Evaluation of Developments in PET Methodology

Pike, Lucy

Awarding institution:
King's College London

Evaluation of Developments in PET Methodology

Lucy Pike

School of Biomedical Engineering and Imaging Sciences

King's College London

This dissertation is submitted for the degree of

Doctor of Philosophy

May 2023

Abstract

Commercial positron emission tomography (PET) systems are continually evolving as manufacturers strive to improve image quality, whilst reducing radiation dose and acquisition times. The impact of these developments on clinically relevant measures must be evaluated to ensure PET imaging is used appropriately. Clinical trials designed specifically to evaluate new developments in PET technology or methodology are expensive and time consuming, so alternative evaluation approaches are required. One potential way to evaluate new PET developments is to use simulated PET data along with forward modelling in a 'virtual imaging trial' framework. The purpose of this project was to develop and validate the methodology for generating PET datasets created by insertion of simulated lesions into real clinical PET datasets. This approach required an accurate and reliable method for combining simulated lesions where the ground truth was known, with the physiological uptake in real patient data with a realistic range of weight, body mass indices and image noise levels. The validated methodology was then used for the evaluation of new PET technologies, in particular use of time-of-flight (TOF), point spread function (PSF) modelling and a Bayesian penalised likelihood (BPL) reconstruction algorithm.

The first stage of the work in this thesis involved designing a model of the General Electric (GE) Healthcare Discovery 710 PET scanner to perform Monte Carlo simulations of realistic lesions and to develop the methodology for inserting the lesions into real PET datasets in projection space. To validate the scanner model, measurements of spatial resolution and sensitivity were performed according to the National Electrical Manufacturers Association (NEMA) NU-2 standard using simulated phantoms and results were compared to those measured for the real

scanner. Additionally, accuracy of corrections was assessed using a simulated acquisition of a uniform cylinder.

The validity of the insertion technique was tested by comparing recovery coefficients derived from simulated spheres inserted into real PET datasets of the background compartment of the NEMA image quality phantom to those from real acquisitions of the phantom containing identical sized spheres. Anonymised PET datasets from 10 patients who had a measurable pulmonary lesion were used to assess the ability of the technique to generate realistic lesions. Characteristics from the real lesions were used to generate simulated lesions that were then inserted into the contralateral lung of the same patient. To demonstrate the simulated lesions were indistinguishable from real lesions, a two-alternative forced choice task was performed by an experienced PET physician whereby they were asked to review the reconstructed PET images for each patient and choose which of the two lesions they thought was simulated and rate their confidence in identifying it.

To ensure that PET images created using the insertion technique could be used in place of real PET data and a real PET imaging system for performing clinically relevant tasks, PET images were generated for a cohort of 97 patients consisting of simulated lesions with characteristics matching those from a population of real patients with known solitary pulmonary nodules (SPNs). Quantitative measures of ^{18}Fluorine-labelled fluorodeoxyglucose (^{18}F-FDG) uptake and diagnostic accuracy for the assessment of malignancy risk were then compared between the cohorts of patients with real and simulated lesions.

Finally, the impact of new PET technologies on image quality and task-based measures was evaluated using phantom and patient images generated using the validated methodology. The influence of TOF, PSF modelling and BPL reconstruction on image quality was investigated using technical image-based measures for the NEMA image quality phantom. For the task-based assessment, 194 lesions were simulated with characteristics representative of real benign and malignant SPNs and inserted into the anonymised raw PET datasets from 194 patients. The PET data were reconstructed using parameters in use across clinical PET Centres in the UK that incorporated TOF, PSF modelling and BPL. The resultant datasets were used to determine

the influence of incorporating new technologies on measurements of ^{18}F-FDG uptake using standardised uptake values (SUVs) and the subsequent impact on the diagnostic accuracy for categorising the lesions as malignant or benign. The diagnostic accuracy for each reconstruction was assessed using lesion SUV_{max}, lesion SUV_{mean} and the Herder score which combines the ^{18}F-FDG uptake with the patient clinical and radiological characteristics to determine risk of malignancy.

Acknowledgements

I would like to thank my supervisors, Prof. Sally Barrington and Prof. Paul Marsden for all their support and guidance throughout this PhD and during my time working at King's College London.

I would also like to acknowledge all the people who provided guidance and support throughout this PhD. Manil Subesinghe for answering all my questions and providing data for the real SPN data set used for validation in Chapter 4. The staff at the King's College London & Guy's and St Thomas' PET Centre, especially Joemon John and Armidita Jacob, who allowed me time on the scanners for performing the phantom experiments and for performing the selection and anonymisation of the data sets used for insertion in this thesis. My thesis committee panel Lefteris Livieratos, Joel Dunn, Shaihan Malik and Robert Eckersley for their challenging and helpful discussions throughout my PhD.

