scan variability and compare PET image values between subjects or time points under specific assumptions.

Standardized Uptake Value (SUV) units are broadly used to express PET images in semiquantitative values. SUV units (SUV$_{BW}$) are calculated as ratio between the measured radioactivity concentration, and the administered dose divided by body weight. The radioactivity concentration and administered dose are decay corrected to the same time point. Instead of body weight, the administered dose may be divided by the lean body mass to calculate SUV units (SUV$_{LBM}$). Alternatively, the administered dose may be divided by body surface area to obtain SUV units (SUV$_{BSA}$). SUV$_{LBM}$ and SUV$_{BSA}$ units have been introduced to account for the nonuniform relative tissue composition in patients with different body weights. For instance, in ^{18}F-FDG PET, tracer uptake in adipose tissue is low, and SUV$_{LBM}$ units, also referred to as SUL units, are recommended compared to SUV$_{BW}$ units to avoid overestimation of tracer uptake in obese patients. Since membrane transporters driving ^{18}F-FDG uptake may be saturated by glucose, SUV units may be corrected in ^{18}F-FDG PET for the plasma glucose level, by multiplication with the measured plasma glucose concentration divided by the population average of $5.0\,\mathrm{nmol/L}$ [7]. Overall, SUV values are influenced by several factors in ^{18}F-FDG PET, including patient habitus, tracer uptake time and plasma glucose levels [70]. Guidelines have been defined to standardize the imaging procedure and facilitate reproducibility [7]. SUV values are typically reported to describe the tracer uptake within a region of interest (ROI) in the PET image, for instance corresponding to an organ or tumor site. Following the definition of an ROI, parameters frequently used to describe the tracer uptake within the ROI are the mean SUV value (SUV$_{mean}$), the maximum SUV value (SUV$_{max}$) and the average SUV value obtained positioning a $1\,\mathrm{mL}$ sphere such that the average SUV value within the sphere and the ROI is maximized (SUV$_{peak}$). SUV$_{mean}$, SUV$_{max}$ and SUV$_{peak}$ parameters are all dependent on the exact ROI delineation and image noise level, each SUV parameter to a different extent, depending on the ROI size and tracer uptake pattern. SUV parameters derived from ^{18}F-FDG PET can be used to evaluate treatment response in cancer patients. In patients with non-Hodgkin lymphoma undergoing chemotherapy, the percent change between baseline and interim PET scan of the highest SUV within manifestations of lymphoma (ΔSUV$_{max}$) is an independent determinant of treatment outcome [31]. The established PET Response Criteria in Solid Tumors (PERCIST) defines objective response classes based on percent changes, between baseline and follow up ^{18}F-FDG PET scans, of the highest SUV$_{peak}$ within lesions, changes in lesions size, and presence of new lesions [130].

Standardized Uptake Ratio (SUR), also referred to as SUV ratio (SUVR), are dimensionless units used to express PET images in semiquantitative values of tracer uptake in relation to the uptake in a reference body region. SUR units are calculated in any PET image region as ratio between the measured radioactivity concentration and the radioactivity concentration measured in a specific region of interest used as reference. Assuming no specific uptake occurs in the reference region, SUR units are used to partially correct for factors inter-scan variability of the measured radioactivity concentration, such body habitus, administered dose, and plasma clearance of the radiotracer. SUR can be referred to as tumor-to-liver ratio (TLR), tumor-to-blood ratio (TBR) or tumor-to-muscle ratio (T/M) depending on region of interest used as reference. In ^{18}F-FDG PET, tumor-to-blood SUR has been shown to significantly correlate with metabolic rate of FDG in tumors [126]. A one to five score, based on tumor-to-mediastinum ratio, tumor-to-liver ratio, and presence of new lesions in a follow up ^{18}F-FDG PET scan, named Deauville score is broadly used to evaluate treatment response in patients with lymphoma [83].

1.4.4 Tumor burden parameters

Semiquantitative parameters to characterize tumor burden from PET images have been proposed as image-derived biomarkers. Compared to PET parameters used to describe tracer uptake for a single region of interest, tumor burden parameters are used to describe the total spatial extent or intensity of malignant tracer uptake in an organ or in the entire body region covered by the PET scan. Numerous research investigations have demonstrated that tumor burden parameters have significant prognostic value. Nevertheless, tumor burden parameters are not yet used in the clinical routine, in part because they require systematic delineation of image regions with tracer uptake suspicious for cancer, which can be time consuming and have limited accuracy when performed with commonly used semi-automated techniques, while no consensus yet exist on the use of a specific delineation method.

Total Metabolic Tumor Volume (TMTV), also referred to as Metabolic Tumor Volume (MTV), is a semiquantitative tumor burden parameter derived from ^{18}F-FDG PET images expressed in SUV units. TMTV is calculated by delineating PET image regions with tracer uptake suspicious for cancer and summing the volume of all delineated regions. Typically, semi-automated delineation methods are used, in which an expert reader gives initial input to or adjusts results of an automated segmentation algorithm, commonly based on a fixed SUV threshold, a fixed-percentage threshold relative to the SUVmax of each suspicious region, region growing, or a combination of them. The use of different methods to determine TMTV has been found to yield appreciably different TMTV values when applied to the same cohort, while the different values obtained retained a comparable prognostic value [24][62]. In retrospective analyses, TMTV has been shown to have significant prognostic value for overall survival in multiple cancer types, including head and neck cancer [87], esophageal carcinoma [61], follicular lymphoma [82], and diffuse large B-cell lymphoma [25]. Analogous to TMTV, PSMA Total Volume (PSMA-TV) has been proposed as tumor burden parameter derived from PSMA-ligand PET images expressed in SUV units. In recent retrospective analyses, PSMA-TV was found to have significant prognostic value for risk stratification and response assessment of prostate cancer patients [51][109][106][110].

Total Lesion Glycolysis (TLG) is a semiquantitative tumor burden parameter derived from ^{18}F-FDG PET images expressed in SUV units. Similar to TMTV, TLG is determined by delineating PET image regions with tracer uptake suspicious for cancer. TLG is calculated by summing for each delineated region the volume of the region multiplied by its SUV$_{mean}$. In retrospective analyses, TLG was found to have significant prognostic value for risk stratification and response assessment in multiple solid tumor types [125]. With a definition analogous to TLG, PSMA Total Lesion uptake (PSMA-TL) has been proposed as tumor burden parameter derived from PSMA-ligand PET images expressed in SUV units [107].

Skeletal tumor burden parameters are image-derived biomarkers used to characterize tumor load in bone as indicator of disease burden. Bone PET Index Volume (BPI$_{vol}$) and Bone PET Index SUV (BPI$_{suv}$) are semiquantitative tumor burden parameters derived from ^{68}Ga-PSMA-11 PET/CT images [5]. BPI$_{vol}$ is calculated as ratio, expressed as percentage, between the volume of bone metastases, determined from the PET image by delineation of regions with uptake suspicious for cancer, and the volume of the skeleton, determined by delineation of bone from the CT image. BPI$_{suv}$ is calculated as ratio between the sum for each bone metastasis of the region volume multiplied by its SUV$_{mean}$, and the volume of the skeleton. bonePSMA-TV and bonePSMA-TL are tumor burden parameters with a definition analogous to PSMA-TV and PSMA-TL, while limited to image regions suspicious for cancer in the skeleton [42]. Fluoride Tumor Volume (FTV) and Total Lesion Fluoride (TLF) are skeletal tumor burden parameters derived from ^{18}F-fluoride PET images, with a definition analogous to bonePSMA-TV and bonePSMA-TL [100].

1.5 Image classification

This section introduces elements of image classification. The purpose is to provide essential background information related to the contributions presented in the dissertation, while references to more detailed descriptions of the discussed methods are provided. A brief introduction to the image classification problem is provided. Deep learning methods used for image classification, including convolutional neural networks and transfer learning, are concisely described. Metrics to evaluate the performance of a classification algorithm are outlined.

1.5.1 Introduction

In computer vision, image classification is the task of assigning to an image a single label or multiple labels from a fixed set of predefined classes. The term binary classification is used to indicate the task of assigning a single label between two predefined classes. The term multiclass classification is used to indicate the task of assigning a single label from three or more predefined classes. The term multi-label classification is used to indicate the task of assigning any set of labels from a fixed set of predefined classes.

Since the introduction of digital images, methods for automated image classification gained interest in research and found application in numerous fields such as factory automation, office automation, surveillance, medical imaging, and biometrics. In any real-word setting, numerous variability factors pose challenges to the image classification task, such as variations in viewpoint, contrast, scale, deformation, occlusion and background. Moreover, high intra-class variation can result in significantly dissimilar appearances for instances of the same class. To address these challenges, several image classification methods employ supervised machine learning algorithms, whereby the visual appearance of each class is modeled trough a set of examples with a known label. Methods which rely on a predefined set of image features are concisely described in this subsection, while convolutional neural networks, employing image features learned directly from the training examples, are described in section 1.5.2.

Methods employing predefined image features have been described for the task of object recognition, which requires both identifying the presence of a given object in an image and determining its position. The task of object recognition is closely related to image classification, since it can be considered as the task of classifying for each image subregion whether a given object is tightly enclosed by the image subregion. Local image features invariant to scale, translation, and rotation, determined at defined key image locations through a Scale Invariant Feature Transform (SIFT) were described and proven useful for object recognition given an example model image of the same object [78]. Haar-like image features employed by a cascade of classifiers, each trained with a boosting algorithm, were shown to be effective for rapid face detection [129]. Features based on Histograms of Oriented Gradients (HOG), determined for all image regions were employed by a support vector machine classifier, used to scan all image positions and scales to identify the presence of a person, for the task of pedestrian detection [30]. Deformable Parts Models (DPM) were used for detection of objects of multiple categories and consisted in part-based models with a star structure, composed of a root filter associated to the entire object, plus a set of parts filters and corresponding deformation models. Filters were based on HOG features and trained using latent support vector machines classifiers, where a latent variable was used to account for different object configurations while having partially labeled training data [37].

1.5.2 Convolutional neural networks

Convolutional neural networks (CNNs) are a subclass of artificial neural networks used as machine learning algorithm in several computer vision applications including image classification, object localization and instance segmentation. Compared to image processing methods relying on predefined image features, CNNs employ a set of convolution operations whose parameters are optimized directly to minimize the error for the task of interest on the training examples. The distinctive ability to identify useful image features through the algorithm training has been termed feature learning.

A convolutional neural network for image classification was first described for the task of handwritten digits recognition [76]. The CNN was composed of four sequential layers; each of the first two layers included convolution, subsampling and pointwise hyperbolic tangent operations, while each of the latter two layers included linear mapping and pointwise hyperbolic tangent operations. The network parameters were optimized to minimize a mean squared error cost function based on the expected output and CNN output on a set of labeled training examples, using stochastic gradient descent and backpropagation. The network proved to be effective in classifying handwritten digits with a broad variety of styles, size, and graphical completeness. A method for face detection in grayscale images was described employing a CNN to classify whether an image window contains a face. The network was trained with a bootstrap algorithm, which included falsely detected examples in the training set as the training progressed. Moreover, predictions from multiple identical networks trained with different random weight initializations were combined, and the method was able to detect faces in a wide variety of images, while producing a limited number of false detections [102].

The availability of large, annotated visual datasets, and the efficient implementation of the convolution operation in powerful graphical processing units supported the introduction of a convolutional neural network used to classify real-word images in one thousand object categories [73]. The network comprised five convolutional layers and three fully connected layers. The CNN employed rectified linear units as activation function enabling faster training compared to the saturating hyperbolic tangent activation function. Moreover, a regularization technique named dropout was used, by which a fraction of the outputs of hidden neurons in the fully connected layers was randomly set to zero at each training iteration. The CNN significantly outperformed other classification methods based on predefined image features. CNN architectures were described which further improved the accuracy of image classification, including the use of a small filter size with more convolutional layers [116], inception modules combining filters of different sizes [120], and residual connections performing identity mapping between layers blocks to facilitate optimization of networks with numerous convolutional layers [57]. In the narrow well-defined task of classifying real-word images in a predefined benchmark dataset with a fixed set of object categories, convolutional neural networks approached the performance previously reported for a single human annotator [58].

1.5.3 Transfer learning

Machine learning methods traditionally rely on the assumption that training and test data are drawn from the same distribution and feature space. However, in many applications collecting a sufficiently large training dataset is expensive and impractical or infeasible. In this context, knowledge transfer from a related application for which training data is available can be beneficial. Formally, a domain can be defined as a feature space paired with a marginal probability distribution, while a task can be defined as a label space paired with a predictive function, which can be considered as the conditional probability distribution of labels given a feature vector. Given a source domain and task, transfer learning aims at improving the learning in a target application with a different domain or task, using the knowledge from the source information [88].

The term inductive transfer learning is used for cases where the target task is different from the source task, and labeled data is available in the target domain. Moreover, in the particular case in which no labeled data is available in the source domain, the term self-taught learning has been proposed to describe a framework in which a feature representation is derived from unlabeled data in the source domain and then used to train a predictive model in the target domain [95]. The term unsupervised transfer learning is used for cases where the source and target task are different, and no labeled data is available in the source and target domains. The term transductive transfer learning is used for cases where source and target tasks are the same, but source and target domains are different. Furthermore, the particular case in which source and target feature spaces are the same but probability distributions are different is referred to as domain adaptation.

For the task of image classification, different transfer learning approaches based on convolutional neural networks have been proposed, relying on the transfer of feature representations and network parameters. For instance, CNNs trained on large image datasets for object category classification can be used as fixed feature extractors to obtain meaningful image representations and employed by machine learning methods in a different target domain or task. Depending on level of similarity between the source and target domains and tasks, more generic low-level CNN features may be used from activation maps of early layers, or more dataset-specific high-level CNN features may be used from activation maps of later layers. It was shown that training a linear classifier using high-level features of a CNN developed for object classification improved results compared to highly tuned task-specific techniques in several visual recognition applications [98]. An alternative approach to using the source CNN layers for fixed feature extraction consists in fine-tuning the source CNN layers by further optimizing their parameters with training examples for the target application.

1.5.4 Evaluation metrics

Multiple parameters can be used to evaluate the performance of a classification algorithm. For machine learning methods, performance parameters are determined based on the ability of the method to correctly classify items in a dataset used specifically for evaluation purposes, here referred to as the test set. Ideally, the composition of a test set should appropriately reflect the distribution of cases in which the method is intended to be applied. In practice, test datasets with a limited number of items are used, and a selection bias may be present due to limitations in data collection. The relevance of the different evaluation metrics depends on the context in which a classification method is applied and the composition of the test dataset used for the evaluation. In this section, parameters used to evaluate the performance of classification algorithms discussed in the dissertation are concisely described.

In the task of binary classification, the two labels which can be assigned are often defined as the presence or absence of a given attribute of interest. In this context, the instances used to evaluate the classifier performance are conventionally named with either the term ‘positive’ or ‘negative’, based on whether or not they were classified as having the attribute of interest by the algorithm, preceded by the term ‘true’ or ‘false’, based on whether or not they were correctly classified compared to the ground truth label. A two-by-two confusion matrix can be used to represent the classification of instances in the test set, in which the row index indicates the ground truth label, and the column index indicates the label assigned by the classification algorithm (or vice versa), such that entries indicate the total number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Tab. 1.3 reports parameters that can be derived to describe binary classification performance.

Moreover, in cases where the prediction of a binary classifier is determined by thresholding of a positivity score employed by the classification algorithm, different parametric curves can be drawn as a function of the prediction threshold (θ), whereby a positivity score higher than the threshold is associated with a positive prediction. For each parametric curve, a corresponding metric can be determined to assess the classifier performance (Tab. 1.3).

Parameter	Description	Computation
Accuracy	fraction of test set instances correctly classified	(TP+TN)/(TP+FP+TN+FN)
Sensitivity, recall or true positive rate (TPR)	fraction of test set instances correctly classified among the ones having the attribute of interest according to the ground truth label	TP/(TP+FN)
False negative rate (FNR)	complementary to sensitivity	FN/(TP+FN)=1-sensitivity
Specificity or true negative rate (TNR)	fraction of test set instances correctly classified among the ones which do not have the attribute of interest according to the ground truth label	TN/(TN+FP)
False positive rate (FPR)	complementary to specificity	FP/(TN+FP)=1-specificity
Positive predictive value (PPV) or precision	fraction of test set instances correctly classified among the ones which were predicted as having the attribute of interest	TP/(TP+FP)
Negative predictive value (NPV)	fraction of test set instances correctly classified among the ones which were predicted as not having the attribute of interest	TN/(TN+FN)
Youden's J statistic (J)	combines sensitivity and specificity	sensitivity+specificity–1

Tab. 1.2. Parameters used to describe the performance of a binary classification algorithm

In the task of multiclass classification, a n-by-n confusion matrix can be used to represent the classification of instances in the test set, where n is number of possible labels, the row index indicates the ground truth label, and the column index indicates the label assigned by the classification algorithm (or vice versa). Accuracy is defined as the fraction of test set instances correctly classified by the algorithm. Binary classification metrics can be determined relative to a single class by considering only the membership to a specific class as attribute of interest for the prediction. Moreover, binary classification metrics determined separately for each class can be aggregated by average or by sum weighted for the prevalence of each class in the test set, to obtain multiclass performance metrics. For instance, balanced accuracy (BA) can be determined as the mean of true positive rate scores relative to each class, while the mean average precision (mAP) can be determined as the mean of average precision scores relative to each class.

Parameter	Description	Parametric curve	Computation
Area under the Receiver Operating Characteristic Curve (AUCROC)	area under the parametric curve representing false positive rate and sensitivity as a function of the prediction threshold	$(x, y)_{ROC} = (FPR(\theta), TPR(\theta))$	$\int_0^1 y\, dx = \int_{+\infty}^{-\infty} TPR(\theta) FPR'(\theta)\, d\theta$
Area under the Free-response Receiver Operating Characteristic curve (AUCFROC)	area under the parametric curve representing the number of false positives per relevant parent test unit, for instance a single image or subject comprising multiple test instances, and sensitivity as a function of the prediction threshold, limited upwards to one false positive per test unit	$(x, y)_{FROC} = (FPR_{per-unit}(\theta), TPR(\theta))$	$\int_0^1 y\, dx = \int_{+\infty}^{FPR^{-1}_{per-unit}(1)} TPR(\theta) \cdot FPR'_{per-unit}(\theta)\, d\theta$
Area under the precision-recall curve or average precision (AP)	area under the parametric curve representing sensitivity and precision as a function of the prediction threshold	$(x, y)_{PR} = (TPR(\theta), PPV(\theta))$	$\int_0^1 y\, dx = \int_{+\infty}^{-\infty} PPV(\theta) TPR'(\theta)\, d\theta$

Tab. 1.3. Parameters used to describe the performance of a binary classification algorithm for which the prediction can be obtained as function of a threshold (θ) applied to a positivity score employed by the algorithm, such that a positivity score higher than the threshold is associated with a positive prediction.

Part II

Contributions

Radiotracer uptake classification 2

This chapter presents contributions in image analysis methods for automated classification of foci with elevated radiotracer uptake in PET/CT images. The methods described are aimed at supporting time-efficient, reproducible, and accurate measurements for the assessment of image-derived cancer biomarkers. While the motivation and methods described are applicable to different cancer types and PET radiotracers, the analysis presented focuses primarily on prostate cancer and PSMA-ligand PET/CT. The use of ^{18}F-FDG PET/CT images of lymphoma and lung cancer patients as additional source of training information is investigated given the context of limited availability of PSMA-ligand PET/CT images with expert-annotated ground truth. This chapter concisely introduces background information on prostate cancer and PSMA-ligand image analysis. Methods for radiotracer uptake classification using a convolutional neural network are described, corresponding results are reported and discussed. The evaluation of the presented image analysis method for prostate cancer staging support is discussed in Chapter 3.

Substantial parts of this chapter have already been published and are quoted verbatim:

[12] N. Capobianco et al. “Transfer Learning of AI-Based Uptake Classification from ^{18}F-FDG PET/CT to ^{68}Ga-PSMA-11 PET/CT for Whole-Body Tumor Burden Assessment”. In: *Journal of Nuclear Medicine* 61.supplement 1 (May 2020), p. 1411

[16] N. Capobianco et al. “Whole-Body Uptake Classification and Prostate Cancer Staging in 68Ga-PSMA-11 PET/CT Using Dual-Tracer Learning”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* (July 2021)

2.1 Introduction

Accurate staging has a pivotal role in the management of prostate cancer, a disease with generally favorable outcome when confined to the prostate, while having poor prognosis if metastasized at the time of diagnosis [115]. As a plethora of management strategies is available, ranging from watchful waiting to localized and systemic treatments, reliable information on the disease spread pattern and overall burden is crucial in the clinical decision-making process [85]. While the gold standard for prostate cancer staging remains histopathology, imaging is increasingly being utilized as noninvasive assessment [79]. Notably, Prostate-Specific Membrane Antigen (PSMA)-targeted PET/CT has shown high accuracy, superior to other imaging modalities, for primary staging of high-risk prostate cancer patients [60][81] as well as for staging after biochemical recurrence [96][32]. The ^{68}Ga-PSMA-11 compound manufactured by the University of California, San Francisco and the University of California, Los Angeles has recently received approval by the U.S. Food and Drug Administration.