Finally, I'd like to thank Jane, Sammy and my family for all their love and encouragement throughout this PhD.

The Duetto PET Toolbox used for reconstruction of the PET data in this thesis is a General Electric (GE) Healthcare proprietary software development tool and was provided by GE Company through a Software Development Kit License.

This work was undertaken as part of a study with Health Research Authority and Health and Care Research Wales approval, Integrated Research Application System (IRAS) project ID: 251611. Lesion insertion using patient scans was limited to use of previously collected, non-identifiable data. Therefore, approval by the research ethics committee was not required.

The author acknowledges support from the National Institute for Health and Care Research (NIHR) [RP-2-16-07-001]. King's College London and University College London Comprehensive Cancer Imaging Centre is funded by Cancer Research UK and the Engineering and Physical Sciences Research Council in association with the Medical Research Council and Department of Health and Social Care (DHSC) (England). The views expressed are those of the author and not necessarily those of the National Health Service, the NIHR or the DHSC.

Table of contents

List of figures

List of tables

Acronyms

^{18}F-FDG . . . . . ^{18}Fluorine-labelled fluorodeoxyglucose

2AFC . . . . . two-alternative forced choice

2D . . . . . two-dimensional

3D . . . . . three-dimensional

4D . . . . . four-dimensional

ACF . . . . . attenuation correction factor

ACR . . . . . American College of Radiology

AI . . . . . artificial intelligence

ASIM . . . . . PET analytic simulation package

ASiR . . . . . adaptive statistical iterative reconstruction

AUC . . . . . area under the ROC curve

BGO . . . . . bismuth germanate

BMI . . . . . body mass index

BPL . . . . . Bayesian penalised likelihood

BREP . . . . . boundary representation

BTS . . . . . British Thoracic Society

BV . . . . . background variability

CAD . . . . . computer-aided design

CC . . . . . non-stem cancer cell

CI . . . . . confidence interval

CNR . . . . . contrast-to-noise ratio

COV . . . . . coefficient of variation

cps . . . . . counts per second

CSC . . . . . cancer stem cell

CSG . . . . . constructive solid geometry

CT . . . . . computed tomography

CVIT . . . . . Centre for Virtual Imaging Trials

DDG . . . . . data-driven respiratory gating

DHSC . . . . . Department of Health and Social Care

DICOM . . . . . Digital Imaging and Communications in Medicine

DLA . . . . . diffusion limited aggregation

DV . . . . . distribution volume

EARL . . . . . European Association of Nuclear Medicine Research Ltd

EN . . . . . ensemble noise

FBP . . . . . filtered backprojection

FORE-FBP . . . . . Fourier rebinning followed by 2D FBP

FOV . . . . . field of view

FWHM . . . . . full-width at half-maximum

FWTM . . . . . full width at tenth maximum

GAN . . . . . generative adversarial network

GATE . . . . . Geant4 application for tomographic emission

GE . . . . . General Electric

Geant4 . . . . . for geometry and tracking

HU . . . . . Hounsfield unit

ICC . . . . . intraclass correlation coefficient

ICRP . . . . . International Commission on Radiological Protection

IR . . . . . image roughness

IRAS . . . . . Integrated Research Application System

LCP . . . . . lung cancer prediction

LOR . . . . . line of response

LYSO . . . . . lutetium-yttrium-orthosilicate

MAP . . . . . maximum *a posteriori*

MBP . . . . . mediastinal blood pool

MIM . . . . . Medical Image Merge

ML-EM . . . . . maximum likelihood-expectation maximisation

MRI . . . . . magnetic resonance imaging

MTV . . . . . metabolic tumour volume

NECR . . . . . noise equivalent count rate

NEMA . . . . . National Electrical Manufacturers Association

NIHR . . . . . National Institute for Health and Care Research

NOPR . . . . . National Oncologic PET Registry

NPV . . . . . negative predictive values

NURBS . . . . . non-uniform rational B-spline

OS-EM . . . . . ordered subsets-expectation maximisation

PERCIST . . . . . Positron Emission Tomography Response Criteria In Solid Tumors

PET . . . . . positron emission tomography

PMT . . . . . photomultiplier tube

PPV . . . . . positive predictive values

PSF . . . . . point spread function

PVC . . . . . partial volume correction

PVE . . . . . partial volume effect

QC . . . . . quality control

QCFX . . . . . BPL with TOF and PSF

QIBA . . . . . Quantitative Imaging Biomarkers Alliance

RAM . . . . . random access memory

RC . . . . . recovery coefficient

ROC . . . . . receiver operating characteristic

RW . . . . . random walk

SBR . . . . . sphere-to-background ratio

SD . . . . . standard deviation

Se . . . . . sensitivity

SiPM . . . . . silicon photomultiplier

SNMMI . . . . . Society of Nuclear Medicine and Molecular Imaging

SNR . . . . . signal-to-noise ratio

Sp . . . . . specificity

SPN . . . . . solitary pulmonary nodule

SSRB . . . . . single-slice rebinning

STIR . . . . . software for tomographic image reconstruction

SUV . . . . . standardised uptake value

TLG . . . . . total lesion glycolysis

TOF . . . . . time-of-flight

VCT . . . . . virtual clinical trial

VIT . . . . . virtual imaging trial

VOI . . . . . volumes of interest

VPFX . . . . . 3D OS-EM with TOF

VPFXS . . . . . 3D OS-EM with TOF and PSF

VPHD . . . . . 3D OS-EM

XCAT . . . . . 4D extended cardiac-torso

Chapter 1

Introduction

1.1 The Role of Positron Emission Tomography in Oncology

Positron emission tomography (PET) is a non-invasive functional imaging technique used widely in the management of patients with cancer from initial diagnosis and staging, response assessment during and after treatment and for guiding radiotherapy treatment [1]. PET is inherently quantitative and, with careful equipment calibration and consistent methodology, measurements of tracer uptake by tumours or metabolic tumour volume have been shown to correlate well with patient outcomes [2]. Functional imaging with PET can often pick up changes before anatomical changes are seen on computed tomography (CT) and magnetic resonance imaging (MRI), allowing clinicians to adapt treatment earlier and spare patients from unnecessary treatments and side-effects [3, 4]. For example, in recent lymphoma clinical trials patient treatment was adapted (escalated/de-escalated) based on results of ^{18}Fluorine-labelled fluorodeoxyglucose (^{18}F-FDG)-PET scans performed during chemotherapy [5, 6]. These studies showed better outcomes for patients with fewer side-effects and improved survival.

The clinical focus of this thesis was on the use of PET in the characterisation of solitary pulmonary nodules (SPNs) therefore a full review of oncological applications of PET is beyond the scope of this work. However, there is a strong-evidence base for the use of PET-CT in a variety of oncological and non-oncological diseases as summarised in the current UK evidence-based

guidance for the use of PET-CT [1]. For indeterminate SPNs, the British Thoracic Society (BTS) guidelines recommend ^{18}F-FDG-PET-CT imaging as part of the diagnostic pathway [7]. Any solid non-calcified pulmonary nodules identified on CT, either through screening or incidental findings, are initially assessed based on their radiological size and appearance. Nodules $\geq$ 8 mm diameter (or $\geq$ 300 mm^3 volume) are then assessed for risk of malignancy using the Brock model [8] with nodules $\geq$ 10 % risk undergoing further investigation using ^{18}F-FDG-PET-CT. The addition of ^{18}F-FDG-PET-CT in determining the risk of malignancy of indeterminate SPNs has been shown to improve the diagnostic accuracy compared to CT alone [9, 10] and a meta-analysis by Gould *et al* found a mean sensitivity and specificity of 93.9 % and 85.8 % respectively [11].

1.2 Methodological Advances in PET

Compared to other imaging modalities, such as CT or MRI, PET suffers from relatively low spatial resolution. Commercial PET systems are continually evolving as manufacturers strive to improve the image quality and reduce imaging times. The first clinical PET scanner in the UK was installed in 1992 [12], at this time PET systems did not include an integrated CT scanner and only acquired PET data in two-dimensional (2D) mode. Typical intrinsic spatial resolution for these early clinical systems was in the order of 6 mm full-width at half-maximum (FWHM) and modest improvements are seen in the current wholebody PET systems with FWHM measured as 4.1 mm to 5.9 mm [13, 14, 15]. With further improvements in detector design, it is theoretically possible to achieve clinical PET scanners with a spatial resolution of 2.36 mm [16]. Notable advances in PET resulting in improved sensitivity and signal-to-noise ratio (SNR) include three-dimensional (3D) acquisition modes, faster scintillation crystals allowing the use of time-of-flight (TOF) information and improved reconstruction algorithms with more accurate corrections incorporated in the system model [17, 18].