In addition to procedure guidelines [38], pitfalls reviews [111][59][97] and case reports [35][44], standardized reporting frameworks for PSMA-ligand PET have been proposed to support replicable and rigorous image assessment [34][101][18]. Moreover, the use of quantitative image-derived biomarkers, such as the total tumor volume, has shown promising results for risk stratification and response assessment [51][109][106][110]. Nevertheless, the application in clinical routine of detailed reporting schemes and image-derived biomarkers remains labor intensive, subject to error and operator dependent in cases where a high number of manual measurements is required, such as when categorical or quantitative variables have to be determined for all suspected lesions. In this context, the use of semi-automated and automated image analysis methods is promising to support accurate, reproducible, and time-efficient assessment.

Recently, semi-automated and automated methods for image analysis in ^{68}Ga-PSMA-11 PET/CT have been developed. A convolutional neural network (CNN) was trained to predict the PSMA-ligand PET positivity status of lymph nodes from CT alone [54], showing a performance comparable to trained radiologists. To support semi-automated quantification of tumor burden, masks of organs that exhibit physiological uptake and bone were obtained from CT images using thresholding methods [5][52], machine learning methods [42], and deep learning methods [109]. While the CT information alone can be used to aid semi-automated identification and anatomical localization of suspicious elevated uptake sites, including the PET information in a machine learning system for whole-body PSMA-ligand image analysis could be beneficial. In particular, the identification of elevated uptake regions as physiological based on automated analysis of the sole CT information is particularly challenging for regions such as small intestines or ureters and would require manual corrections. A machine learning algorithm trained on multimodal PET/CT information may more accurately identify such regions of physiological uptake limiting the number of manual corrections required, as well as potentially being able to identify further challenging patterns of nonsuspicious uptake, such as uptake in ganglia and unspecific uptake in lymph nodes and bone. Recently, a convolutional neural network was trained with multimodal PET/CT information to identify tracer uptake regions suspicious for prostate cancer within the pelvis [132] with promising results.

In the current analysis, we developed and evaluated a multi-task convolutional neural network trained on the PET and CT information for the identification and anatomical location classification of suspicious tracer uptake sites in the entire axial body coverage of the scan. We employ multi-task training, previously evaluated in ^{18}F-FDG PET/CT [114] with encouraging results, for assessment of ^{68}Ga-PSMA-11 images. In addition, we explore two strategies to leverage training information from both radiotracers: transfer learning by pretraining on ^{18}F-FDG images with fine-tuning on ^{68}Ga-PSMA-11 images, and a modified network architecture for synergistic dual-tracer learning.

2.2 Materials and Methods

2.2.1 Patients

Two groups of subjects who underwent ^{68}Ga-PSMA-11 PET/CT at the Klinikum rechts der Isar (Technical University of Munich) were retrospectively analyzed. The rationale for the definition and inclusion of the two groups was to allow the representation of different disease stages in the training dataset while keeping an acceptable expected annotation workload for the expert readers, employing a different annotation scheme for each group. The first group, referred to as group A, consisted of 123 consecutive subjects referred to PSMA-ligand PET/CT for primary staging or for assessment of biochemical recurrence. The second group, referred to as group B, consisted of 50 consecutive subjects referred to PSMA-ligand PET/CT for all indications of prostate cancer. PET/CT images were acquired on a Biograph mCT scanner (Siemens Medical Solutions). Diagnostic CT scans were acquired after intravenous injection of contrast agent (Imeron 300), followed by PET acquisition. PET scans were acquired $(54 \pm 10)\,\mathrm{min}$ (mean ± std) after injection of ^{68}Ga-PSMA-11 ligand solution $(149 \pm 26)\,\mathrm{MBq}$, with acquisition time of 3–$4\,\mathrm{min}$ per bed position.

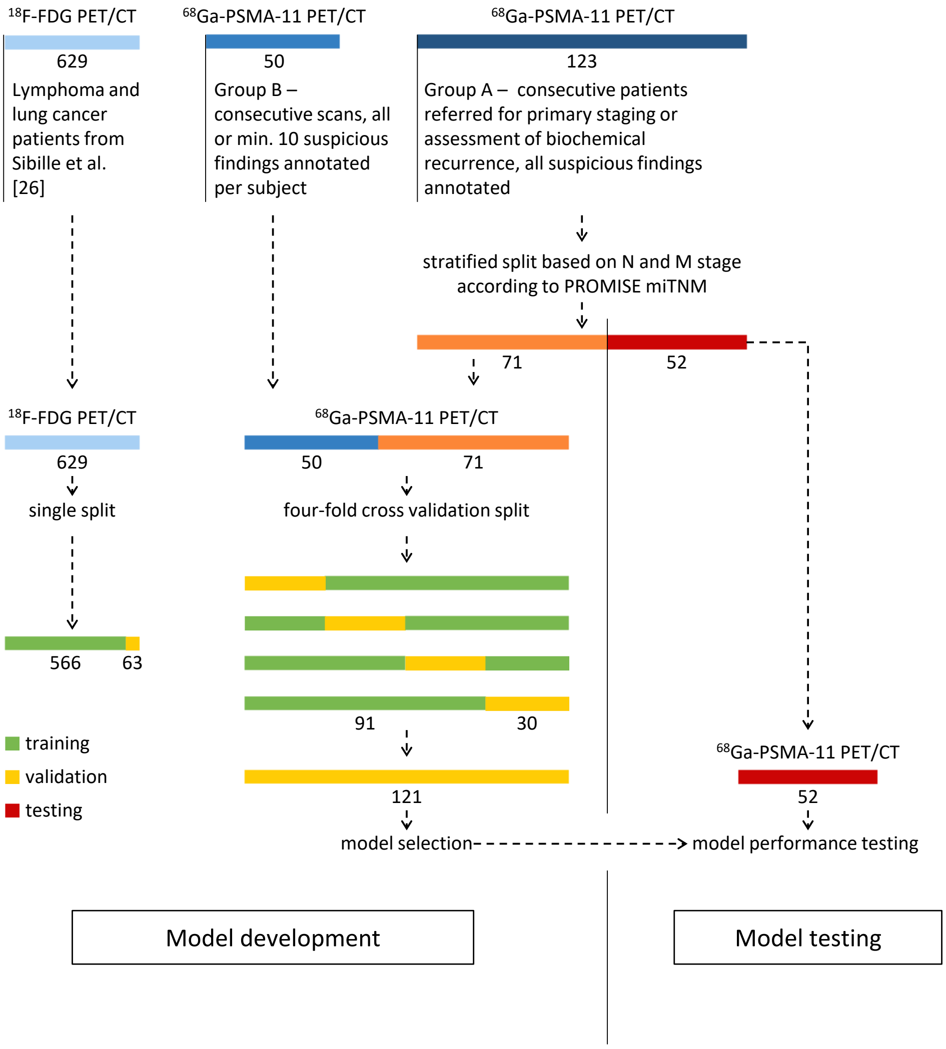
2.2.2 Image analysis

Data annotation

PET/CT images were reviewed by expert nuclear medicine physicians who segmented sites of elevated tracer uptake, labeled them as nonsuspicious or suspicious for prostate cancer, and assigned them an anatomical localization from a set of physiological uptake sites and sites relevant for staging. Due to differences in patient tumor burden and to maintain an acceptable annotation workload, different annotation schemes were used for subjects in group A and group B, which were then considered in the deep learning model development and validation. For subjects in group A, having a low tumor burden, all regions of elevated tracer uptake were segmented semi-automatically using 45% of region SUV$_{max}$ thresholding [106]. For subjects in group B, which included cases of high tumor burden, all high-uptake sites with SUV$_{max}$ above the average liver uptake within a PERCIST-based reference region [130] were segmented with an incremental connected component algorithm [11] using 45% of SUV$_{max}$ thresholding, of which up to one hundred sites per subject with the highest SUV$_{max}$ were annotated. For each subject in group B at least ten suspicious uptake sites were annotated when present, additionally labeling sites with lower SUV$_{max}$ if necessary.

Model development

Subjects of group A (n=123) were assigned to an N and M stage based on expert annotations and following the PROMISE miTNM framework. A stratified split of subjects in group A based on stage was then performed forming a development (n=71) and a hold-out test set (n=52). All subjects of group B (n=50) were added to the development set. Four-fold cross validation on the final development set (n=121) was used to evaluate different model training schemes. The hold-out test set was used exclusively to report results of the model testing and was not employed for the model development. A diagram summarizing the dataset split is reported in Fig. 2.1.



^{18}F-FDG PET/CT
629
Lymphoma and lung cancer patients from Sibille et al. [26]
^{68}Ga-PSMA-11 PET/CT
50
Group B – consecutive scans, all or min. 10 suspicious findings annotated per subject
^{68}Ga-PSMA-11 PET/CT
123
Group A – consecutive patients referred for primary staging or assessment of biochemical recurrence, all suspicious findings annotated
stratified split based on N and M stage according to PROMISE miTNM
71
52
^{18}F-FDG PET/CT
629
single split
566
63
^{68}Ga-PSMA-11 PET/CT
50
71
four-fold cross validation split
91
30
121
model selection
^{68}Ga-PSMA-11 PET/CT
52
model performance testing
🟩 training
🟨 validation
🟥 testing
Model development
Model testing

Fig. 2.1. Diagram summarizing the PET/CT datasets used in the analysis and the data split performed for the deep learning model development and testing.

A multi-task convolutional neural network was trained to both classify PET/CT regions of interest as uptake suspicious or nonsuspicious for cancer and assign them an anatomical location classification. In addition to expert-annotated findings, regions of interest with SUV$_{max}$ above 1 which were not labeled by the experts as suspicious were generated automatically with an incremental connected component algorithm [11], labelled as nonsuspicious, and used for training. These were generated using with 45% of SUV$_{max}$ thresholding and only for subjects in group A or subjects in group B with up to nine suspicious findings, i.e. for PET/CT images where all suspicious findings were annotated and remaining image regions could be considered as physiological uptake. The network architecture and hyperparameters are illustrated in Fig. 2.2a. Inputs to the network are thirteen PET/CT coronal (192 mm x 192 mm) reformations extracted with offsets (-144, -96, -48, -24, -12, -6, 0, +6, +12, +24, +48, +96, +144 mm) from the region of interest SUV$_{max}$ position, after resampling of PET at CT at 3mm isotropic resolution, PET windowing between 0 and 15 SUV and CT windowing between -300 and 300 HU.

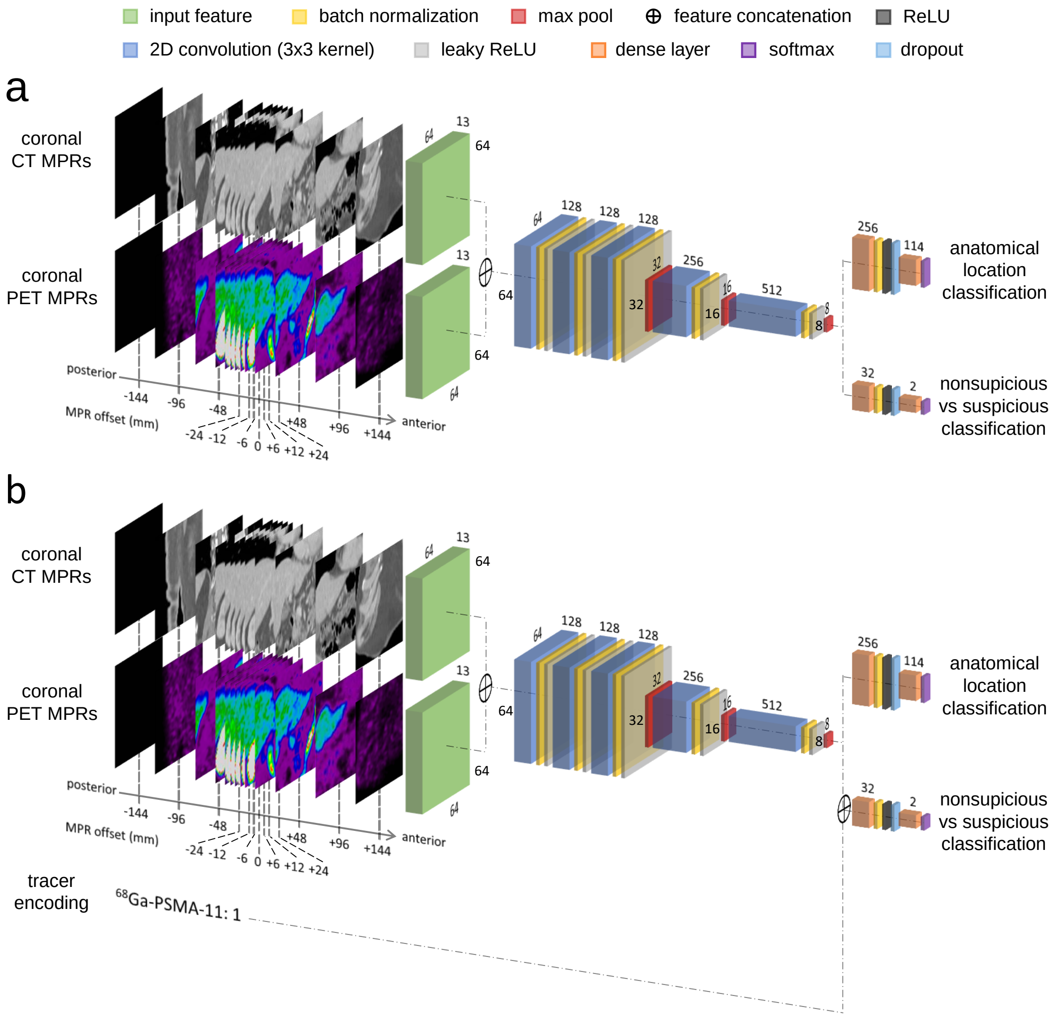


Fig. 2.2. Convolutional neural network architecture used for PET uptake classification when training (a) with data from a single radiotracer and (b) with data from two radiotracers, encoding the tracer type as input to the network. Multiplanar Reformations (MPRs) extracted from a region of interest being classified are represented as exemplar input to the network. In the three-dimensional illustration, numbers along layers' edges indicate the size of the feature maps resulting as output of the corresponding layers.

We evaluated different training strategies to improve the algorithm performance. First (I), the model was trained with sequential sampling of all the training examples. Second (II), a balanced sampling of the training examples was performed, where a fixed maximum number of training examples per class per subject was randomly sampled at each training epoch (maximum of 32 physiological and 32 suspicious findings, 4 findings for each anatomical location class). Third (III), regions of interest used for training were augmented through affine transformations of the PET/CT randomly generated at each training epoch with isotropic scaling between 0.8 and 1.2 and rotations between -17.2 and 17.2 degrees in all directions, to obtain additional regions with plausible pose and size. Forth (IV), to leverage expert knowledge of the same task in ^{18}F-FDG PET/CT images, we trained the network as in (III) with datasets from [114], with a single split between training (90%) and validation (10%). The rationale for the ^{18}F-FDG dataset split was to maximize the training data for knowledge transfer to PSMA-ligand PET/CT, while evaluation on a hold-out ^{18}F-FDG test set was considered outside the scope of the analysis, which is mainly focused on assessing the proposed method for staging support in PSMA-ligand PET/CT. Fifth (V), we evaluated transfer learning by fine tuning on PSMA-ligand PET/CT data the network weights initially trained with ^{18}F-FDG PET/CT images. Sixth (VI), we evaluated simultaneous training with PSMA-ligand and FDG PET/CT images by adding a binary input encoding the tracer type to the first fully connected layer and only for the output branch of the network responsible for classifying nonsuspicious vs suspicious uptake, as illustrated in Fig. 2.2b. Finally, we validated the network with highest performance by training it on the entire development set and evaluating it on the test set.

2.2.3 Statistical analysis

The main metrics used to evaluate the network performance were the area under the precision-recall curve, which accounts for marked class imbalance, referred to as average precision (AP), for the classification of regions as suspicious or nonsuspicious, and the classification accuracy of regions labeled as suspicious by the experts, for the anatomical location classification. The performance metrics were evaluated by pooling findings of all subjects together, and a 95% confidence interval was calculated by 2000 bootstrap resampling of the subjects. To compare different training schemes on the development set, a two-sided paired z-test was performed based on the bootstrap replicates with a significance level set to 5%. Bonferroni correction was used to account for multiple comparisons. For the test set, additional performance metrics were evaluated: number of true positives, false positives, false negatives, recall and positive predictive value for the classification as suspicious or nonsuspicious, classification accuracy of all findings labeled by the experts for the anatomical location classification. For the test set, except for the average precision, the performance metrics were also evaluated and reported on a per-subject basis.

2.3 Results

In total 173 subjects were included in the analysis of which 123 in group A and 50 in group B. A total of 5,577 high uptake regions were annotated, of which 4,520 were physiological uptake and 1,057 were suspicious uptake. The median volume of regions annotated as suspicious was $1.3\,\mathrm{mL}$ (interquartile rage 0.6–$3.0\,\mathrm{mL}$. In addition to the expert-annotated findings, more than 160,000 regions with nonsuspicious uptake were automatically generated for subjects in the development set. A summary of the findings and expert annotations is reported in Tab. 2.1.

Tab. 2.1. Summary of the findings annotated by an expert reader in ^{68}Ga-PSMA-11 PET/CT images using semi-automated segmentation methods, reported by anatomical location label assigned, in descending order of occurrence.

	Number of findings	Number of suspicious findings (%)
Total	5577	1057 (19)
Anatomical location		

Tab. 2.1. (continued).

abdomen, liver	1875	22 (1)
abdomen, small intestine	870	8 (1)
abdomen, kidney	357	0 (0)
neck, submandibular gland	334	0 (0)
neck, parotid gland	310	0 (0)
abdomen, spleen	277	0 (0)
abdomen, lymph nodes, iliacal	164	159 (97)
neck, sublingual gland	149	0 (0)
abdomen, bladder	146	0 (0)
abdomen, lymph nodes, para-aortic	134	133 (99)
abdomen, bones, pelvis	123	121 (98)
thorax, bones, spine	109	109 (100)
abdomen, bones, spine	86	77 (90)
thorax, lymph nodes	79	75 (95)
thorax, bones, ribs	72	72 (100)
neck, glottis	50	0 (0)
abdomen, bones, sacrum	50	47 (94)
abdomen, ureter	38	1 (3)
abdomen, prostate	30	30 (100)
abdomen, lymph nodes, presacral	29	28 (97)
neck, tonsils	27	0 (0)
lower limb, bones, femur	23	22 (96)
thorax, bones, scapula	23	23 (100)
neck, bones, spine	20	19 (95)
cranium, nose	17	0 (0)
thorax, bones, sternum	17	17 (100)
abdomen, lymph nodes, obturator	15	15 (100)
neck, thyroid	15	0 (0)
neck, cervical lymph nodes	13	11 (85)

Tab. 2.1. (continued).

upper limb, bones, humerus	12	12 (100)
abdomen, lymph nodes, inguinal	10	9 (90)
thorax, lung	10	9 (90)
cranium, mouth, teeth	9	0 (0)
abdomen, colon	8	0 (0)
abdomen, lymph nodes, inguinal / femoral	8	8 (100)
thorax, oesophagus	7	0 (0)
thorax, bones, clavicle	6	6 (100)
thorax, skin	5	5 (100)
abdomen, ganglia	5	1 (20)
abdomen, rectum	4	2 (50)
cranium, skull	4	3 (75)
thorax, ganglia	4	1 (25)
cranium, mouth, palate	4	0 (0)
lower limb, lymph nodes, femoral	3	3 (100)
cranium, mouth, floor of mouth	3	0 (0)
abdomen, skin	3	2 (67)
cranium, eye	3	0 (0)
abdomen, lymph nodes, mesenterial	3	2 (67)
abdomen, lymph nodes, peri-hepatic	2	1 (50)
thorax, mediastinum, hilum	2	0 (0)
abdomen, penis	2	0 (0)
abdomen, testis	2	0 (0)
abdomen, adrenal gland	1	1 (100)
neck, accessory sinuses, maxillary sinus	1	0 (0)
thorax, pleura	1	1 (100)
abdomen, stomach, cardia / fundus / body	1	0 (0)
abdomen, lymph nodes, peri-splenic	1	1 (100)
neck, bones, clavicle	1	1 (100)

Fig. 2.3 illustrates results obtained using different methods to train the CNN, evaluated with cross validation on the development dataset of ^{68}Ga-PSMA-11 PET/CT images. The corresponding main performance metrics are summarized in Tab. 2.2. For the classification of findings as suspicious or nonsuspicious, using sequential sampling (I) as baseline [AP: 84.1, confidence interval (CI): 76.2-89.3], a performance improvement not statistically significant after applying Bonferroni correction was found with other training schemes including: balanced sampling (II) (AP: 85.0, CI: 77.5-89.8, p=0.197), its combination with affine (III) data augmentation (AP: 87.0, CI: 81.0-91.3, p=0.067) and their combination with transfer learning (V) (AP: 87.7, CI: 82.3-91.8, p=0.072) or combined training with ^{18}F-FDG data (VI) (AP: 87.9 CI:82.3-91.7, p=0.047). Balanced sampling allowed markedly lower training time due to fewer training examples being processed on average per epoch (3584 vs 128640). For the anatomical location classification of suspicious findings, compared to sequential (I) sampling (accuracy: 64.9, CI: 59.8-70.9), affine data augmentation (III) significantly improved performance (accuracy: 72.7 CI: 68.5-77.1, p<0.001) while balanced sampling (II) alone did not (accuracy: 66.8, CI: 61.5-73.4, p=0.095). Compared to affine data augmentation (III), transfer learning (V) showed a further significant improvement (accuracy: 79.2 CI:75.1-82.7, p=0.001), with a performance not significantly different compared to combined training with ^{18}F-FDG data (VI) (accuracy: 80.0 CI:74.8-84.1, p=0.489), which overall scored highest for both classification tasks.

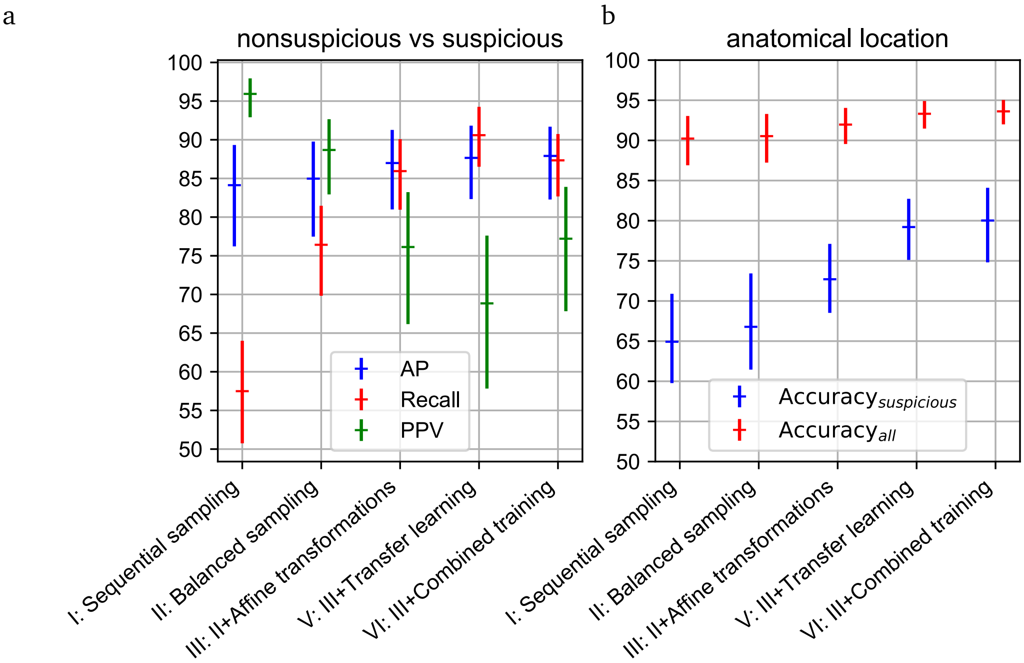


Fig. 2.3. Performance obtained for (a) classification of PET uptake sites as nonsuspicious or suspicious and (b) classification of their anatomical location, using different strategies to train a convolutional network evaluated with four-fold cross validation on the development set of ^{68}Ga-PSMA-11 PET/CT scans. Performance metrics are determined by pooling findings of all subjects together. Error bars indicate the 95% confidence interval obtained via bootstrap resampling at subject level.

Following combined training using ^{18}F-FDG images together with ^{68}Ga-PSMA-11 scans of the entire development set and evaluation on the ^{68}Ga-PSMA-11 test set, an average precision of 80.4 (CI: 71.1-87.8), a sensitivity of 81.1% (CI: 70.6-90.1) and a positive predictive value of 66.8% (CI: 60.3-72.7) were obtained (Tab. 2.3). Anatomical location classification

Tracer	^{68}Ga-PSMA-11		^{18}F-FDG	
Classification output	nonsuspicious vs suspicious	anatomical location	nonsuspicious vs suspicious	anatomical location
Performance metric	AP^{a}	Accuracy$_{suspicious}$	AP^{a}	Accuracy$_{suspicious}$
Model training				
I: Sequential sampling	84.1 (76.2,89.3)	64.9 (59.8,70.9)	-	-
II: Balanced sampling	85.0 (77.5,89.8) p=0.197 vs. I	66.8 (61.5,73.4) p=0.095 vs I	-	-
III: II + Affine transformations	87.0 (81.0,91.3) p=0.067 vs. I	72.7 (68.5,77.1) p<0.001* vs. I	-	-
IV: III on ^{18}F-FDG data	-	-	77.1 (70.8,87.3)	76.7 (71.2,81.3)
V: Transfer learning by fine tuning of IV	87.7 (82.3,91.8) p=0.072 vs. I	79.2 (75.1,82.7) p=0.001* vs. III	-	-
VI: III + combined training on ^{68}Ga-PSMA-11 and ^{18}F-FDG data	87.9 (82.3,91.7) p=0.047 vs. I	80.0 (74.8,84.1) p=0.489 vs. V	77.9 (71.0,86.5) p=0.739 vs. III	77.2 (71.1,81.2) p=0.681 vs. III

^{a}Average Precision
*Significant

Tab. 2.2. PET uptake classification performance using different strategies to train a convolutional network, evaluated with four-fold cross validation on a development set of ^{68}Ga-PSMA-11 PET/CT scans and a fixed validation dataset of ^{18}F-FDG PET/CT scans. Performance metrics are determined by pooling findings of all subjects together, with a 95% confidence interval obtained via bootstrap resampling at subject level, reported in brackets. The P value for a two-sided paired z-test based on bootstrap replicates is reported.

accuracy was 77.0% (CI: 70.0-83.4) for suspicious regions and 94.4% (92.4-96.1) for all expert-annotated regions.

2.4 Discussion

In this analysis we showed that a convolutional neural network can be trained to classify sites of elevated ^{68}Ga-PSMA-11 uptake in the entire axial body coverage of the scan by leveraging both PET and CT information. Having extensively included in the training data regions with uptake above 1 SUV, the network can be used to assess a broad window of the tracer distribution in the body and effectively identify sites suspicious for prostate cancer. Moreover, thanks to the combined identification of suspicious uptake sites and the classification of their anatomical location, the network can be used to assess the spread pattern of suspicious sites in different organs and tissues. Additionally, we found that including training information from PET/CT images and expert annotations obtained with a different PET tracer improved the network performance on ^{68}Ga-PSMA-11 PET/CT images, for which a limited number of reader-annotated cases was available. Previously described methods for ^{68}Ga-PSMA-11 PET/CT image analysis using machine learning were trained on PET/CT information to identify

Tracer	^{68}Ga-PSMA-11						
Classification output	nonsuspicious vs suspicious						
Summary statistic	Pooled (CI)	Per-subject					
		Average	Min	Q1	Median	Q3	Max
Performance metric							
AP[a]	80.4 (71.1,87.8)	-	-	-	-	-	-
Recall	81.1 (70.6,90.1)	85.2	11.1	73.8	100.0	100.0	100.0
PPV[b]	66.8 (60.3,72.7)	65.5	0.0	50.0	68.3	100.0	100.0
True positives	159 (114,209)	3.1	0	1	2	3	19
False positives	79 (58,102)	1.5	0	0	1	2	8
False negatives	37 (18,59)	0.7	0	0	0	1	8
Classification output	anatomical location						
Performance metric							
Accuracy$_{suspicious}$	77.0 (70.0,83.4)	78.4	0.0	57.5	95.0	100.0	100.0
Accuracy$_{all}$	94.4 (92.4,96.1)	94.1	78.6	90.9	94.9	100.0	100.0

[a] Average Precision
[b] Positive predictive value

Tab. 2.3. PET uptake classification performance obtained with combined training of a convolutional neural network using ^{68}Ga-PSMA-11 PET/CT and ^{18}F-FDG PET/CT scans, evaluated on a hold-out test dataset of ^{68}Ga-PSMA-11 PET/CT scans. Performance metrics determined by pooling findings of all subjects together are reported. Summary statistics for performance metrics determined at per-subject level are also reported. 95% confidence intervals obtained via bootstrap resampling at subject level are reported in brackets.

suspicious uptake regions limited to the pelvis [132] or were trained on CT-only information to segment a predefined set of organs and then used to guide semi-automated identification of suspicious high uptake regions in the whole body [42][109].

In the current analysis, regions of interest were segmented both by the expert reader as well as for the network training and validation using methods based on thresholding, which allow limited flexibility and accuracy in delineating contours. Although the threshold-based segmentation methods used have limited accuracy, these offer a practical solution for rapid semi-automated annotation by an expert reader, they are often used in clinical practice and research investigations on metabolic tumor volume [3], as well as mentioned in procedure guidelines [7]. Nonetheless, efforts for standardizing and advancing segmentation techniques are ongoing, and machine learning based methods are promising for improving automated segmentation accuracy. Notably, for tumor segmentation in ^{18}F-FDG PET/CT images, machine learning methods have recently shown improved test-retest repeatability [90] and accuracy [56][131] compared to thresholding methods, as well as ability to delineate tumor regions in the whole body [64][6]. While our results with ^{68}Ga-PSMA-11 PET/CT images support the use of machine learning methods for identification and anatomical location classification of suspicious uptake sites, future analyses are required to evaluate the accuracy and repeatability of different segmentation methods in PSMA-ligand images for varying tumor sites in the whole body. Moreover, while different software implementations of threshold-based segmentation methods were reported to yield comparable results for metabolic tumor volume quantification in ^{68}Ga-PSMA-11 PET/CT scans [55], there may be variations in machine learning-based segmentation methods and the concordance and potential standardization of these should also be investigated.

A limited number of subjects with advanced prostate cancer was included in the analysis and these were used solely for the network training. Given the very low tumor burden of subjects in the test set, it was not possible within the context of this analysis to evaluate the ability of the proposed method to estimate total tumor volume within a wide range and in particular for subjects at an advanced stage, for which tumor burden may be more informative. Furthermore, the majority of uptake regions annotated as suspicious for prostate cancer were in lymph nodes or bone, while suspicious findings in other organs were limited. Since the network was trained to evaluate single regions of interest, it was possible to use PET/CT scans with only partial annotation of suspicious sites for training. This is beneficial since labeling can be highly time consuming in cases where a large number of lesions need to be fully annotated. Moreover, as the network was trained with a variety of regions of interest in the whole body, it may prove useful for the staging and tumor burden assessment also in subjects with an advanced disease, but this will need to be confirmed in future analyses.

The ground truth used to train and evaluate the proposed algorithm was determined by visual assessment of the images by an expert physician, while neither histopathology nor follow-up information was considered. Additionally, PET/CT image quality characteristics, such as pitfalls due to motion or artifacts, reconstruction settings and partial volume effects may influence the output of the network and results will require expert supervision for the use in clinical context. Despite the above limitations, the network showed good ability to identify even small suspicious sites with a limited number of false positives, compared to the expert evaluation. In this analysis, we found that combining training information from ^{18}F-FDG PET/CT and ^{68}Ga-PSMA-11 PET/CT led to improved accuracy for the identification and anatomical location classification of suspicious uptake sites. This result brings forward the promising perspective of a deep learning framework for supporting staging and tumor burden assessment in multiple cancer types with PET/CT images obtained using different tracers. Notably, an increasing variety of PET radiotracers is being clinically used and developed in oncology, with multiple alternative compounds undergoing clinical trials for PSMA-targeted imaging alone. On the one hand, the lesion anatomical spread pattern and tumor volume are meaningful biomarkers in different cancer types independently of the PET tracer used. On the other hand, with each compound having a different biodistribution, training distinct networks de novo as a separate solution for each tracer would require a significant number of image datasets and expert annotations to reach sufficient accuracy. Ideally, combining information from multiple diseases and tracers in a single network could create synergies, leveraging similarities in physiological uptake, tracer excretion patterns and tumor spread, while still accounting for differences based on the provided input encoding the tracer type. In the current analysis performance improvements when training with information from both ^{68}Ga-PSMA-11 PET/CT and ^{18}F-FDG PET/CT images were found mainly for ^{68}Ga-PSMA-11 PET/CT scans, having a smaller training dataset. Moreover, a significant improvement was found for the task of anatomical location classification, possibly driven mainly by the larger CT training information, while the performance increase in identification of suspicious uptake was less pronounced. The overall benefit of a combined training approach may depend on the level of similarity and the relative frequency of the different imaging findings between tracers, and future analyses will be required to evaluate the extensibility of the proposed framework to additional patient cohorts and radiotracers.

3 Prostate cancer staging based on PSMA-ligand imaging

This chapter discusses the evaluation of the uptake classification method described in Chapter 2 for automated assessment of prostate cancer nodal and metastatic stage based on PSMA-ligand PET/CT imaging. A convolutional neural network is used to identify sites with uptake suspicious for cancer and assign them an anatomical location classification relevant for staging. Based on the identified anatomical pattern of cancer spread, a stage category relevant for disease management is assigned following a standardized framework concordant with established staging guidelines. The agreement between the automated image analysis method and the visual evaluation by an expert reader is assessed. The chapter concisely introduces background information on prostate cancer staging. Methods used to evaluate the stage classification based on automated image analysis and expert assessment are detailed, corresponding results are described, findings and limitations are discussed.

Parts of this chapter have already been published and are partially quoted:

[13] N. Capobianco et al. “Whole-Body Lesion Detection and Prostate Cancer Staging in 68Ga-PSMA-11 PET/CT Using Deep Learning”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 47.S1 (Sept. 2020), pp. 273–274

[16] N. Capobianco et al. “Whole-Body Uptake Classification and Prostate Cancer Staging in 68Ga-PSMA-11 PET/CT Using Dual-Tracer Learning”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* (July 2021)

3.1 Introduction

In prostate cancer, the anatomical extent of disease spread is an important factor to evaluate prognosis as well as to select and tailor treatment. While localized disease confined to the prostate is associated with a generally favorable outcome and can be indolent in a subset of patients with low-grade tumors [118], in presence of metastases at the time of diagnosis prostate cancer is associated with a poor outcome and can constitute a threat to life [115]. Several clinical examinations are used to evaluate prostate cancer presence and progression including serum Prostate Specific Antigen (PSA) analysis, digital rectal examination, imaging, and histopathology. Based on the clinical information available, the anatomical extent of prostate cancer spread can be classified in fixed categories defined by staging guidelines, with respect to the primary tumor (T), local lymph nodes (N) and distant metastases (M). The categories defined by the American Joint Committee on Cancer (AJCC) staging manual for prostate cancer are reported in Tab. 3.1.

Primary Tumor, clinical (cT)	
TX	Primary tumor cannot be assessed
T0	No evidence of primary tumor
T1	Clinically inapparent tumor that is not palpable
T1a	Tumor incidental histologic finding in 5% or less of tissue resected
T1b	Tumor incidental histologic finding in more than 5% of tissue resected
T1c	Tumor identified by needle biopsy found in one or both sides, but not palpable
T2	Tumor is palpable and confined within prostate
T2a	Tumor involves one-half of one side or less
T2b	Tumor involves more than one-half of one side but not both sides
T2c	Tumor involves both sides
T3	Extraprostatic tumor that is not fixed or does not invade adjacent structures
T3a	Extraprostatic extension (unilateral or bilateral)
T3b	Tumor invades seminal vesicle(s)
T4	Tumor is fixed or invades adjacent structures other than seminal vesicles, such as external sphincter, rectum, bladder, levator muscles, and/or pelvic wall
Regional lymph nodes (N)	
NX	Regional lymph nodes were not assessed
N0	No positive regional lymph nodes
N1	Metastases in regional lymph node(s)
Distant metastases (M)	
M0	No distant metastasis
M1	Distant metastasis
M1a	Nonregional lymph node(s)
M1b	Bone(s)
M1c	Other site(s) with or without bone disease

Tab. 3.1. Prostate cancer TNM stage categories defined in the AJCC 8th edition cancer staging manual. Pathological T stage (pT) categories are omitted. Adapted from [9].

While histopathology remains the gold standard for prostate cancer staging, imaging is increasingly used for noninvasive disease evaluation. Notably, PSMA-ligand hybrid imaging has shown superior accuracy for prostate cancer staging compared to conventional imaging. In response to the increased adoption of PSMA-ligand imaging, the Prostate Cancer Molecular Imaging Standardized Evaluation (PROMISE) criteria has been proposed [34], defining a standardized reporting framework which includes the molecular imaging TNM system (miTNM) for image-based staging. The miTNM system is designed to parallel the clinicopathologic TNM framework. Additionally, categories describing the PSMA expression level for each lesion (miPSMA score) and the pattern of bone involvement are introduced to aid image interpretation, prognostication, and PSMA-targeted treatment planning.

The PROMISE system provides a framework to aid reproducibility of image interpretation and reporting, defining clinically meaningful image-based categorical variables and related diagnostic flowcharts. On the one hand, image evaluation may be operator-dependent and error-prone if it relies solely on visual inspection. On the other hand, the systematic use of manual measurements for each suspicious image region can be highly time consuming. We evaluated the use of automated image analysis, based on the uptake classification algorithm described in Chapter 2, to support prostate cancer staging following the PROMISE miTNM framework.

3.2 Materials and Methods

We assessed the ability of the uptake classification algorithm described in Chapter 2 to determine the N and M stage from ^{68}Ga-PSMA-11 PET/CT images fully automatically. For this purpose, annotations of elevated uptake sites assigned by an expert physician in 123 consecutive PET/CT scans of prostate cancer patients (group A), referred for primary staging or assessment of biochemical recurrence, were used to determine a stage category for each subject. Following PROMISE recommendations, a distinction between patterns of bone metastases was considered. No subject presented diffuse bone marrow involvement. This resulted in three N stage categories, related to regional lymph node metastases: N0 (none), N1 (single), N2 (multiple) and six M stage categories, related to distant metastases: M0 (none), M1a (extrapelvic lymph nodes), M1b/u (single bone lesion, unifocal), M1b/o (up to three multiple bone lesions, oligometastatic), M1b/d (four or more bone lesions, disseminated), M1c (other organs). The stage category was employed both as ground truth and for performing a stratified split determining a hold-out test set of 52 subjects. The remaining 71 subjects were used for training the uptake classification algorithm as detailed in Chapter 2. An illustration of the dataset split for the subjects in group A is represented in Fig. 2.1. A summary of the N and M stage for subjects in group A is reported in Tab. 3.2.

For each test set subject, we first segmented all regions with SUV$_{max}$ above 1 using an incremental connected component algorithm [11] and 45% of SUV$_{max}$ thresholding. These regions were then processed by the convolutional neural network, classified as nonsuspicious or suspicious and assigned an anatomical location label. The anatomical location labels of regions classified as suspicious were used to obtain a prediction of the N and M stage according to the PROMISE miTNM framework. Predictions of the N and M stage were then compared to the ones based on expert annotations.

Agreement between the N and M stage estimated using the CNN and determined from the expert labels was assessed using percent agreement and confusion matrices.

	Development (group A)	Test
Total	71	52
Stage		
miN0M0	8	6
miN1M0	11	8
miN2M0	5	3
miN0M1a	2	1
miN1M1a	3	3
miN2M1a	7	5
miN0M1b/u	11	8
miN1M1b/u	7	4
miN2M1b/u	1	1
miN0M1b/o	3	2
miN1M1b/o	1	1
miN2M1b/o	2	1
miN0M1b/d	5	4
miN1M1b/d	1	1
miN2M1b/d	2	2
miN0M1c	1	1
miN2M1c	1	1

Tab. 3.2. Summary of the N and M stage assigned based on expert reader annotation of PSMA-ligand PET/CT images according to the PROMISE miTNM framework, for subjects in group A.

3.3 Results

Based on the expert annotations of subjects in group A, 52 patients had miN0 stage, 40 had miN1 stage, 31 had miN2 stage, whereas 41 subjects had miM0 stage, 21 had miM1a stage, 57 had miM1b stage and 4 had miM1c stage.

Fig. 3.1 shows an example subject in the test dataset assessed using the CNN. After assigning an N stage based on CNN annotations and based on expert annotations, agreement was 67%, while agreement for identification of any pelvic nodal involvement (N0 vs N1/N2) was 81%. The confusion matrix for the N stage assessment is shown in Tab. 3.3. After assigning an M stage based on CNN annotations and based on expert annotations, agreement was 62%, agreement excluding discrimination of bone involvement pattern was 73% and agreement for identification of any distant metastases (M0 vs M1) was 77%. The confusion matrix for the M stage assessment is shown in Tab. 3.4.