1.3 Challenges in Assessment of Advances in PET

As a result of these advances, there are a range of software and hardware options available in the clinical setting, not all of which have been independently evaluated in clinical trials. To ensure any developments are put into practice appropriately, the impact of changes on clinically relevant measures must be carefully evaluated and understood. However, it is difficult to quantify the clinical impact of improved hardware and reconstruction methods as they are not necessarily related to the more simple technical parameters associated with physical detector characteristics [17]. Clinical trials, particularly large phase II/III randomised trials, are expensive, complex and take a long time to complete [19]. Results of clinical trials often take years to be published and implemented into clinical practice in which time the technology is likely to have moved forward. As such, it is not feasible to conduct clinical trials every time there is a change in PET scanner technology or methodology, and there is a need for more efficient, validated methods of technology assessment to better inform clinicians and researchers how to incorporate new PET developments into clinical trials and how this translates into clinical practice.

The accurate diagnosis and staging of cancer is crucial in selecting the appropriate course of treatment to improve outcomes for patients whilst reducing unnecessary risks of morbidity and costs associated with many therapies. PET imaging plays a significant role in this decision-making process through detecting the presence and spread of cancer and characterising the nature of the disease. Assessment of the clinical and cost-effectiveness of PET or indeed any new imaging technology is complex. Several projects, such as the National Oncologic PET Registry (NOPR) in the United States, have collected prospective data on the referring physicians' intended patient management before and after the ^{18}F-FDG-PET study to assess the impact of including the PET information on patient management [20]. This type of study requires significant logistical planning and takes considerable time and expense to accrue the data. Additionally, the assessment doesn't take account of whether the clinician or patient opts to act on the information provided by the PET, the efficacy of the treatments available at the time, or the natural history of the disease itself [21]. Randomised controlled trials that typically apply

the intention-to-treat methodology are also inappropriate in this context as the statistical analysis includes all participants who are randomized to have a PET scan according to the group they were originally assigned, regardless of whether or not the findings on PET were included in the final treatment decision [21].

The performance of new PET technology and methodology, such as a new image reconstruction algorithm, is often measured in terms of technical image-based metrics like SNR, noise equivalent count rate (NECR) or quantitative accuracy. These measurements are often made with physical or simulated phantoms, where the ground truth is known, providing an objective measure of the technical scanner performance that can be replicated and compared across systems. However, these measures of image quality do not take into account the complex anatomy and physiology present in real patients and often do not correlate well with perceived image quality by clinicians. Additionally, phantom and simulation assessments do not provide a task-specific evaluation, and do not include many relevant variables that are likely to affect the properties of the final image. These include patient-related factors such as body habitus, variation in tracer biodistribution related to medication and/or sub-optimal pre-scan preparation and variation in attenuation properties e.g. as a result of patient movement during the acquisition.

The performance may also depend strongly on the method of analysis used. In clinical practice, diagnostic tasks include lesion detection, characterisation of tumours using measurements of radiopharmaceutical uptake or functional volume at a single time point and measurements of changes in uptake or functional volume to assess response to therapy from a series of scans at different stages of treatment. These diagnostic tasks are broadly divided into two types: classification and estimation tasks. Classification tasks would include lesion detection and/or localisation whilst estimation tasks include measurements of parameters that help to characterise the lesion such as the degree of radiotracer uptake or metabolic tumour volume (MTV) [22]. The characterisation of task-specific performance indices such as these in terms of bias and variability is in general very hard to determine since the ground truth is unknown and is not necessarily related in a simple way to the technical image-based metrics.

1.4 Virtual Imaging Trials

Use of modelling tools combining real clinical data and simulated image data in a virtual (or '*in silico*') imaging trial framework has been proposed as a method for evaluating new imaging technologies. In a virtual imaging trial (VIT) framework, a validated model of the real imaging system is used to generate images for a virtual population of patients using anatomically and physiologically realistic phantoms. By accurately modelling the acquisition process, images can be generated where the ground truth is known and that incorporate the noise and resolution characteristics of the physical imaging system. Systematic comparisons can then be made through controlling different aspects of the patient population or imaging system to optimise the imaging methodology for specific clinical applications. This makes VITs much more efficient and cost effective than traditional clinical trials using real patients.