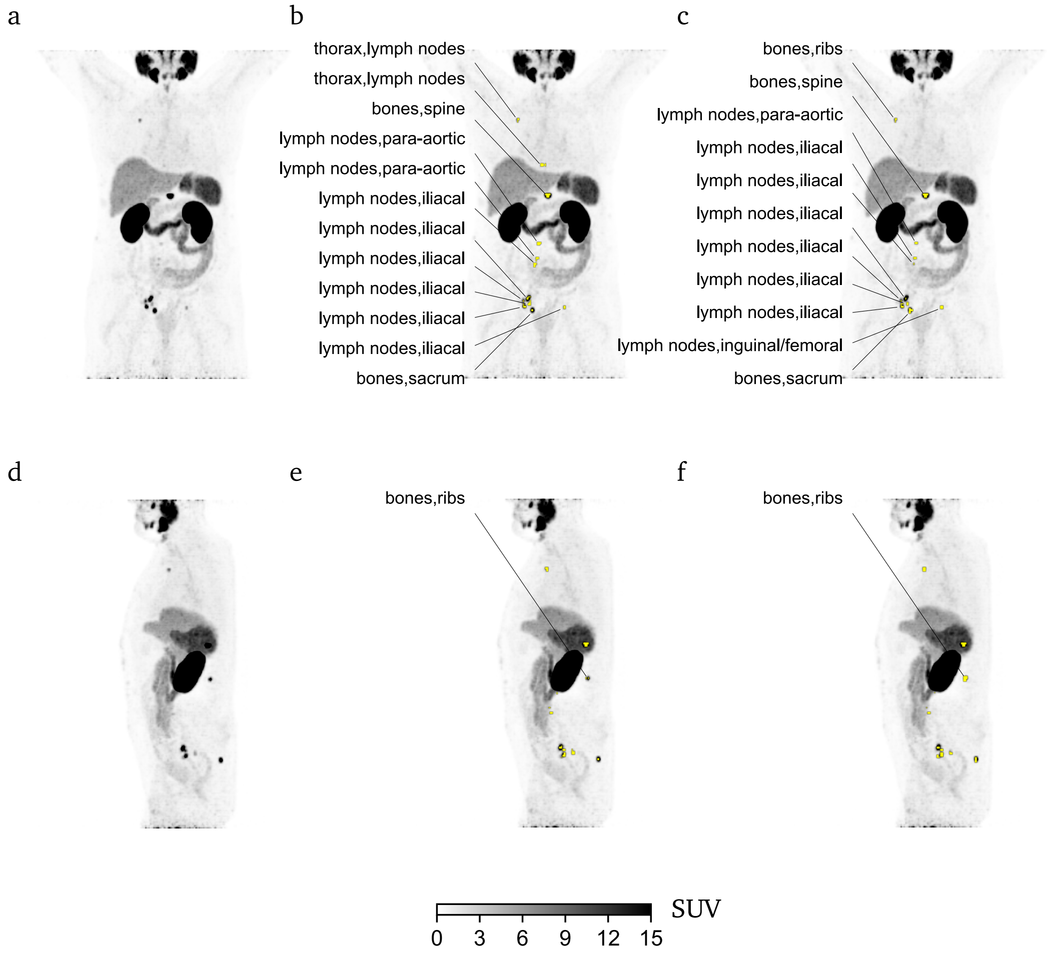


Fig. 3.1. (a) coronal and (d) sagittal Maximum Intensity Projections (MIP) of a ^{68}Ga-PSMA-11 PET scan for a subject in the test set. (b, e) Regions of interest classified by the convolutional neural network as suspicious overlayed to the PET MIP in yellow, together with the anatomical location label assigned by the network. (c, f) Regions of interest identified by an expert physician as suspicious uptake overlayed to the PET MIP in yellow, together with the anatomical location label assigned by the expert.

Predicted stage	miN0	miN1	miN2	Total
Annotations stage				
miN0	14	7	1	22
miN1	2	11	4	17
miN2	0	3	10	13
Total	16	21	15	52

Tab. 3.3. Confusion matrix comparing the N stage determined according to the PROMISE miTNM framework based on expert annotations and based on convolutional neural network annotations.

Predicted stage	miM0	miM1a	miM1b/u	miM1b/o	miM1b/d	miM1c	Total
Annotations stage							
miM0	8	5	4	0	0	0	17
miM1a	2	6	1	0	0	0	9
miM1b/u	1	0	7	4	0	1	13
miM1b/o	0	0	0	3	1	0	4
miM1b/d	0	0	0	1	6	0	7
miM1c	0	0	0	0	0	2	2
Total	11	11	12	8	7	3	52

Tab. 3.4. Confusion matrix comparing the M stage determined according to the PROMISE miTNM framework based on expert annotations and based on convolutional neural network annotations.

3.4 Discussion

In this survey we showed that an automated image analysis method, employing a convolutional neural network to identify sites of suspicious tracer uptake and classify their anatomical location, was able to determine the N and M prostate cancer stage according to the PROMISE miTNM framework fully automatically in fair concordance with the expert evaluation.

Our analysis focused on prostate cancer staging with respect to regional lymph nodes and distant metastases, while the assessment of the primary tumor extent was not addressed. For clinical staging of the primary tumor, Digital Rectal Examination is the recommended examination given its wide accessibility [9]. Moreover, PSMA-ligand imaging is typically employed in cases where there is a suspicion of metastases prior to initial treatment, or in cases of suspected recurrence based on high PSA levels after primary treatment, to identify potential secondary sites of tumor spread relevant for treatment planning. Nevertheless, combined PSMA-ligand and anatomical imaging can provide useful clinical information for the primary tumor. Notably, PSMA-ligand PET/CT [77] and PET/MRI [40] have shown the ability to detect intraprostatic tumors, suggesting a potential role in guiding biopsy in a subset of patients with improved accuracy compared to conventional approaches. Furthermore, invasion of the primary tumor in adjacent organs for T staging can be assessed with PSMA-ligand PET/MRI [50]. While threshold-based image analysis methods are more frequently used, machine learning methods also showed utility for segmentation of intraprostatic tumors in PSMA-ligand PET [72], and additional analyses are required to further investigate the potential of machine learning methods for primary prostate cancer delineation and T staging based on multimodal imaging.

We found a moderate level of agreement between N and M stage assigned based on the neural network annotations and expert annotations, which decreased when considering a higher number of stage categories corresponding to increased anatomical granularity. In terms of automated image analysis, correctly identifying stage is a particularly challenging task given that the misclassification of a single elevated uptake region, either as false positive, false negative, or improper anatomical location, can result in up-staging or under-staging.

Equivocal tracer uptake potentially resulting from known pitfalls of PSMA expression in non-prostate cancer tissue or due to artifacts may be incorrectly classified by the network and results will require supervision by an expert physician to guide clinical decisions. Nevertheless, the proposed method was able to correctly label a good majority of suspicious uptake sites relevant for staging and could be employed to assist a human reader in evaluating images as well as in performing detailed measurements, only requiring manual input for a limited set of improperly classified high uptake regions.

The present analysis was retrospective and employed a restricted single-center test cohort of prostate cancer patients referred to PET/CT imaging for primary staging or assessment of biochemical recurrence. While a stratified split based on stage was performed to determine the hold-out test set, the stage categories were nonuniformly represented, and distant metastases other than in bone or lymph nodes were strongly underrepresented. Moreover, PET and CT images had relatively uniform image quality characteristics and were acquired with a single scanner. Future analyses should further evaluate the proposed method with respect to different image quality characteristics and including a broader range of disease progression, ideally considering a multi-center cohort.

The ground truth used to determine the N and M stage in the current analysis was neither based on histopathology nor follow-up information but only on visual assessment by a single physician. Recent analyses have found good intrareader and inter-reader agreement for visual interpretation of ^{68}Ga-PSMA-11 PET images [39][124], supporting the reproducibility of image evaluation when performed by an expert reader following standardized criteria. In the present analysis we reported results based on the recently described PROMISE framework, while multiple other evaluation systems have been proposed [101][36]. Nevertheless, the miTNM framework has the advantage of mirroring the broadly recognized TNM classification and was recently supported by consensus guidelines for PSMA PET reporting [18].

4 Total Metabolic Tumor Volume estimation in lymphoma

This chapter describes the evaluation of deep learning-based tracer uptake classification in ^{18}F-FDG PET/CT images for automated estimation of Total Metabolic Tumor Volume (TMTV) in patients with lymphoma. Baseline TMTV has potential as prognostic factor to identify high-risk patients who may benefit from more intensive treatment strategies. However, determining TMTV involves the segmentation of all malignant foci throughout the body, a task which requires extensive manual input from an experienced reader when performed with commonly employed image analysis methods. In the present retrospective analysis, we evaluated an automated method for TMTV estimation in a cohort of patients with diffuse large B-cell lymphoma from a multi-center clinical trial. Background information on TMTV estimation in patients with lymphoma is first introduced. Methods used to determine TMTV, by combining a high-uptake segmentation algorithm with deep learning-based classification to discard physiological uptake regions, are described. Techniques employed to compare the TMTV estimations obtained automatically and determined semi-automatically by experts, as well as their respective prognostic value are detailed. Results related to the uptake classification, TMTV segmentation, and survival analysis are reported. Finally, key findings and limitations are discussed.

Substantial parts of this chapter have already been published and are quoted verbatim:

[15] N. Capobianco et al. "Fully Automated Deep Learning FDG Uptake Classification Enables Total Metabolic Tumor Volume (MTV) Estimation in Diffuse Large B-Cell Lymphoma with Similar Predictive Value as Expert MTV Measurements". In: *Journal of Nuclear Medicine* 61.supplement 1 (May 2020), p. 504

[14] N. Capobianco et al. "Deep-Learning ^{18} F-FDG Uptake Classification Enables Total Metabolic Tumor Volume Estimation in Diffuse Large B-Cell Lymphoma". en. In: *Journal of Nuclear Medicine* 62.1 (Jan. 2021), pp. 30–36

4.1 Introduction

Total metabolic tumor volume derived from ^{18}F-FDG PET/CT baseline scans is a promising prognostic factor in diffuse large B-cell lymphoma (DLBCL) [104][117] and other types of lymphoma [68][27][82]. DLBCL is the most frequent non-Hodgkin lymphoma, being present in about 30%–40% of non-Hodgkin lymphoma cases worldwide. Although the prognosis of DLBCL can be improved with immunochemotherapy, more than 30% of patients are refractory or relapse after first-line treatment, with a poor outcome [46][28]. Therefore, there is a need to identify high-risk patients who could benefit from intensive or novel therapies early.

Unfortunately, the role of current prognostic factors such as the International Prognostic Index [63], Revised International Prognostic Index [108], and National Comprehensive Cancer Network International Prognostic Index [133], based on tumor burden surrogates is limited. Thus, baseline TMTV, which estimates the total metabolic tumor burden at diagnosis, has been proposed as an alternative prognostic tool for early risk stratification.

To date, TMTV is not yet routinely used in clinical lymphoma patient management, in part because of a lack of consensus throughout the literature. Several methods have been proposed to calculate TMTV [24][62][3], and the cutoffs reported to detect high risk patients differed among methods and investigations. However, recent analyses have suggested that, despite these differences, most methods yielded similar accuracy in predicting patient prognosis when applied in similar patient groups [24][62], emphasizing the strong prognostic power of baseline TMTV.

Regardless of the criteria used for delineating tumor regions, all methods for deriving TMTV require extensive and time-consuming manual input from an experienced reader. The reader either manually segments the tumor regions or, more commonly, uses an automated method to detect all regions with increased uptake and then manually eliminates the regions of physiologic uptake and adds in undetected tumor regions [3]. Recently, a machine-learning algorithm using a convolutional neural network (CNN) was trained to differentiate physiologic from nonphysiologic uptake regions in whole-body ^{18}F-FDG PET scans acquired from an unselected population of more than 600 patients, including half who were lymphoma patients with different subtypes of diseases [113][114]. This CNN achieved a high degree of accuracy in characterizing increased tracer uptake in the whole body as physiologic or nonphysiologic. Such automated identification of nonphysiologic regions would facilitate TMTV measurement and clinical adoption. This analysis therefore sought to assess the ability of this CNN to identify regions from which TMTV could be automatically calculated and to evaluate the ability of the resulting TMTV in predicting patient outcome among a large group of DLBCL patients included in an international phase III trial wherein TMTV has already been demonstrated to be a strong predictor of 4-y progression-free survival (PFS) and overall survival (OS). To evaluate the CNN performance, regions with elevated tracer uptake automatically identified as physiologic or suspicious were compared with regions attributed to suspicious uptake by an expert reader using a semiautomatic method.

4.2 Materials and Methods

4.2.1 Patients

Patients from an ancillary analysis [21][128] of the REMARC trial (NCT01122472) were retrospectively analyzed. This trial is a phase III investigation that was designed to assess the efficacy of lenalidomide versus placebo in responding elderly DLBCL patients (60–80 y old) treated with the standard first-line rituximab, cyclophosphamide, doxorubicin hydrochloride (hydroxydaunorubicin), vincristine sulfate, and prednisone (R-CHOP) therapy approach [123]. The institutional review board approval and the informed consent of the REMARC trial included all the ancillary investigations. The ancillary analysis was conducted by involving

301 patients who underwent baseline PET/CT before R-CHOP and showed that TMTV was a strong prognosticator of outcome in patients responding to first-line chemotherapy combined with monoclonal antibody treatment.

4.2.2 Image Acquisition and Analysis

All baseline ^{18}F-FDG PET/CT images from the ancillary analysis were collected in an anonymized DICOM format. Patients whose PET or CT DICOM series had incomplete axial slices or irregular slice intervals were excluded. PET images were expressed in SUV units, accounting for injected dose and patient body weight.

PET/CT images were analyzed using an investigational software prototype (PET Assisted Reporting System [PARS]; Siemens Medical Solutions USA, Inc.) that uses artificial intelligence. The prototype first automatically located a cylindric reference region at the center of the proximal descending aorta by applying a landmarking algorithm to the CT image [122]. This region was used to determine the mean blood pool SUV and mean blood pool SUV standard deviation (SD), following PERCIST recommendations [130]. The 3-dimensional regions of the PET image with increased tracer uptake were identified for each subject using an automated whole-body high-uptake segmentation algorithm (multi-foci segmentation, MFS) [127]. In line with the PERCIST recommendations, only the regions with SUV$_{peak}$ greater than twice the mean blood pool SUV plus twice the mean blood pool SUV standard deviation were included. Those regions were then further segmented according to 42% of the SUV$_{max}$ threshold, and the ones with volumes below $2\,\mathrm{cm}^3$ were discarded. The resulting regions of interest (ROIs), called ROI$_{PARS}$, were then automatically processed by a CNN. Details of the training and validation of this CNN were previously reported [114]. The input of the CNN was the PET/CT data together with the set of ROI$_{PARS}$ sites. For each ROI$_{PARS}$, the output of the CNN was the anatomic localization among a set of possible anatomic sites relevant for staging and whether the ROI$_{PARS}$ uptake was physiologic (e.g., due to unspecific bowel uptake, muscle activation, inflammation, infection, or bone degeneration) or suspicious (i.e., due to lymphoma). The volumes of all ROI$_{PARS}$ sites classified as suspicious uptake were then summed to obtain the TMTV$_{PARS}$.

The CNN was also used in combination with two other settings of the initial high-uptake ROI segmentation: the first used an initial threshold of 2.5 SUV instead of the blood-pool–based threshold, followed by thresholding with 41% of SUV$_{max}$; the second also included ROIs with a volume between 0.1 and $2\,\mathrm{cm}^3$.

The TMTV obtained by 2 experienced nuclear medicine physicians in the context of a previous analysis [21][128] was used as a reference (TMTV$_{REF}$). The TMTV$_{REF}$ was obtained using the semiautomatic version of the Beth Israel Fiji (ImageJ) software plugin [69], which was previously used to demonstrate the prognostic value of TMTV in various lymphoma subtypes [82][26]. To calculate TMTV$_{REF}$, the physician combined automated and manual steps as follows. First, volumes of interest with high uptake in the PET images were segmented using an automated method, which applied in sequence an algorithm based on component trees and shape priors [48], a region growing, and a final region delineation using 41% of the region SUV$_{max}$ threshold [84]. Second, the resulting ROIs were manually reviewed by the reader, who selected only the regions corresponding to lymphoma (ROI$_{REF}$), adding an ROI$_{REF}$ wherever a lymphoma lesion had been missed by the algorithm by drawing a prism around that lesion and applying a 41% SUV$_{max}$ threshold. The volumes of all lymphoma ROI$_{REF}$ sites were summed to obtain the reference TMTV (TMTV$_{REF}$).

4.2.3 Statistical Analysis

To evaluate the performance of the CNN classification, for each patient, each ROI$_{PARS}$, having been labeled as presenting suspicious or physiologic uptake by the CNN, was compared with all the ROI$_{REF}$ sites of that patient taken together. The ROI$_{PARS}$ was considered to match the ROI$_{REF}$ if at least 50% of its volume overlapped with one or several ROI$_{REF}$ sites. ROI$_{PARS}$ sites classified as suspicious and matching one or several ROI$_{REF}$ sites were considered true-positives, ROI$_{PARS}$ sites classified as physiologic and matching one or several ROI$_{REF}$ sites were considered false-negatives, ROI$_{PARS}$ sites classified as physiologic and not matching any ROI$_{REF}$ sites were considered true-negatives, and ROI$_{PARS}$ sites classified as suspicious and not matching any ROI$_{REF}$ sites were considered false-positives. The sensitivity, specificity, and accuracy of the uptake classification were calculated. The performance of the CNN classification was also assessed in case a minimum overlap of 25% and 75% was required to consider an ROI$_{PARS}$ as matching the ROI$_{REF}$.

To evaluate differences between TMTV$_{PARS}$ and TMTV$_{REF}$, Bland– Altman analysis was performed. Since the Shapiro–Wilk test revealed a significant nonnormal distribution of the differences between TMTV$_{PARS}$ and TMTV$_{REF}$ ($P < 0.001$), the median bias and limits of agreement at the 2.5 and 97.5 percentiles were reported in the Bland– Altman plot. To assess the correlation between ranked TMTV values, the Spearman rank correlation coefficient was used. For each patient, the agreement between the patient set of ROI$_{PARS}$ sites classified as suspicious and the patient set of ROI$_{REF}$ sites was characterized using the Dice score, precision (the fraction of voxels in the set of ROI$_{PARS}$ sites classified as suspicious that were also present in the set of ROI$_{REF}$ sites), and recall (the fraction of voxels in the set of ROI$_{REF}$ sites that were also present in the set of ROI$_{PARS}$ sites classified as suspicious).

Survival analysis was performed for both TMTV$_{PARS}$ and TMTV$_{REF}$ with respect to PFS and OS. Receiver-operating-characteristic curves were used to determine TMTV cutoffs to predict the occurrence of events within 4 y for both PFS and OS, by maximizing the Youden index (sensitivity + specificity - 1). Survival functions were computed by Kaplan–Meier analyses and used to estimate survival time statistics (such as 4-y PFS rate and 4-y OS rate) for low- and high-TMTV groups. A log-rank test was used to assess whether differences between Kaplan–Meier survival curves were significant. Univariate Cox regression was used to calculate hazard ratios between survival groups. Statistical significance was set at a P value of less than 0.05. Statistical analysis was performed using R, version 3.6.1, with survivalROC, version 1.0.3, and pROC, version 1.15.3 [99].

4.3 Results

In total, 280 patients from 124 centers were included in the analysis. Patient characteristics are reported in Tab. 4.1. All received first-line treatment with R-CHOP and were responders at the time of inclusion in the trial, 142 received a lenalidomide regimen afterward as maintenance, and 138 received placebo. After a median follow-up of 5 y, 86 patients presented with a PFS event and 51 patients had an OS event; the 4-y survival rates were 69% for PFS and 83% for OS. The 4-y survival rates were comparable to those of the entire trial.

Patient characteristics	Data (%)
Sex	
Female	119 (42.5)
Male	161 (57.5)
Age (median, ranges) years	68 (58-80)
Ann Arbor Stage	
I	1 (0.4)
II	25 (8.9)
III	57 (20.4)
IV	197 (70.4)
Performance status (ECOG)	
0	113 (40.4)
1	119 (42.5)
2	39 (13.9)
3	2 (0.7)
4	2 (0.7)
Missing	5 (1.8)
IPI	
1	6 (2.1)
2	73 (26.1)
3	97 (34.6)
4	81 (28.9)
5	19 (6.8)
Missing	4 (1.4)
Elevated LDH (>Upper limit of normal*)	
No	111 (39.6)
Yes	165 (58.9)
Missing	4 (1.4)

*LDH upper limit set specifically for each laboratory

Tab. 4.1. Patient characteristics.

PET/CT images were acquired using different scanner models from different vendors as summarized in Tab. 4.2. The delay between injection and acquisition time was $(71.7 \pm 14.1)\,\mathrm{min}$ (mean ± std). The SUV$_{mean}$ in the proximal descending aorta cylindric region was 1.6 ± 0.5 (mean ± std across subjects), resulting in an SUV$_{peak}$ threshold of 3.6 ± 1.2 for detecting ROIs with increased tracer uptake.

PET/CT scan characteristics	Data
Injected dose (MBq)	309 ± 87 (mean ± std)
Post injection scan delay (min)	71.7 ± 14.1 (mean ± std)
PET slice thickness (mm)	Median: 3.7; min-max: 2.0–5.0
PET pixel spacing (mm)	Median: 4.0; min-max: 2.3–5.5
CT slice thickness (mm)	Median: 3.00; min-max: 1.25–8.00
CT pixel spacing (mm)	Median: 1.17; min-max: 0.86–1.52
PET/CT scanner model	
General Electric (all)	72
Discovery 690	40
Discovery STE	14
Discovery ST	8
Discovery RX	4
Discovery 600	3
Discovery 710	2
Discovery LS	1
Siemens (all)	105
Biograph HiRez (1080)	40
Biograph Truepoint (1093,1094)	27
Biograph mCT	25
Biograph LSO (1023,1024)	8
Biograph BGO (1062)	5
Philips (all)	103
Gemini TF	38
Gemini GXL	36
Allegro Body	19
Unspecified (Philips)	10

Tab. 4.2. PET/CT scan characteristics.

The results below are described for the PERCIST-based setting of the initial high-uptake ROI segmentation, whereas changes observed with other settings are reported in Tab. 4.5, Tab. 4.7 and Tab. 4.8.

4.3.1 Uptake Classification

In total, 6,737 ROI$_{PARS}$ sites exhibiting increased uptake were obtained from the 280 subjects. There were 7,996 ROI$_{REF}$ sites in the 280 subjects. Descriptive statistics for the number of ROI$_{PARS}$ and ROI$_{REF}$ sites per subject are summarized in Tab. 4.3. Among the 6,737 ROI$_{PARS}$ sites with increased uptake, 2,831 (42%) were classified as having suspicious uptake by the CNN.

	ROI$_{PARS}$	ROI$_{REF}$
Total number of ROI	6737	7996
Average number of ROI per subject (min–max)	24.1 (2–91)	28.6 (1-201)
Median number of ROI per subject (IQR)	19.0 (13.0–31.2)	16.0 (6.0-38.0)

Tab. 4.3. Descriptive statistics related to the number of ROI$_{PARS}$ and ROI$_{REF}$ in the 280 subjects included in the analysis.

When compared with the ROI$_{REF}$ sites obtained by the experienced reader, the identification of the ROI$_{PARS}$ sites with suspicious uptake by the CNN yielded 3,317 true-negatives, 2,399 true-positives, 589 false-negatives, and 432 false-positives. Corresponding sensitivity was 80%, specificity was 88%, and accuracy was 85%.

Additionally, the mean per-subject ROI$_{PARS}$ classification accuracy was 87% (median, 89%; interquartile range [IQR], 81%– 96%). There were an average of 20 correctly classified ROI$_{PARS}$ sites per subject (median, 17 ROI$_{PARS}$ sites; IQR, 11–27 ROI$_{PARS}$ sites) and an average of 4 incorrectly classified ROI$_{PARS}$ sites per subject (median, 2 ROI$_{PARS}$ sites; IQR, 1–5 ROI$_{PARS}$ sites), which were regions classified as suspicious by the CNN that did not overlap with the set of ROI$_{REF}$ sites or regions classified as physiologic by the CNN but overlapped with the set of ROI$_{REF}$ sites. Two examples of uptake classification of ROI$_{PARS}$ sites with corresponding ROI$_{REF}$ are shown in Fig. 4.1. Results with a minimum overlap of 25% and 75% required to consider a ROI$_{PARS}$ as matching the ROI$_{REF}$ are reported in Tab. 4.4.

4.3.2 TMTV

After discarding the ROI$_{PARS}$ sites classified as physiologic uptake by the CNN, a median TMTV$_{PARS}$ of $110\,\mathrm{cm}^3$ was obtained (IQR, 33–$281\,\mathrm{cm}^3$). The median TMTV$_{REF}$ was $240\,\mathrm{cm}^3$ (IQR, 80–$529\,\mathrm{cm}^3$) (Tab. 4.6).

There was a significant correlation between ranked TMTV estimates (ρ=0.76; P < 0.001). The median Dice score across all patients between the patient set of ROI$_{PARS}$ sites labeled as suspicious and the patient set of ROI$_{REF}$ sites was 0.73 (IQR, 0.33– 0.86), the median recall of the patient set of ROI$_{PARS}$ sites labeled as suspicious with respect to the patient set of ROI$_{REF}$ sites was 0.62 (IQR, 0.20–0.81), and the median precision was 0.96 (IQR, 0.86–0.99). The Bland–Altman plot comparing TMTV$_{PARS}$ and TMTV$_{REF}$ (Fig. 4.2) showed wide limits of agreement.

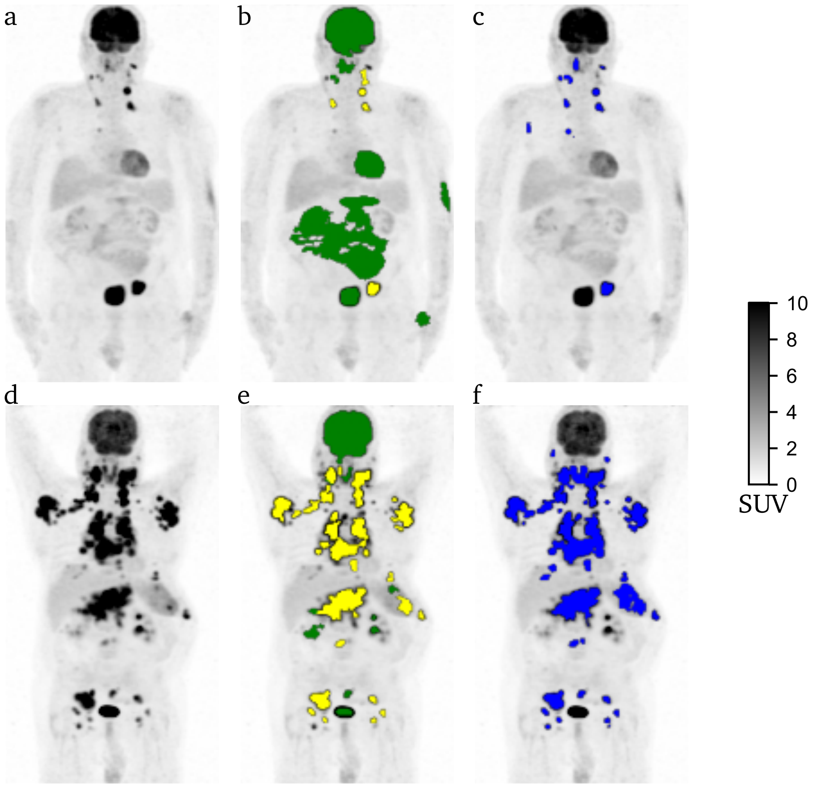


Fig. 4.1. Detection and classification of high ^{18}F-FDG uptake regions as physiological or suspicious. (a, d) Maximum-intensity projection (MIP) PET images of two subjects with low TMTV (a) and high TMTV (d). (b, e) ROI$_{PARS}$ obtained automatically using the PARS software prototype. ROI$_{PARS}$ detected by the MFS algorithm are overlaid on to the PET MIP. ROI$_{PARS}$ classified by the deep learning algorithm as physiological are shown in green, and ROI$_{PARS}$ classified as suspicious are shown in yellow. (c, f) ROI$_{REF}$ regions obtained by an experienced nuclear medicine physician using a semiautomatic software.

4.3.3 Survival Analysis

The area under the receiver-operating-characteristic curve for predicting the 4-y PFS was 0.63 for TMTV$_{PARS}$ and 0.69 for TMTV$_{REF}$ (Fig. 4.3). The optimal cutoffs for predicting the 4-y PFS were $171\,\mathrm{cm}^3$ for TMTV$_{PARS}$ and $242\,\mathrm{cm}^3$ for TMTV$_{REF}$. Kaplan–Meier survival curves are shown in Fig. 4.4. The 4-y PFS rates were 79% and 54% for the low- and high-TMTV$_{PARS}$ groups and 83% and 55% for the low- and high-TMTV$_{REF}$ groups, respectively. The log-rank test indicated a significantly longer PFS time in the low-TMTV patient group for both TMTV estimation methods ($P < 0.001$ for TMTV$_{PARS}$ and TMTV$_{REF}$). Cox regression for PFS resulted in hazard ratios (high-TMTV group vs. low-TMTV group) of 2.3 (95% confidence interval, 1.5–3.6; $P < 0.001$ for Wald test) for TMTV$_{PARS}$ and 2.6 (95% confidence interval, 1.6–4.1; $P < 0.001$) for TMTV$_{REF}$. The survival results are summarized in Tab. 4.9.

For the 4-y OS, the area under the receiver-operating-characteristic curve was 0.65 for TMTV$_{PARS}$ and 0.68 for TMTV$_{REF}$. The optimal TMTV cutoffs for predicting the 4-y OS were $148\,\mathrm{cm}^3$ for TMTV$_{PARS}$ and $223\,\mathrm{cm}^3$ for TMTV$_{REF}$. The 4-y OS rates were 90% and 74% for the low- and high-TMTV$_{PARS}$ groups and 93% and 74% for the low- and high-TMTV$_{REF}$ groups, respectively. The log-rank test revealed a significantly higher OS time in the low-TMTV patient group for both TMTV estimation methods ($P < 0.001$ for TMTV$_{PARS}$ and TMTV$_{REF}$).

	Overlap≥25%	Overlap≥50%	Overlap≥75%
Overall accuracy	0.85	0.85	0.84
Overall sensitivity	0.79	0.80	0.81
Overall specificity	0.91	0.88	0.85
Average misclassified number of ROI$_{PARS}$ per subject (min–max)	3.5 (0-61)	3.6 (0-60)	3.9 (0-53)
Median misclassified number of ROI$_{PARS}$ per subject (IQR)	2.0 (1.0-4.0)	2.0 (1.0-5.0)	2.0 (1.0-5.0)
Average classification accuracy per subject (min–max)	0.88 (0.33-1.00)	0.87 (0.34-1.00)	0.86 (0.42-1.00)
Median classification accuracy per subject (IQR)	0.90 (0.82-0.96)	0.89 (0.81-0.96)	0.88 (0.80-0.94)

Tab. 4.4. Results associated with the classification of ROIs with uptake significantly above the blood pool and volume above $2\,\mathrm{mL}$, when different levels of overlap are required to consider a ROI as matching the reference TMTV region.

Cox regression for OS resulted in hazard ratios (high-TMTV group vs. low-TMTV group) of 2.8 (95% confidence interval, 1.6–5.1; P < 0.001) for TMTV$_{PARS}$ and 3.7 (95% confidence interval, 1.9–7.2; P < 0.001) for TMTV$_{REF}$.

The sensitivity, specificity, negative predictive value, positive predictive value, and accuracy for predicting the occurrence of survival events within 4 y, determined at the optimal TMTV cutoff for each method, are reported in Tab. 4.10 and were similar for both PFS and OS.

4.4 Discussion

Our main result was that a fully automated method combining a region delineation method based on PERCIST recommendations and a CNN-based algorithm to distinguish between regions with elevated physiologic uptake and nonphysiologic regions was able to generate, in a uniform population of DLBCL patients, TMTV values predictive of 4-y PFS and OS with an accuracy comparable to that obtained when TMTV is calculated by manual selection of the tumor regions by medical experts. Although the CNN-based algorithm was trained using images obtained on only 2 scanner models from the same vendor, the algorithm was highly accurate in classifying increased uptake in patients from an international trial involving 124 centers that obtained images on different scanner models from different vendors and with variable reconstruction settings. This accuracy underlines the robustness of the CNN despite different image quality. Moreover, this algorithm was not originally trained for TMTV computation and outcome prediction and was developed with data from patients with different lymphoma subtypes and lung cancer who underwent PET at baseline and for response assessment. However, we showed that the algorithm was successful in a group of patients with a homogeneous lymphoma subtype scanned at baseline, enabling the identification of a TMTV cutoff separating high-risk from low-risk patients and predicting prognosis with accuracy comparable to that of the reference method. No subject was excluded because of failure of the initial high-uptake ROI segmentation, which identified at least one high-uptake region in

	SUV_{max} >2.5 (vol>2 mL)	SUV_{peak} >Blood Pool (vol>2 mL)
Total number of MFS* findings (ROI_{PARS})	18674	6737
Average number of ROI_{PARS} per subject (min–max)	66.7 (6–242)	24.1 (2–91)
Median number of ROI_{PARS} findings per subject (IQR)	59.0 (39.0–86.0)	19.0 (13.0–31.2)
Average misclassified number of ROI_{PARS} per subject (min–max)	6.9 (0–73)	3.6 (0–60)
Median misclassified number of ROI_{PARS} per subject (IQR)	4.0 (2.0–9.0)	2.0 (1.0–5.0)
Overall accuracy	0.90	0.85
Overall sensitivity	0.79	0.80
Overall specificity	0.92	0.88
Average classification accuracy per subject (min–max)	0.90 (0.40–1.00)	0.87 (0.34–1.00)
Median classification accuracy per subject (IQR)	0.93 (0.86–0.97)	0.89 (0.81–0.96)

*Multi-foci segmentation

Tab. 4.5. Results associated with the classification of high-uptake ROIs for two different groups of ROIs obtained with two different settings of the multi-foci segmentation algorithm.

TMTV Estimation	Mean	STD	Min	Q1 (25%)	Median	Q3 (75%)	Max
$TMTV_{PARS}$ (mL)	235.2	347.6	0.0	32.9	110.2	280.8	2471.9
$TMTV_{REF}$ (mL)	433.7	571.3	2.27	80.0	240.0	529.3	3832.7

Tab. 4.6. Descriptive statistics of TMTV obtained using the software prototype PARS ($TMTV_{PARS}$) and the reference method for the 280 subjects included in the analysis.

all subjects. Furthermore, comparable results were obtained when different settings of the initial high-uptake ROI segmentation were applied using a lower threshold (2.5 SUV) than the PERCIST-recommended blood-pool–based threshold (Tab. 4.5 and Tab. 4.7), suggesting the robustness of the algorithm to the initial segmentation results. Additionally, the accuracy of the high-uptake ROI classification was not substantially impacted when a different level of overlap was required to consider an ROI as matching the $TMTV_{REF}$ and when ROIs with volumes of less than $2\,\mathrm{cm}^3$ were included in the analysis (Tab. 4.4 and Tab. 4.8).

The median $TMTV_{PARS}$ and the resulting cutoff were lower than those observed for $TMTV_{REF}$. This finding could be due to multiple factors, including the higher initial SUV threshold used for $TMTV_{PARS}$ relative to the one used for $TMTV_{REF}$, the manual addition of suspicious regions with low uptake in $TMTV_{REF}$, regions being classified as physiologic in $TMTV_{PARS}$ but considered suspicious in $TMTV_{REF}$, and differences in the contouring of suspicious regions between $TMTV_{PARS}$ and $TMTV_{REF}$. However, the ability of the $TMTV_{PARS}$ estimates to be predictive of PFS and OS despite involving a TMTV range different from that of $TMTV_{REF}$ is consistent with what has already been reported [24][62] when comparing different TMTV estimation methods. This result confirms both the validity of the CNN method and the value of TMTV as a prognostic indicator.

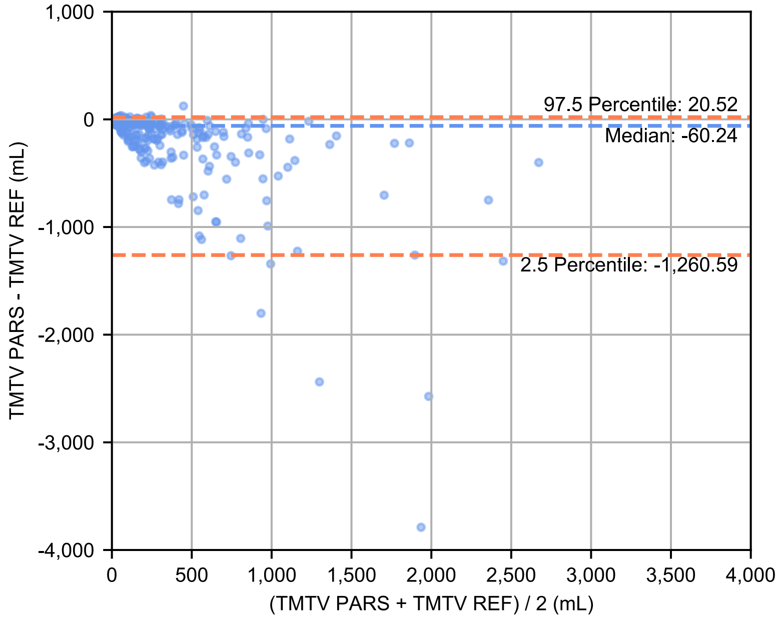


Fig. 4.2. Bland–Altman plot comparing fully automated and reference TMTV estimations. Bland–Altman plot comparing the TMTV obtained using the software prototype PARS (TMTV$_{PARS}$) and the reference TMTV (TMTV$_{REF}$) obtained by a nuclear medicine physician using a semiautomatic software.

Our analysis had limitations. Results of the receiver-operating-characteristic curve analysis and survival rates were reported for a four-year follow up time, whereas for longer follow-up intervals survival information was censored in the majority of subjects, resulting in larger confidence intervals for survival rates as visualized in the Kaplan-Meier curves. Since there is currently no gold standard method for TMTV calculation from ^{18}F-FDG PET/ CT images [23], the reported figures of merit supporting the uptake classification performance and accuracy of the TMTV segmentation are limited to the comparison with the reference method considered in the analysis. Moreover, a uniform cohort of lymphoma patients was evaluated in the current investigation, and results may differ for different lymphoma subtypes or different cancer types.

In the present work, we evaluated a fully automated application of PARS. However, PARS was initially intended to be used in a supervised manner, allowing the reader to correct for potentially misclassified regions when appropriate. In particular, pitfalls in PET/CT image quality, such as misalignment due to motion or image artifacts, may influence the classification output of the CNN algorithm, and the results should be validated by an expert. This is especially true when the labeling results are used to derive a prognostic index such as TMTV that can be used to stratify the risk and guide personalized therapy. Nevertheless, this approach could be used by expert readers to efficiently estimate TMTV, as the deep learning–based method is able to automatically identify several relevant suspicious uptake sites and automatically discard physiologic uptake sites, with the expert only having to correct the potential improper classification of a limited number of regions per subject, requiring limited user interaction and potentially improving inter-reader variability. This approach may introduce bias in the TMTV estimation process by relying on pre-generated results. However, this risk should be marginal, especially when a careful revision of the results is performed by an experienced reader.

	SUV$_{max}$ >2.5 (vol>2 mL)	SUV$_{peak}$ >Blood Pool (vol>2 mL)
Mean TMTV (min–max)	258.2 (0.0–2544.1)	235.2 (0.0–2471.9)
Median TMTV (IQR)	126.8 (37.8–295.0)	110.2 (32.9–280.8)
Average Dice with respect to the patient set of ROI$_{REF}$ (min-max)	0.59 (0.00–0.99)	0.60 (0.00–0.99)
Median Dice with respect to the patient set of ROI$_{REF}$ (IQR)	0.71 (0.31–0.86)	0.73 (0.33–0.86)
Average recall with respect to the patient set of ROI$_{REF}$ (min-max)	0.56 (0.00–1.00)	0.53 (0.00–0.99)
Median recall with respect to the patient set of ROI$_{REF}$ (IQR)	0.66 (0.23–0.85)	0.62 (0.20–0.81)
Average precision with respect to the patient set of ROI$_{REF}$ (min-max)	0.79 (0.00–1.00)	0.89 (0.00–1.00)
Median precision with respect to the patient set of ROI$_{REF}$ (IQR)	0.88 (0.72–0.96)	0.96 (0.86–0.99)
Spearman correlation coefficient with respect to reference TMTV	0.73	0.76

Tab. 4.7. Results associated with total metabolic tumor volume obtained using two different settings of the high-uptake region detection algorithm (multi-foci segmentation).

To our knowledge, this was the first analysis showing that an artificial intelligence method can generate a TMTV value prognostic of outcome in a large series of patients with DLBCL, with results comparable to other currently used methodologies. Other machine-learning–based approaches for TMTV estimation in lymphoma patients, including some involving CNN, are being developed and evaluated [65]. The automated method for TMTV segmentation assessed in the present analysis combined a region delineation method based on PERCIST recommendations and a deep-learning–based classification scheme for rapidly discarding physiologic uptake. Further efforts toward developing a stricter definition of TMTV, standardizing volume-segmentation methods, and establishing guidelines for the inclusion of tumor-bearing anatomic regions are ongoing, and these will constitute a prerequisite for the optimization of a complete automated method [3].

	SUV$_{max}$ >2.5 (vol>$2\,\mathrm{mL}$)	SUV$_{max}$ >2.5 (vol>$0.1\,\mathrm{mL}$)	SUV$_{peak}$ >Blood Pool (vol>$2\,\mathrm{mL}$)	SUV$_{peak}$ >Blood Pool (vol>$0.1\,\mathrm{mL}$)
Total number of MFS* findings (ROI$_{PARS}$)	18674	82114	6737	16717
Number of ROI$_{PARS}$ per subject, average (min-max)	66.7 (6-242)	293.3 (11-1952)†	24.1 (2-91)	59.7 (2-689)†
Number of ROI$_{PARS}$ per subject, median (IQR)	59.0 (39.0-86.0)	191.0 (91.8-428.5)†	19.0 (13.0-31.2)	39.5 (23.0-72.5)†
Classification accuracy per subject, average (min-max)	0.90 (0.40-1.00)	0.89 (0.46-1.00)‡	0.87 (0.34-1.00)	0.85 (0.38-1.00)†
Classification accuracy per subject, median (IQR)	0.93 (0.86-0.97)	0.93 (0.83-0.97)‡	0.89 (0.81-0.96)	0.87 (0.78-0.94)†
TMTV, average (min-max)	258.2 (0.0-2544.1)	275.9 (0.0-2571.9)†	235.2 (0.0-2471.9)	244.8 (0.0-2488.3)†
TMTV, median (IQR)	126.8 (37.8-295.0)	142.2 (42.9-340.1)†	110.2 (32.9-280.8)	123.3 (35.9-295.6)†
Dice with respect to the patient set of ROI$_{REF}$, average (min-max)	0.59 (0.00-0.99)	0.60 (0.00-0.99)†	0.60 (0.00-0.99)	0.62 (0.00-0.99)†
Dice with respect to the patient set of ROI$_{REF}$, median (IQR)	0.71 (0.31-0.86)	0.71 (0.35-0.85)†	0.73 (0.33-0.86)	0.74 (0.39-0.88)†

*Multi-foci segmentation

$^{\dagger}p < 0.05$,$^{\ddagger}p > 0.05$, Wilcoxon signed-rank test compared to the same variable obtained by neglecting ROIs with a volume below $2\,\mathrm{mL}$

Tab. 4.8. Results associated with the classification of high-uptake ROIs for four groups of ROIs obtained with two different settings of the multi-foci segmentation algorithm both with and without the neglection of ROIs with a volume between $0.1\,\mathrm{mL}$ and $2\,\mathrm{mL}$.

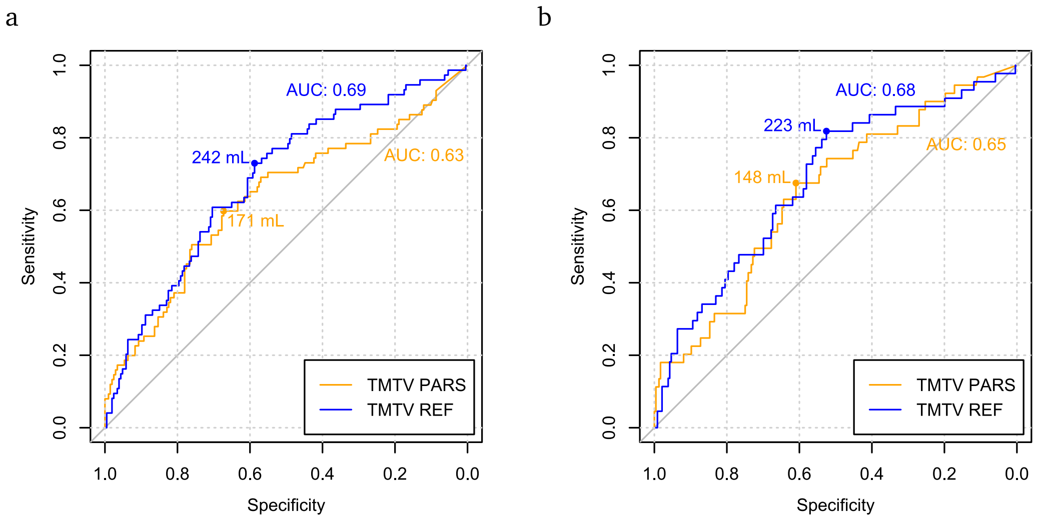


Fig. 4.3. ROC curves for determining the occurrence of PFS or OS events using a TMTV threshold. ROC curves for TMTV$_{PARS}$ and TMTV$_{REF}$ for (a) 4-y PFS and (b) 4-y OS. Areas under the ROC curves (AUC) and optimal TMTV cutoff thresholds are reported.

TMTV estimation	AUC*	Cutoff (mL)	Hazard ratio (95% CI)	High TMTV 4-y Survival	Low TMTV 4-y Survival	P
Progression-free Survival						
TMTV$_{PARS}$	0.61	110	2.4 (1.5–3.7)	58%	81%	0.00016
TMTV$_{REF}$	0.64	242	2.6 (1.6–4.1)	55%	83%	0.00004
Overall survival						
TMTV$_{PARS}$	0.64	148	2.8 (1.6–5.1)	74%	90%	0.00044
TMTV$_{REF}$	0.66	223	3.7 (1.9–7.2)	74%	93%	0.00012

*Area under the ROC curve

Tab. 4.9. Results associated with ROC analysis of TMTV, Kaplan–Meier estimation of four-year survival rates, Cox regression hazard ratio, and Wald test p-values for PFS and OS for the 280 subjects included in the analysis.

	Accuracy	Sensitivity	Specificity	NPV*	PPV†
TMTV$_{PARS}$ PFS	0.60	0.66	0.57	0.79	0.41
TMTV$_{REF}$ PFS	0.61	0.67	0.58	0.80	0.42
TMTV$_{PARS}$ OS	0.63	0.67	0.62	0.89	0.28
TMTV$_{REF}$ OS	0.58	0.78	0.53	0.92	0.27

*Negative predictive value
†Positive predictive value

Tab. 4.10. Performance of the prediction of the occurrence of an event for both PFS and OS based on the TMTV cutoff thresholds selected by maximizing Youden's J index.

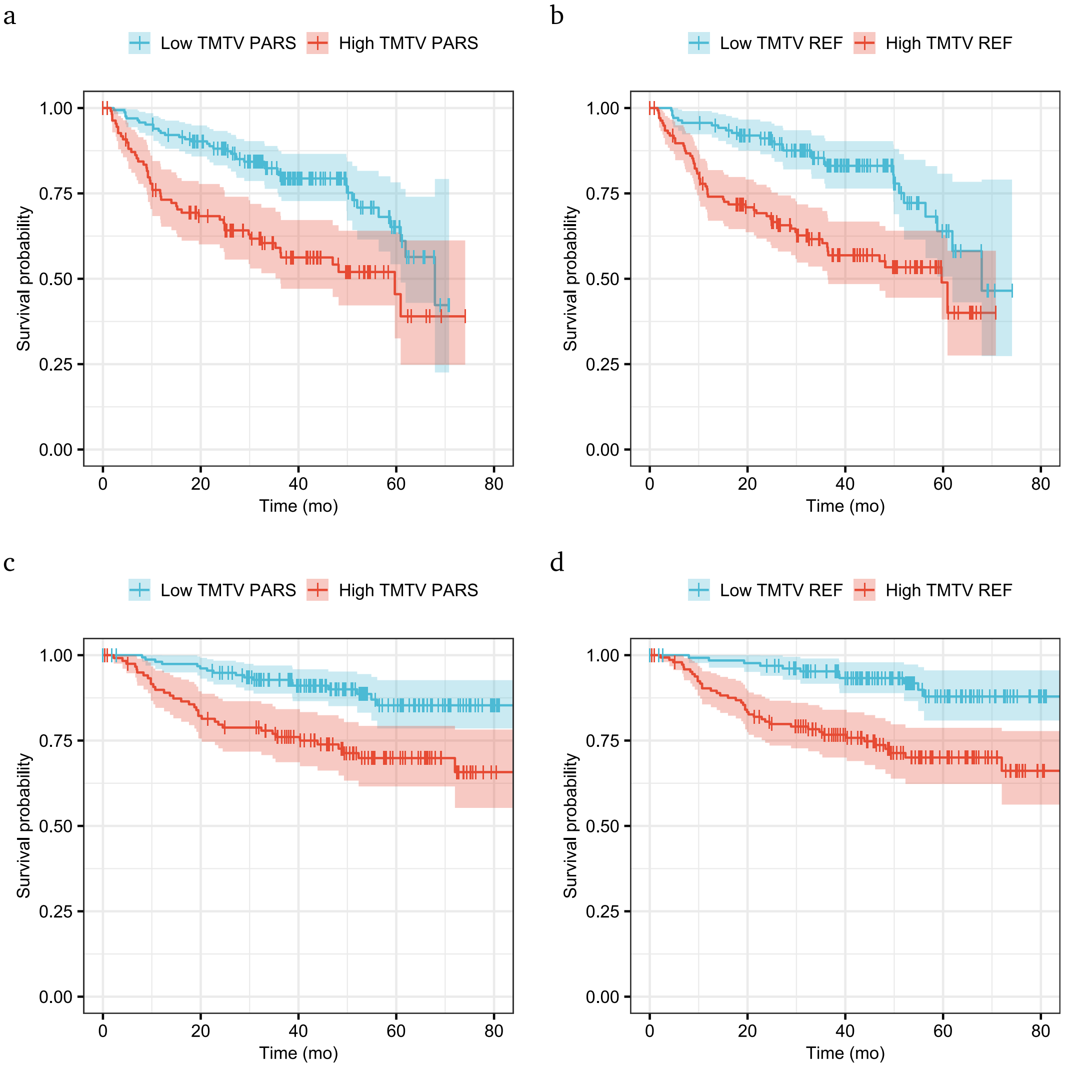


Fig. 4.4. Survival curves for the low- and high-TMTV groups for fully automated and reference TMTV estimations. Kaplan–Meier survival curves for PFS (a: TMTV_{PARS}, b: TMTV_{REF}) and OS (c: TMTV_{PARS}, d: TMTV_{REF}).

Part III

Conclusion

5 Conclusion and outlook

The research presented in this dissertation aimed at developing and evaluating methods to enable image-derived biomarkers in oncology. This chapter summarizes the main results obtained and discusses possible directions for future research.

In Chapter 2 a method for automated identification and anatomical location classification of sites suspicious for cancer in PET/CT images was described. The method was found to have good agreement with visual assessment by an expert physician. Considering the context of limited availability of image data with expert-annotated ground truth, training the described algorithm with information from both ^{68}Ga-PSMA-11 PET/CT and ^{18}F-FDG PET/CT scans was found to improve performance. Given the ability to automatically identify relevant imaging findings, the investigated techniques are promising to support physicians for efficient assessment of cancer spread and overall burden.

The analysis described considered as ground truth the visual image evaluation by a single physician. To fully validate the proposed method, the accuracy, reproducibility, and time-efficiency for a defined task with and without employing the algorithm should be compared for several users in a prospective analysis, using as ground truth the consensus from multiple experts, or ideally a composite standard of truth including follow-up or histology information. Moreover, dedicated surveys may in the future critically evaluate the proposed method, or investigate new solutions, for the analysis of challenging imaging findings, such as in presence of artifacts, tracer uptake pitfalls, or cases requiring contextual patient information beyond the PET/CT scan for correct interpretation. Notably, PET/CT image quality characteristics can vary significantly due to different acquisition systems, scan protocols, and image formation algorithms employed. Future analyses should evaluate the impact of different image quality characteristics on the performance of deep learning methods used to identify suspicious tracer uptake, and to which extent dedicated harmonization or algorithm training strategies can mitigate this impact.

In this dissertation, an uptake classification algorithm jointly trained with information from ^{68}Ga-PSMA-11 PET/CT and ^{18}F-FDG PET/CT scans was described. Currently, several radiotracers are in use and in development for clinical applications in oncology. Future analyses should further investigate the combination of training information from multiple radiotracers, to reduce the need for extensive sets of image data and expert-annotated ground truth for each tracer with potentially partly redundant information whose collection requires significant resources. With the same objective of maximizing the value of training information while minimizing the burden of its collection, approaches to acquire selected informative examples, such as active learning, or limit the need for expert-annotated ground truth, such as weakly supervised, semi-supervised or unsupervised learning, could be valuable for PET/CT image analysis. Moreover, given the frequently incumbent restrictions on data sharing, including the constraints imposed by data protection regulations, methods to train PET/CT image analysis algorithms across multiple centers without exchanging local training samples, such as federated learning, could be investigated, together with the impact of combining possibly heterogeneous training information originating from different centers.

The uptake classification algorithm described in the dissertation was trained with PET images expressed in SUV units, which have several limitations. Future analyses could investigate the use of parametric PET images for automated identification of suspicious uptake, combining multiple relevant kinetic parameters. Furthermore, while several recent analyses have evaluated deep learning techniques for identification of suspicious tracer uptake in PET/CT images, future analyses could investigate whether similar results can be obtained with other multimodal images such as SPECT/CT or PET/MRI, and whether training information for the same image analysis task may be efficiently shared across different hybrid images. While the method described in this dissertation focuses on a dichotomous classification of tracer uptake as nonsuspicious or suspicious, future analyses may consider multiple categories related to increased tracer uptake including for instance brown fat activation, muscle activation, inflammation, or infection, in a multiclass classification task.

The method described in this dissertation for automated identification of suspicious tracer uptake combined an initial segmentation of high uptake regions based on thresholding, having limited accuracy, with a deep learning-based classification algorithm to discard physiological uptake regions. Future analyses could evaluate the use of entirely deep learning-based segmentation algorithms for contouring regions suspicious for cancer in the whole body in PET/CT images, and compare their accuracy with commonly employed threshold-based methods, for different cancer types and radiotracers.

Notably, deep learning algorithms may be subject to bias and lack interpretability, posing a significant challenge for their use in guiding clinical decisions. Future investigations could further highlight limitations and sources of bias of deep learning algorithms for PET/CT image analysis; while methods developed in the active research field referred to as explainable artificial intelligence may aid interpretability, one example being the identification of relevant image features used for classification.

In Chapter 3 the use of the uptake classification method described in Chapter 2 to support prostate cancer staging based on PSMA-ligand PET/CT images was evaluated. The method was found to determine the extent of cancer spread with respect to local lymph nodes and distant metastases, according to the standardized PROMISE miTNM framework, in fair concordance with the visual assessment by a physician.

While the cohort analyzed included subjects referred to PSMA-ligand PET/CT for primary staging or assessment of biochemical recurrence, future analyses could evaluate the proposed method in a patient cohort representative of a broader spectrum of prostate cancer progression, including subjects with advanced disease with distant lesions other than in bone or lymph nodes, as well as subjects with diffuse bone marrow involvement. Future investigations may also evaluate the extent to which the proposed method could support users in determining prostate cancer stage, and whether further advances in the algorithm performance are required for this purpose. While the analysis described in the dissertation focused on PSMA-ligand tracer uptake suspicious for prostate cancer, future surveys could evaluate cases with PSMA-ligand tracer uptake related to other cancer types, and whether these cases can be identified through image analysis methods for appropriately supporting prostate cancer staging.

The analysis described in this thesis focused on staging with respect to local lymph nodes and distant metastases. In prostate cancer, the majority of subjects present disease localized to the prostate at diagnosis, and staging the primary tumor has a significant role for patient management, with diagnostic information increasingly being obtained through imaging. Future investigations could evaluate image analysis methods to support T staging, characterizing the primary tumor extent and invasion of adjacent anatomical structures based on hybrid imaging with PET/CT or PET/MRI. Besides the TNM classification, histological biomarkers such as the International Society of Urological Pathology (ISUP) Grade Group are highly relevant for risk stratification. Future analyses may further investigate whether PET/CT or PET/MR image-derived biomarkers might allow to reliably characterize the primary tumor grade noninvasively. Furthermore, while the present dissertation evaluated the use of the described uptake classification method for prostate cancer staging, future investigations may evaluate PET/CT image analysis methods for supporting the staging of other cancer types according to established staging criteria.

In Chapter 4 the use of a deep learning-based ^{18}F-FDG uptake classification algorithm for estimation of Total Metabolic Tumor Volume in patients with diffuse large B-cell lymphoma was described. The method was found to determine a fully automated estimation of baseline TMTV from PET/CT images which was significantly correlated and had comparable prognostic value with a TMTV estimation obtained semi-automatically by an expert physician. Classification of high-uptake regions using deep learning for rapidly discarding physiologic uptake may considerably simplify TMTV estimation and facilitate its use as prognostic indicator in patients with diffuse large B-cell lymphoma.

Future prospective analyses may further evaluate the accuracy, reproducibility, and time-efficiency of TMTV estimation by multiple users with and without employing the proposed algorithm. Since no consensus currently exists regarding the most appropriate method for TMTV segmentation, future investigations and consensus guidelines may identify a suitable recommended method allowing sufficient accuracy and reproducibility. In this context, entirely deep learning-based algorithms for segmentation of suspicious uptake regions are promising and may be further evaluated for TMTV segmentation in future analyses. The survey described in this dissertation equally evaluated PET/CT images acquired with different scanners models from multiple vendors, and future analyses may investigate the extent to which various image quality characteristics impact the performance of described method and the estimation of TMTV. While the analysis described in this thesis focused on estimation of TMTV based on ^{18}F-FDG PET/CT images in patients with diffuse large B-cell lymphoma, future investigations may further evaluate the proposed method for TMTV estimation in other lymphoma subtypes, other cancer types, or based on PET/CT images acquired with other radiotracers.

For supporting the use of TMTV in clinical settings, future analyses may evaluate the integration of TMTV with well-established disease indicators, propose viable related criteria to guide decisions in defined clinical applications, such as risk stratification or response assessment, and evaluate the impact of applying the proposed criteria for patient management.

The image analysis methods described in this dissertation allow to derive relevant findings for determining meaningful image-derived biomarkers, including cancer stage and overall burden. Moreover, by identification and anatomical location classification of regions suspicious for cancer in whole-body images, the described methods allow to extract extended information, which may support the investigation of novel image-derived biomarkers to characterize disease in oncology. In conclusion, image analysis methods based on deep learning may aid physicians in evaluating informative and actionable image-derived cancer biomarkers and, together with diligent validation in multicenter trials, drive their establishment in the clinical routine.

Part IV

Appendix

A List of Authored and Co-authored Publications

Primary author

Journal articles

N. Capobianco, M. Meignan, A.-S. Cottereau, L. Vercellino, L. Sibille, B. Spottiswoode, S. Zuehlsdorff, O. Casasnovas, C. Thieblemont, and I. Buvat. "Deep-Learning ^{18} F-FDG Uptake Classification Enables Total Metabolic Tumor Volume Estimation in Diffuse Large B-Cell Lymphoma". en. In: *Journal of Nuclear Medicine* 62.1 (Jan. 2021), pp. 30–36

N. Capobianco, L. Sibille, M. Chantadisai, A. Gafita, T. Langbein, G. Platsch, E. L. Solari, V. Shah, B. Spottiswoode, M. Eiber, W. A. Weber, N. Navab, and S. G. Nekolla. "Whole-Body Uptake Classification and Prostate Cancer Staging in 68Ga-PSMA-11 PET/CT Using Dual-Tracer Learning". en. In: *European Journal of Nuclear Medicine and Molecular Imaging* (July 2021)

Conference abstracts

N. Capobianco, A. Gafita, G. Platsch, L. Sibille, B. Spottiswoode, M. Eiber, W. Weber, N. Navab, and S. Nekolla. "Transfer Learning of AI-Based Uptake Classification from ^{18}F-FDG PET/CT to ^{68}Ga-PSMA-11 PET/CT for Whole-Body Tumor Burden Assessment". In: *Journal of Nuclear Medicine* 61.supplement 1 (May 2020), p. 1411

N. Capobianco, M. Meignan, A. S. Cottereau, L. Vercellino, L. Sibille, B. Spottiswoode, S. Zuehlsdorff, O. Casasnovas, C. Thieblemont, and I. Buvat. "Fully Automated Deep Learning FDG Uptake Classification Enables Total Metabolic Tumor Volume (MTV) Estimation in Diffuse Large B-Cell Lymphoma with Similar Predictive Value as Expert MTV Measurements". In: *Journal of Nuclear Medicine* 61.supplement 1 (May 2020), p. 504

N. Capobianco, A. Gafita, G. Platsch, L. Sibille, B. Spottiswoode, M. Eiber, W. A. Weber, N. Navab, and S. G. Nekolla. "Whole-Body Lesion Detection and Prostate Cancer Staging in 68Ga-PSMA-11 PET/CT Using Deep Learning". en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 47.S1 (Sept. 2020), pp. 273–274

Co-author

Journal articles

A.-S. Cottereau, M. Meignan, C. Nioche, N. Capobianco, J. Clerc, L. Chartier, L. Vercellino, O. Casasnovas, C. Thieblemont, and I. Buvat. "Risk Stratification in Diffuse Large B-Cell Lymphoma Using Lesion Dissemination and Metabolic Tumor Burden Calculated from Baseline PET/CT†". en. In: *Annals of Oncology* 32.3 (Mar. 2021), pp. 404–411

Conference abstracts

F. Orlhac, N. Capobianco, A.-S. Cottereau, L. Vercellino, S. Zuehlsdorff, O. Casasnovas, C. Thieblemont, M. Meignan, and I. Buvat. "Refining the Stratification of Diffuse Large B-Cell Lymphoma Patients Based on Metabolic Tumor Volume (MTV) by Automatically Adapting the MTV Cut-off Value to the Segmentation Method". In: *Journal of Nuclear Medicine* 61.supplement 1 (2020), pp. 274–274

Bibliography

[1] M. B. Amin, American Joint Committee on Cancer, and A. C. Society, eds. *AJCC Cancer Staging Manual*. Eight edition / editor-in-chief, Mahul B. Amin, MD, FCAP ; editors, Stephen B. Edge, MD, FACS [and 16 others] ; Donna M. Gress, RHIT, CTR - Technical editor ; Laura R. Meyer, CAPM - Managing editor. Chicago IL: American Joint Committee on Cancer, Springer, 2017 (cit. on p. 15).

[2] D. L. Bailey, D. W. Townsend, P. E. Valk, and M. N. Maisey. *Positron Emission Tomography Basic Sciences*. English. Springer, London, 2005 (cit. on pp. 10–12).

[3] S. F. Barrington and M. Meignan. "Time to Prepare for Risk Adaptation in Lymphoma by Standardizing Measurement of Metabolic Tumor Burden". en. In: *Journal of Nuclear Medicine* 60.8 (Aug. 2019), pp. 1096–1102 (cit. on pp. 37, 50, 60).

[4] S. Bastawrous, P. Bhargava, F. Behnia, D. S. W. Djang, and D. R. Haseley. "Newer PET Application with an Old Tracer: Role of ^{18} F-NaF Skeletal PET/CT in Oncologic Practice". en. In: *RadioGraphics* 34.5 (Sept. 2014), pp. 1295–1316 (cit. on p. 14).

[5] M. Bieth, M. Krönke, R. Tauber, et al. "Exploring New Multimodal Quantitative Imaging Indices for the Assessment of Osseous Tumor Burden in Prostate Cancer Using ^{68} Ga-PSMA PET/CT". en. In: *Journal of Nuclear Medicine* 58.10 (Oct. 2017), pp. 1632–1637 (cit. on pp. 18, 28).

[6] P. Blanc-Durand, S. Jégou, S. Kanoun, et al. "Fully Automatic Segmentation of Diffuse Large B Cell Lymphoma Lesions on 3D FDG-PET/CT for Total Metabolic Tumour Volume Prediction Using a Convolutional Neural Network." en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 48.5 (May 2021), pp. 1362–1370 (cit. on p. 37).

[7] R. Boellaard, R. Delgado-Bolton, W. J. G. Oyen, et al. "FDG PET/CT: EANM Procedure Guidelines for Tumour Imaging: Version 2.0". en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 42.2 (Feb. 2015), pp. 328–354 (cit. on pp. 16, 37).

[8] J. Brierley, M. K. Gospodarowicz, and C. Wittekind, eds. *TNM Classification of Malignant Tumours*. Eighth edition. Chichester, West Sussex, UK ; Hoboken, NJ: John Wiley & Sons, Inc, 2017 (cit. on p. 15).

[9] M. K. Buyyounouski, P. L. Choyke, J. K. McKenney, et al. "Prostate Cancer - Major Changes in the American Joint Committee on Cancer Eighth Edition Cancer Staging Manual: Prostate Cancer-Major 8th Edition Changes". en. In: *CA: A Cancer Journal for Clinicians* 67.3 (May 2017), pp. 245–253 (cit. on pp. 42, 46).

[10] T. M. Buzug. *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*. Berlin: Springer, 2008 (cit. on pp. 5, 7).

[11] M. R. Camacho, E. Etchebehere, N. Tardelli, et al. "Validation of a Multifocal Segmentation Method for Measuring Metabolic Tumor Volume in Hodgkin Lymphoma". en. In: *Journal of Nuclear Medicine Technology* 48.1 (Mar. 2020), pp. 30–35 (cit. on pp. 29, 30, 43).

[12] N. Capobianco, A. Gafita, G. Platsch, et al. “Transfer Learning of AI-Based Uptake Classification from ^{18}F-FDG PET/CT to ^{68}Ga-PSMA-11 PET/CT for Whole-Body Tumor Burden Assessment”. In: *Journal of Nuclear Medicine* 61.supplement 1 (May 2020), p. 1411 (cit. on pp. 27, 73).

[13] N. Capobianco, A. Gafita, G. Platsch, et al. “Whole-Body Lesion Detection and Prostate Cancer Staging in 68Ga-PSMA-11 PET/CT Using Deep Learning”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 47.S1 (Sept. 2020), pp. 273–274 (cit. on pp. 41, 73).

[14] N. Capobianco, M. Meignan, A.-S. Cottereau, et al. “Deep-Learning ^{18} F-FDG Uptake Classification Enables Total Metabolic Tumor Volume Estimation in Diffuse Large B-Cell Lymphoma”. en. In: *Journal of Nuclear Medicine* 62.1 (Jan. 2021), pp. 30–36 (cit. on pp. 49, 73).

[15] N. Capobianco, M. Meignan, A. S. Cottereau, et al. “Fully Automated Deep Learning FDG Uptake Classification Enables Total Metabolic Tumor Volume (MTV) Estimation in Diffuse Large B-Cell Lymphoma with Similar Predictive Value as Expert MTV Measurements”. In: *Journal of Nuclear Medicine* 61.supplement 1 (May 2020), p. 504 (cit. on pp. 49, 73).

[16] N. Capobianco, L. Sibille, M. Chantadisai, et al. “Whole-Body Uptake Classification and Prostate Cancer Staging in 68Ga-PSMA-11 PET/CT Using Dual-Tracer Learning”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* (July 2021) (cit. on pp. 27, 41, 73).

[17] P. P. Carbone, H. S. Kaplan, K. Musshoff, D. W. Smithers, and M. Tubiana. “Report of the Committee on Hodgkin’s Disease Staging Classification”. eng. In: *Cancer Research* 31.11 (Nov. 1971), pp. 1860–1861 (cit. on p. 15).

[18] F. Ceci, D. E. Oprea-Lager, L. Emmett, et al. “E-PSMA: The EANM Standardized Reporting Guidelines v1.0 for PSMA-PET”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* (Feb. 2021) (cit. on pp. 28, 47).

[19] C. A. Chang, D. A. Pattison, R. W. Tothill, et al. “68Ga-DOTATATE and 18F-FDG PET/CT in Paraganglioma and Pheochromocytoma: Utility, Patterns and Heterogeneity”. en. In: *Cancer Imaging* 16.1 (Dec. 2016), p. 22 (cit. on p. 13).

[20] B. D. Cheson, R. I. Fisher, S. F. Barrington, et al. “Recommendations for Initial Evaluation, Staging, and Response Assessment of Hodgkin and Non-Hodgkin Lymphoma: The Lugano Classification”. en. In: *Journal of Clinical Oncology* 32.27 (Sept. 2014), pp. 3059–3067 (cit. on p. 15).

[21] A. Cottereau, L. Vercellino, O. Casasnovas, et al. “High Total Metabolic Tumor Volume at Baseline Allows to Discriminate for Survival Patients in Response after R-Chop: An Ancillary Analysis of the Remarc Study”. en. In: *Hematological Oncology* 37 (June 2019), pp. 49–50 (cit. on pp. 50, 51).

[22] A.-S. Cottereau, M. Meignan, C. Nioche, et al. “Risk Stratification in Diffuse Large B-Cell Lymphoma Using Lesion Dissemination and Metabolic Tumor Burden Calculated from Baseline PET/CT†”. en. In: *Annals of Oncology* 32.3 (Mar. 2021), pp. 404–411 (cit. on p. 74).

[23] A.-S. Cottereau, I. Buvat, S. Kanoun, et al. “Is There an Optimal Method for Measuring Baseline Metabolic Tumor Volume in Diffuse Large B Cell Lymphoma?” en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 45.8 (July 2018), pp. 1463–1464 (cit. on p. 59).

[24] A.-S. Cottereau, S. Hapdey, L. Chartier, et al. “Baseline Total Metabolic Tumor Volume Measured with Fixed or Different Adaptive Thresholding Methods Equally Predicts Outcome in Peripheral T Cell Lymphoma”. en. In: *Journal of Nuclear Medicine* 58.2 (Feb. 2017), pp. 276–281 (cit. on pp. 17, 50, 58).

[25] A.-S. Cottereau, H. Lanic, S. Mareschal, et al. “Molecular Profile and FDG-PET/CT Total Metabolic Tumor Volume Improve Risk Classification at Diagnosis for Patients with Diffuse Large B-Cell Lymphoma”. en. In: *Clinical Cancer Research* 22.15 (Aug. 2016), pp. 3801–3809 (cit. on p. 17).

[26] A.-S. Cottereau, A. Versari, A. Loft, et al. “Prognostic Value of Baseline Metabolic Tumor Volume in Early-Stage Hodgkin Lymphoma in the Standard Arm of the H10 Trial”. en. In: *Blood* 131.13 (Mar. 2018), pp. 1456–1463 (cit. on p. 51).

[27] A. Cottereau, S. Becker, F. Broussais, et al. “Prognostic Value of Baseline Total Metabolic Tumor Volume (TMTV0) Measured on FDG-PET/CT in Patients with Peripheral T-Cell Lymphoma (PTCL)”. en. In: *Annals of Oncology* 27.4 (Apr. 2016), pp. 719–724 (cit. on p. 49).

[28] M. Crump, S. S. Neelapu, U. Farooq, et al. “Outcomes in Refractory Diffuse Large B-Cell Lymphoma: Results from the International SCHOLAR-1 Study”. en. In: *Blood* 130.16 (Oct. 2017), pp. 1800–1808 (cit. on p. 49).

[29] J. Czernin, N. Satyamurthy, and C. Schiepers. “Molecular Mechanisms of Bone 18F-NaF Deposition”. en. In: *Journal of Nuclear Medicine* 51.12 (Dec. 2010), pp. 1826–1829 (cit. on p. 14).

[30] N. Dalal and B. Triggs. “Histograms of Oriented Gradients for Human Detection”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*. Vol. 1. San Diego, CA, USA: IEEE, 2005, pp. 886–893 (cit. on p. 19).

[31] U. Dührsen, S. Müller, B. Hertenstein, et al. “Positron Emission Tomography–Guided Therapy of Aggressive Non-Hodgkin Lymphomas (PETAL): A Multicenter, Randomized Phase III Trial”. en. In: *Journal of Clinical Oncology* 36.20 (July 2018), pp. 2024–2034 (cit. on p. 16).

[32] M. Eiber, T. Maurer, M. Souvatzoglou, et al. “Evaluation of Hybrid 68Ga-PSMA Ligand PET/CT in 248 Patients with Biochemical Recurrence After Radical Prostatectomy”. en. In: *Journal of Nuclear Medicine* 56.5 (May 2015), pp. 668–674 (cit. on p. 27).

[33] M. Eiber, W. P. Fendler, S. P. Rowe, et al. “Prostate-Specific Membrane Antigen Ligands for Imaging and Therapy”. en. In: *Journal of Nuclear Medicine* 58.Supplement 2 (Sept. 2017), 67S–76S (cit. on p. 13).

[34] M. Eiber, K. Herrmann, J. Calais, et al. “Prostate Cancer Molecular Imaging Standardized Evaluation (PROMISE): Proposed miTNM Classification for the Interpretation of PSMA-Ligand PET/CT”. en. In: *Journal of Nuclear Medicine* 59.3 (Mar. 2018), pp. 469–478 (cit. on pp. 28, 42).

[35] M. Eiber, S. G. Nekolla, T. Maurer, G. Weirich, H.-J. Wester, and M. Schwaiger. “68Ga-PSMA PET/MR with Multimodality Image Analysis for Primary Prostate Cancer”. en. In: *Abdominal Imaging* 40.6 (Aug. 2015), pp. 1769–1771 (cit. on p. 28).

[36] S. Fanti, S. Minozzi, J. J. Morigi, et al. “Development of Standardized Image Interpretation for 68Ga-PSMA PET/CT to Detect Prostate Cancer Recurrent Lesions”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 44.10 (Sept. 2017), pp. 1622–1635 (cit. on p. 47).

[37] P. F. Felzenszwalb, R. B. Girshick, D McAllester, and D Ramanan. “Object Detection with Discriminatively Trained Part-Based Models”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9 (Sept. 2010), pp. 1627–1645 (cit. on p. 19).

[38] W. P. Fendler, M. Eiber, M. Beheshti, et al. “68Ga-PSMA PET/CT: Joint EANM and SNMMI Procedure Guideline for Prostate Cancer Imaging: Version 1.0”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 44.6 (June 2017), pp. 1014–1024 (cit. on pp. 13, 14, 28).

[39] W. P. Fendler, J. Calais, M. Allen-Auerbach, et al. “^{68} Ga-PSMA-11 PET/CT Interobserver Agreement for Prostate Cancer Assessments: An International Multicenter Prospective Study”. en. In: *Journal of Nuclear Medicine* 58.10 (Oct. 2017), pp. 1617–1623 (cit. on p. 47).

[40] D. A. Ferraro, A. S. Becker, B. Kranzbühler, et al. “Diagnostic Performance of 68Ga-PSMA-11 PET/MRI-Guided Biopsy in Patients with Suspected Prostate Cancer: A Prospective Single-Center Study”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* (Feb. 2021) (cit. on p. 46).

[41] A. Fourquet, C. Aveline, O. Cussenot, et al. “68Ga-PSMA-11 PET/CT in Restaging Castration-Resistant Nonmetastatic Prostate Cancer: Detection Rate, Impact on Patients’ Disease Management and Adequacy of Impact”. en. In: *Scientific Reports* 10.1 (Dec. 2020), p. 2104 (cit. on p. 13).

[42] A. Gafita, M. Bieth, M. Krönke, et al. “qPSMA: Semiautomatic Software for Whole-Body Tumor Burden Assessment in Prostate Cancer Using ^{68} Ga-PSMA11 PET/CT”. en. In: *Journal of Nuclear Medicine* 60.9 (Sept. 2019), pp. 1277–1283 (cit. on pp. 18, 28, 37).

[43] F. Giammarile, P. Castellucci, R. Dierckx, et al. “Non-FDG PET/CT in Diagnostic Oncology: A Pictorial Review”. en. In: *European Journal of Hybrid Imaging* 3.1 (Dec. 2019), p. 20 (cit. on p. 13).

[44] F. L. Giesel, C. Kesch, M. Yun, et al. “18F-PSMA-1007 PET/CT Detects Micrometastases in a Patient With Biochemically Recurrent Prostate Cancer”. en. In: *Clinical Genitourinary Cancer* 15.3 (June 2017), e497–e499 (cit. on p. 28).

[45] F. L. Giesel, K. Knorr, F. Spohn, et al. “Detection Efficacy of ^{18} F-PSMA-1007 PET/CT in 251 Patients with Biochemical Recurrence of Prostate Cancer After Radical Prostatectomy”. en. In: *Journal of Nuclear Medicine* 60.3 (Mar. 2019), pp. 362–368 (cit. on p. 13).

[46] C. Gisselbrecht, B. Glass, N. Mounier, et al. “Salvage Regimens With Autologous Transplantation for Relapsed Large B-Cell Lymphoma in the Rituximab Era”. en. In: *Journal of Clinical Oncology* 28.27 (Sept. 2010), pp. 4184–4190 (cit. on p. 49).

[47] P. Goldstraw, K. Chansky, J. Crowley, et al. “The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer”. en. In: *Journal of Thoracic Oncology* 11.1 (Jan. 2016), pp. 39–51 (cit. on p. 15).

[48] E. Grossiord, H. Talbot, N. Passat, M. Meignan, P. Terve, and L. Najman. “Hierarchies and Shape-Space for Pet Image Segmentation”. In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. Brooklyn, NY, USA: IEEE, Apr. 2015, pp. 1118–1121 (cit. on p. 51).

[49] F.-N. B. W. Group. “BEST (Biomarkers, Endpoints, and Other Tools) Resource [Internet]”. In: (2016) (cit. on p. 12).

[50] B. Grubmüller, P. Baltzer, S. Hartenbach, et al. “PSMA Ligand PET/MRI for Primary Prostate Cancer: Staging Performance and Clinical Impact”. en. In: *Clinical Cancer Research* 24.24 (Dec. 2018), pp. 6300–6307 (cit. on p. 46).

[51] B. Grubmüller, D. Senn, G. Kramer, et al. “Response Assessment Using 68Ga-PSMA Ligand PET in Patients Undergoing 177Lu-PSMA Radioligand Therapy for Metastatic Castration-Resistant Prostate Cancer”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 46.5 (May 2019), pp. 1063–1072 (cit. on pp. 17, 28).

[52] J. Hammes, P. Täger, and A. Drzezga. “EBONI: A Tool for Automated Quantification of Bone Metastasis Load in PSMA PET/CT”. eng. In: *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine* 59.7 (July 2018), pp. 1070–1075 (cit. on p. 28).

[53] S. A. Harmon, E. Bergvall, E. Mena, et al. “A Prospective Comparison of ^{18} F-Sodium Fluoride PET/CT and PSMA-Targeted ^{18} F-DCFBC PET/CT in Metastatic Prostate Cancer”. en. In: *Journal of Nuclear Medicine* 59.11 (Nov. 2018), pp. 1665–1671 (cit. on p. 13).

[54] A. Hartenstein, F. Lübbe, A. D. J. Baur, et al. “Prostate Cancer Nodal Staging: Using Deep Learning to Predict 68Ga-PSMA-Positivity from CT Imaging Alone”. en. In: *Scientific Reports* 10.1 (Dec. 2020), p. 3398 (cit. on p. 28).

[55] P. E. Hartrampf, M. Heinrich, A. K. Seitz, et al. “Metabolic Tumour Volume from PSMA PET/CT Scans of Prostate Cancer Patients during Chemotherapy—Do Different Software Solutions Deliver Comparable Results?” en. In: *Journal of Clinical Medicine* 9.5 (May 2020), p. 1390 (cit. on p. 38).

[56] M. Hatt, B. Laurent, A. Ouahabi, et al. “The First MICCAI Challenge on PET Tumor Segmentation”. en. In: *Medical Image Analysis* 44 (Feb. 2018), pp. 177–195 (cit. on p. 37).

[57] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 770–778 (cit. on p. 20).

[58] K. He, X. Zhang, S. Ren, and J. Sun. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, Dec. 2015, pp. 1026–1034 (cit. on p. 20).

[59] M. S. Hofman, R. J. Hicks, T. Maurer, and M. Eiber. “Prostate-Specific Membrane Antigen PET: Clinical Utility in Prostate Cancer, Normal Patterns, Pearls, and Pitfalls”. en. In: *RadioGraphics* 38.1 (Jan. 2018), pp. 200–217 (cit. on pp. 14, 28).

[60] M. S. Hofman, N. Lawrentschuk, R. J. Francis, et al. “Prostate-Specific Membrane Antigen PET-CT in Patients with High-Risk Prostate Cancer before Curative-Intent Surgery or Radiotherapy (proPSMA): A Prospective, Randomised, Multicentre Study”. en. In: *The Lancet* 395.10231 (Apr. 2020), pp. 1208–1216 (cit. on p. 27).

[61] S. H. Hyun, J. Y. Choi, Y. M. Shim, et al. “Prognostic Value of Metabolic Tumor Volume Measured by 18F-Fluorodeoxyglucose Positron Emission Tomography in Patients with Esophageal Carcinoma”. en. In: *Annals of Surgical Oncology* 17.1 (Jan. 2010), pp. 115–122 (cit. on p. 17).

[62] H. Ilyas, N. G. Mikhaeel, J. T. Dunn, et al. “Defining the Optimal Method for Measuring Baseline Metabolic Tumour Volume in Diffuse Large B Cell Lymphoma”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 45.7 (July 2018), pp. 1142–1154 (cit. on pp. 17, 50, 58).

[63] International Non-Hodgkin’s Lymphoma Prognostic Factors Project. “A Predictive Model for Aggressive Non-Hodgkin’s Lymphoma”. en. In: *New England Journal of Medicine* 329.14 (Sept. 1993), pp. 987–994 (cit. on p. 50).

[64] S. Jemaa, J. Fredrickson, R. A. D. Carano, T. Nielsen, A. de Crespigny, and T. Bengtsson. “Tumor Segmentation and Feature Extraction from Whole-Body FDG-PET/CT Using Cascaded 2D and 3D Convolutional Neural Networks”. en. In: *Journal of Digital Imaging* 33.4 (Aug. 2020), pp. 888–894 (cit. on p. 37).

[65] S. Jemaa, J. Fredrickson, A. Coimbra, et al. “A Fully Automated Measurement of Total Metabolic Tumor Burden in Diffuse Large B-Cell Lymphoma and Follicular Lymphoma”. en. In: *Blood* 134.Supplement_1 (Nov. 2019), pp. 4666–4666 (cit. on p. 60).

[66] C. B. Johnbeck, U. Knigge, and A. Kjær. “PET Tracers for Somatostatin Receptor Imaging of Neuroendocrine Tumors: Current Status and Review of the Literature”. en. In: *Future Oncology* 10.14 (Nov. 2014), pp. 2259–2277 (cit. on p. 14).

[67] W. Kalender. *Computed Tomography: Fundamentals, System Technology, Image Quality, Applications*. English. Weinheim: Wiley-VCH, 2011 (cit. on p. 7).

[68] S. Kanoun, C. Rossi, A. Berriolo-Riedinger, et al. “Baseline Metabolic Tumour Volume Is an Independent Prognostic Factor in Hodgkin Lymphoma”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 41.9 (Sept. 2014), pp. 1735–1743 (cit. on p. 49).

[69] S. Kanoun, I. Tal, A. Berriolo-Riedinger, et al. “Influence of Software Tool and Methodological Aspects of Total Metabolic Tumor Volume Calculation on Baseline [18F]FDG PET to Predict Survival in Hodgkin Lymphoma”. en. In: *PLOS ONE* 10.10 (Oct. 2015). Ed. by C.-T. Chen, e0140830 (cit. on p. 51).

[70] J. W. Keyes. “SUV: Standard Uptake or Silly Useless Value?” eng. In: *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine* 36.10 (Oct. 1995), pp. 1836–1839 (cit. on p. 16).

[71] P. E. Kinahan, D. W. Townsend, T. Beyer, and D. Sashin. “Attenuation Correction for a Combined 3D PET/CT Scanner”. en. In: *Medical Physics* 25.10 (Oct. 1998), pp. 2046–2053 (cit. on p. 11).

[72] D. Kostyszyn, T. Fechter, N. Bartl, et al. “Intraprostatic Tumor Segmentation on PSMA PET Images in Patients with Primary Prostate Cancer with a Convolutional Neural Network”. en. In: *Journal of Nuclear Medicine* 62.6 (June 2021), pp. 823–828 (cit. on p. 46).

[73] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. en. In: *Communications of the ACM* 60.6 (May 2017), pp. 84–90 (cit. on p. 20).

[74] H. Labriet, C. Nemoz, M. Renier, et al. “Significant Dose Reduction Using Synchrotron Radiation Computed Tomography: First Clinical Case and Application to High Resolution CT Exams”. en. In: *Scientific Reports* 8.1 (Dec. 2018), p. 12491 (cit. on p. 6).

[75] J. Langner. “Development of a Parallel Computing Optimized Head Movement Correction Method in Positron Emission Tomography”. In: *Master of computer science thesis, university of Applied sciences Dresden and research center Dresden-rossendorf* (2003) (cit. on p. 8).

[76] Y. LeCun, B. Boser, J. S. Denker, et al. “Backpropagation Applied to Handwritten Zip Code Recognition”. en. In: *Neural Computation* 1.4 (Dec. 1989), pp. 541–551 (cit. on p. 19).

[77] C. Liu, T. Liu, Z. Zhang, et al. “^{68} Ga-PSMA PET/CT Combined with PET/Ultrasound-Guided Prostate Biopsy Can Diagnose Clinically Significant Prostate Cancer in Men with Previous Negative Biopsy Results”. en. In: *Journal of Nuclear Medicine* 61.9 (Sept. 2020), pp. 1314–1319 (cit. on p. 46).

[78] D. Lowe. “Object Recognition from Local Scale-Invariant Features”. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Kerkyra, Greece: IEEE, 1999, 1150–1157 vol.2 (cit. on p. 19).

[79] B. R. Mason, J. A. Eastham, B. J. Davis, et al. “Current Status of MRI and PET in the NCCN Guidelines for Prostate Cancer”. In: *Journal of the National Comprehensive Cancer Network* 17.5 (May 2019), pp. 506–513 (cit. on p. 27).

[80] C. D. Mathers, T. Boerma, and D. Ma Fat. “Global and Regional Causes of Death”. en. In: *British Medical Bulletin* 92.1 (Dec. 2009), pp. 7–32 (cit. on p. 3).

[81] T. Maurer, J. E. Gschwend, I. Rauscher, et al. “Diagnostic Efficacy of ^{68} Gallium-PSMA Positron Emission Tomography Compared to Conventional Imaging for Lymph Node Staging of 130 Consecutive Patients with Intermediate to High Risk Prostate Cancer”. en. In: *Journal of Urology* 195.5 (May 2016), pp. 1436–1443 (cit. on p. 27).

[82] M. Meignan, A. S. Cottereau, A. Versari, et al. “Baseline Metabolic Tumor Volume Predicts Outcome in High-Tumor-Burden Follicular Lymphoma: A Pooled Analysis of Three Multicenter Studies”. eng. In: *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 34.30 (Oct. 2016), pp. 3618–3626 (cit. on pp. 17, 49, 51).

[83] M. Meignan, A. Gallamini, M. Meignan, A. Gallamini, and C. Haioun. “Report on the First International Workshop on Interim-PET Scan in Lymphoma”. en. In: *Leukemia & Lymphoma* 50.8 (Jan. 2009), pp. 1257–1260 (cit. on p. 17).

[84] M. Meignan, M. Sasanelli, R. O. Casasnovas, et al. “Metabolic Tumour Volumes Measured at Staging in Lymphoma: Methodological Evaluation on Phantom Experiments and Patients”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 41.6 (June 2014), pp. 1113–1122 (cit. on p. 51).

[85] J. L. Mohler, E. S. Antonarakis, A. J. Armstrong, et al. “Prostate Cancer, Version 2.2019, NCCN Clinical Practice Guidelines in Oncology”. In: *Journal of the National Comprehensive Cancer Network* 17.5 (May 2019), pp. 479–505 (cit. on p. 27).

[86] F. Orlhac, N. Capobianco, A.-S. Cottereau, et al. “Refining the Stratification of Diffuse Large B-Cell Lymphoma Patients Based on Metabolic Tumor Volume (MTV) by Automatically Adapting the MTV Cut-off Value to the Segmentation Method”. In: *Journal of Nuclear Medicine* 61.supplement 1 (2020), pp. 274–274 (cit. on p. 74).

[87] K. Pak, G. J. Cheon, H.-Y. Nam, et al. “Prognostic Value of Metabolic Tumor Volume and Total Lesion Glycolysis in Head and Neck Cancer: A Systematic Review and Meta-Analysis”. en. In: *Journal of Nuclear Medicine* 55.6 (June 2014), pp. 884–890 (cit. on p. 17).

[88] S. J. Pan and Q. Yang. “A Survey on Transfer Learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (Oct. 2010), pp. 1345–1359 (cit. on p. 20).

[89] S. Pecorelli, J. Benedet, W. Creasman, J. Shepherd, and on behalf of the 1994-1997 FIGO Committee on Gynecologic Oncology. “FIGO Staging of Gynecologic Cancer”. en. In: *International Journal of Gynecology & Obstetrics* 65.3 (June 1999), pp. 243–249 (cit. on p. 15).

[90] E. Pfaehler, L. Mesotten, G. Kramer, et al. “Repeatability of Two Semi-Automatic Artificial Intelligence Approaches for Tumor Segmentation in PET”. en. In: *EJNMMI Research* 11.1 (Dec. 2021), p. 4 (cit. on p. 37).

[91] C. Plathow and W. A. Weber. “Tumor Cell Metabolism Imaging”. en. In: *Journal of Nuclear Medicine* 49.Suppl_2 (June 2008), 43S–63S (cit. on p. 13).

[92] T. Poeppel, B. Krause, T. Heusner, C. Boy, A. Bockisch, and G. Antoch. “PET/CT for the Staging and Follow-up of Patients with Malignancies”. en. In: *European Journal of Radiology* 70.3 (June 2009), pp. 382–392 (cit. on p. 15).

[93] J. Radon. “On the Determination of Functions from Their Integral Values along Certain Manifolds”. In: *IEEE Transactions on Medical Imaging* 5.4 (Dec. 1986), pp. 170–176 (cit. on p. 7).

[94] A. Rahmim, M. A. Lodge, N. A. Karakatsanis, et al. “Dynamic Whole-Body PET Imaging: Principles, Potentials and Applications”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 46.2 (Feb. 2019), pp. 501–518 (cit. on p. 11).

[95] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. “Self-Taught Learning: Transfer Learning from Unlabeled Data”. en. In: *Proceedings of the 24th International Conference on Machine Learning - ICML ’07*. Corvalis, Oregon: ACM Press, 2007, pp. 759–766 (cit. on p. 20).

[96] I. Rauscher, T. Maurer, A. J. Beer, et al. “Value of 68Ga-PSMA HBED-CC PET for the Assessment of Lymph Node Metastases in Prostate Cancer Patients with Biochemical Recurrence: Comparison with Histopathology After Salvage Lymphadenectomy”. en. In: *Journal of Nuclear Medicine* 57.11 (Nov. 2016), pp. 1713–1719 (cit. on p. 27).

[97] I. Rauscher, M. Krönke, M. König, et al. “Matched-Pair Comparison of ^{68} Ga-PSMA-11 PET/CT and ^{18} F-PSMA-1007 PET/CT: Frequency of Pitfalls and Detection Efficacy in Biochemical Recurrence After Radical Prostatectomy”. en. In: *Journal of Nuclear Medicine* 61.1 (Jan. 2020), pp. 51–57 (cit. on p. 28).

[98] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. “CNN Features Off-the-Shelf: An Astounding Baseline for Recognition”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Columbus, OH, USA: IEEE, June 2014, pp. 512–519 (cit. on p. 21).

[99] X. Robin, N. Turck, A. Hainard, et al. “pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves”. en. In: *BMC Bioinformatics* 12.1 (Dec. 2011), p. 77 (cit. on p. 52).

[100] E. M. Rohren, E. C. Etchebehere, J. C. Araujo, et al. “Determination of Skeletal Tumor Burden on 18F-Fluoride PET/CT”. en. In: *Journal of Nuclear Medicine* 56.10 (Oct. 2015), pp. 1507–1512 (cit. on p. 18).

[101] S. P. Rowe, K. J. Pienta, M. G. Pomper, and M. A. Gorin. “Proposal for a Structured Reporting System for Prostate-Specific Membrane Antigen–Targeted PET Imaging: PSMA-RADS Version 1.0”. en. In: *Journal of Nuclear Medicine* 59.3 (Mar. 2018), pp. 479–485 (cit. on pp. 28, 47).

[102] H. Rowley, S. Baluja, and T. Kanade. “Neural Network-Based Face Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.1 (Jan./1998), pp. 23–38 (cit. on p. 20).

[103] G. B. Saha. *Basics of PET Imaging*. en. New York, NY: Springer New York, 2010 (cit. on pp. 9, 10, 12).

[104] M. Sasanelli, M. Meignan, C. Haioun, et al. “Pretherapy Metabolic Tumour Volume Is an Independent Predictor of Outcome in Patients with Diffuse Large B-Cell Lymphoma”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 41.11 (Nov. 2014), pp. 2017–2022 (cit. on p. 49).

[105] B. Savir-Baruch, L. Zanoni, and D. M. Schuster. “Imaging of Prostate Cancer Using Fluciclovine”. en. In: *PET Clinics* 12.2 (Apr. 2017), pp. 145–157 (cit. on p. 13).

[106] C. Schmidkonz, M. Cordes, D. Schmidt, et al. “68Ga-PSMA-11 PET/CT-Derived Metabolic Parameters for Determination of Whole-Body Tumor Burden and Treatment Response in Prostate Cancer”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 45.11 (Oct. 2018), pp. 1862–1872 (cit. on pp. 17, 28, 29).

[107] S. Schmuck, C. A. von Klot, C. Henkenberens, et al. “Initial Experience with Volumetric ^{68} Ga-PSMA I&T PET/CT for Assessment of Whole-Body Tumor Burden as a Quantitative Imaging Biomarker in Patients with Prostate Cancer”. en. In: *Journal of Nuclear Medicine* 58.12 (Dec. 2017), pp. 1962–1968 (cit. on p. 18).

[108] L. H. Sehn, B. Berry, M. Chhanabhai, et al. “The Revised International Prognostic Index (R-IPI) Is a Better Predictor of Outcome than the Standard IPI for Patients with Diffuse Large B-Cell Lymphoma Treated with R-CHOP”. en. In: *Blood* 109.5 (Mar. 2007), pp. 1857–1861 (cit. on p. 50).

[109] R. Seifert, K. Herrmann, J. Kleesiek, et al. “Semiautomatically Quantified Tumor Volume Using ^{68} Ga-PSMA-11 PET as a Biomarker for Survival in Patients with Advanced Prostate Cancer”. en. In: *Journal of Nuclear Medicine* 61.12 (Dec. 2020), pp. 1786–1792 (cit. on pp. 17, 28, 37).

[110] R. Seifert, K. Kessel, K. Schlack, et al. “PSMA PET Total Tumor Volume Predicts Outcome of Patients with Advanced Prostate Cancer Receiving [177Lu]Lu-PSMA-617 Radioligand Therapy in a Bicentric Analysis”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 48.4 (Apr. 2021), pp. 1200–1210 (cit. on pp. 17, 28).

[111] S. Sheikhbahaei, A. Afshar-Oromieh, M. Eiber, et al. “Pearls and Pitfalls in Clinical Interpretation of Prostate-Specific Membrane Antigen (PSMA)-Targeted PET Imaging”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 44.12 (Nov. 2017), pp. 2117–2136 (cit. on p. 28).

[112] P. D. Shreve, Y. Anzai, and R. L. Wahl. “Pitfalls in Oncologic Diagnosis with FDG PET Imaging: Physiologic and Benign Variants”. en. In: *RadioGraphics* 19.1 (Jan. 1999), pp. 61–77 (cit. on p. 13).

[113] L. Sibille, N. Avramovic, B. Spottiswoode, M. Schaefers, S. Zuehlsdorff, and J. Declerck. “PET Uptake Classification in Lymphoma and Lung Cancer Using Deep Learning”. In: *Journal of Nuclear Medicine* 59.supplement 1 (May 2018), p. 325 (cit. on p. 50).

[114] L. Sibille, R. Seifert, N. Avramovic, et al. “^{18} F-FDG PET/CT Uptake Classification in Lymphoma and Lung Cancer by Using Deep Convolutional Neural Networks”. en. In: *Radiology* 294.2 (Feb. 2020), pp. 445–452 (cit. on pp. 13, 28, 31, 50, 51).

[115] R. L. Siegel, K. D. Miller, and A. Jemal. “Cancer Statistics, 2020”. en. In: *CA: A Cancer Journal for Clinicians* 70.1 (Jan. 2020), pp. 7–30 (cit. on pp. 27, 41).

[116] K. Simonyan and A. Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *arXiv preprint arXiv:1409.1556* (2014). arXiv: `1409.1556` (cit. on p. 20).

[117] M.-K. Song, J.-S. Chung, H.-J. Shin, et al. “Clinical Significance of Metabolic Tumor Volume by PET/CT in Stages II and III of Diffuse Large B Cell Lymphoma without Extranodal Site Involvement”. en. In: *Annals of Hematology* 91.5 (May 2012), pp. 697–703 (cit. on p. 49).

[118] E. Steyerberg, M. Roobol, M. Kattan, T. van der Kwast, H. de Koning, and F. Schröder. “Prediction of Indolent Prostate Cancer: Validation and Updating of a Prognostic Nomogram”. en. In: *Journal of Urology* 177.1 (Jan. 2007), pp. 107–112 (cit. on p. 41).

[119] J. Strosberg, G. El-Haddad, E. Wolin, et al. “Phase 3 Trial of ^{177} Lu-Dotatate for Midgut Neuroendocrine Tumors”. en. In: *New England Journal of Medicine* 376.2 (Jan. 2017), pp. 125–135 (cit. on p. 14).

[120] C. Szegedy, Wei Liu, Yangqing Jia, et al. “Going Deeper with Convolutions”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, June 2015, pp. 1–9 (cit. on p. 20).

[121] D. Taïeb, R. J. Hicks, E. Hindié, et al. “European Association of Nuclear Medicine Practice Guideline/Society of Nuclear Medicine and Molecular Imaging Procedure Standard 2019 for Radionuclide Imaging of Phaeochromocytoma and Paraganglioma”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 46.10 (Sept. 2019), pp. 2112–2137 (cit. on p. 14).

[122] Y. Tao, Z. Peng, A. Krishnan, and X. S. Zhou. “Robust Learning-Based Parsing and Annotation of Medical Radiographs”. In: *IEEE Transactions on Medical Imaging* 30.2 (Feb. 2011), pp. 338–350 (cit. on p. 51).

[123] C. Thieblemont, H. Tilly, M. Gomes da Silva, et al. “Lenalidomide Maintenance Compared With Placebo in Responding Elderly Patients With Diffuse Large B-Cell Lymphoma Treated With First-Line Rituximab Plus Cyclophosphamide, Doxorubicin, Vincristine, and Prednisone”. en. In: *Journal of Clinical Oncology* 35.22 (Aug. 2017), pp. 2473–2481 (cit. on p. 50).

[124] A. Toriihara, T. Nobashi, L. Baratto, et al. “Comparison of 3 Interpretation Criteria for ^{68} Ga-PSMA11 PET Based on Inter- and Intrareader Agreement”. en. In: *Journal of Nuclear Medicine* 61.4 (Apr. 2020), pp. 533–539 (cit. on p. 47).

[125] C. Van de Wiele, V. Kruse, P. Smeets, M. Sathekge, and A. Maes. “Predictive and Prognostic Value of Metabolic Tumour Volume and Total Lesion Glycolysis in Solid Tumours”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 40.2 (Jan. 2013), pp. 290–301 (cit. on p. 18).

[126] J. van den Hoff, L. Oehme, G. Schramm, et al. “The PET-Derived Tumor-to-Blood Standard Uptake Ratio (SUR) Is Superior to Tumor SUV as a Surrogate Parameter of the Metabolic Rate of FDG”. en. In: *EJNMMI Research* 3.1 (2013), p. 77 (cit. on p. 17).

[127] M. van Kruchten, A. W. J. M. Glaudemans, E. F. J. de Vries, et al. “PET Imaging of Estrogen Receptors as a Diagnostic Tool for Breast Cancer Patients Presenting with a Clinical Dilemma”. en. In: *Journal of Nuclear Medicine* 53.2 (Feb. 2012), pp. 182–190 (cit. on p. 14).

[128] L. Vercellino, A.-S. Cottereau, O. Casasnovas, et al. “High Total Metabolic Tumor Volume at Baseline Predicts Survival Independent of Response to Therapy”. en. In: *Blood* 135.16 (Apr. 2020), pp. 1396–1405 (cit. on pp. 50, 51).

[129] P. Viola and M. Jones. “Rapid Object Detection Using a Boosted Cascade of Simple Features”. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Vol. 1. Kauai, HI, USA: IEEE Comput. Soc, 2001, pp. I–511–I–518 (cit. on p. 19).

[130] R. L. Wahl, H. Jacene, Y. Kasamon, and M. A. Lodge. “From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors”. en. In: *Journal of Nuclear Medicine* 50.Suppl_1 (May 2009), 122S–150S (cit. on pp. 16, 29, 51).

[131] A. J. Weisman, M. W. Kieler, S. Perlman, et al. “Comparison of 11 Automated PET Segmentation Methods in Lymphoma”. In: *Physics in Medicine & Biology* 65.23 (Nov. 2020), p. 235019 (cit. on p. 37).

[132] Y. Zhao, A. Gafita, B. Vollnberg, et al. “Deep Neural Network for Automatic Characterization of Lesions on 68Ga-PSMA-11 PET/CT”. en. In: *European Journal of Nuclear Medicine and Molecular Imaging* 47.3 (Mar. 2020), pp. 603–613 (cit. on pp. 28, 37).

[133] Z. Zhou, L. H. Sehn, A. W. Rademaker, et al. “An Enhanced International Prognostic Index (NCCN-IPI) for Patients with Diffuse Large B-Cell Lymphoma Treated in the Rituximab Era”. en. In: *Blood* 123.6 (Feb. 2014), pp. 837–842 (cit. on p. 50).

[134] C. Zippel, S. C. Ronski, S. Bohnet-Joschko, F. L. Giesel, and K. Kopka. “Current Status of PSMA-Radiotracers for Prostate Cancer: Data Analysis of Prospective Trials Listed on ClinicalTrials.Gov”. en. In: *Pharmaceuticals* 13.1 (Jan. 2020), p. 12 (cit. on p. 13).

List of Figures

List of Tables