As discussed further in Section 2.3, VITs have been used for comparing 2D and 3D mammography in breast cancer screening [23, 24] and to assess the ability of PET to detect a given change [25, 26, 27, 28]. More recently, the Centre for Virtual Imaging Trials (CVIT) was established at Duke University (\textcolor{blue}{https://cvit.duke.edu/about/cvit/}) with the aim of using VITs to improve the diagnostic accuracy of CT in specific clinical applications. To achieve this, the group are developing the tools for conducting VITs including a framework for creating virtual patients using computational phantoms, manufacturer-specific models of CT scanners, and task-based mathematical observers. Use of VITs in this way permits the optimisation of CT protocols for a wide range of clinical applications and technology whist minimising the radiation dose for patients.

The VIT concept can also be used in 'digital twin' studies. A digital twin is a virtual representation of an individual patient generated by combining real-time data from the patient and their environment with a model derived from population-based data. The digital twin can then be used to investigate the influence of different factors and predict outcomes for the real patient to aid in clinical decisions [29]. This concept has potential applications for personalising patient care in nuclear medicine theranostics where the same target is used for radionuclide imaging

and therapy [30]. By developing digital twins using mathematical models that incorporate the pre-therapy radionuclide imaging to perform predictive dosimetry, the therapy plan (injected activity, number of cycles etc) can be optimised for an individual patient.

1.5 Thesis Objectives and Structure

The purpose of this thesis was two-fold. First was to develop and validate the methodology for generating PET images consisting of simulated lesions inserted into PET datasets from real patients. This was achieved through the design and testing of a model of the PET system for performing Monte Carlo simulations and developing the steps for insertion of simulated objects into existing PET datasets in projection space. The methodology was validated through comparing quantitative measures for real solitary pulmonary nodules (SPNs) in a population of patients to simulated lesions with matching characteristics inserted into patient datasets.

The next stage was to use the described methodology to investigate the impact of new technologies, in particular time-of-flight (TOF), point spread function (PSF) modelling and the Bayesian penalised likelihood (BPL) reconstruction algorithm on image quality and task-based measures. Assessments of image quality were performed using the National Electrical Manufacturers Association (NEMA) image quality phantom. The impact of new technologies on ^{18}F-FDG uptake measurements and subsequent determination of risk of malignancy for patients with SPNs was evaluated using PET images generated using the validated methodology in a virtual imaging trial.

The key aims of this thesis were to:

1. to develop and validate a methodology for generating PET images consisting of real patient PET datasets with realistic simulated lesions

2. to use the validated methodology to generate images for use in a virtual imaging trial to evaluate the impact of new PET technologies on task-based measures

Chapter 2 provides an overview of the principles of PET imaging including a description of the detector instrumentation and reconstruction techniques. Methods for assessing image quality are also described covering both technical image-based metrics and task-based assessments. The reader is also introduced to quantification in PET and common terminology used throughout this thesis. This is followed by a review of phantoms and Monte Carlo techniques used for image quality assessment. Lastly the concept of virtual imaging trials is introduced along with a review of VITs in the literature.

Chapter 3 describes the model of the PET system used for simulation of the lesions and the process of lesion insertion in projection space. The methodology and results are presented for a series of phantom experiments performed to validate the model and the insertion technique. Further, a visual review of 10 patient datasets with both a real and simulated SPN was performed by a clinical expert to assess the realism of the inserted lesions.

Chapter 4 describes the process for generating simulated SPNs based on the characteristics of an existing cohort of patients with real SPNs. The simulated lesions were then inserted into real patient datasets to generate a cohort of 97 patients with simulated SPNs that mirrored the data from a retrospective study performed on clinical patients with a known SPN referred for a PET scan at the King's College London & Guy's and St Thomas' PET Centre. Measurements of ^{18}F-FDG uptake and diagnostic accuracy for the cohort of patients with simulated lesions were compared to those for the cohort of patients with real SPNs.

Chapter 5 goes on to evaluate the impact of TOF, PSF modelling and BPL reconstruction on image quality and task-based measures. The NEMA image quality phantom was used to investigate technical image-based measures of image quality. PET images containing simulated lesions with characteristics representative of real benign and malignant SPNs were generated using the validated methodology described in the previous chapters and used in a VIT to determine the impact of incorporating the new technologies on ^{18}F-FDG uptake measurements and subsequent determination of risk of malignancy.

Chapter 6 provides overall conclusions of the work in this thesis and potential areas for further use of the methodology described.

Chapter 2

Background

2.1 Principles of Positron Emission Tomography

During a positron emission tomography (PET) study, patients are injected with a chemical compound that has been labelled with a positron-emitting radionuclide, known as a radiopharmaceutical. The radiopharmaceutical circulates in the blood and may accumulate in tissues or organs dependent on the biological pathway of the labelled compound. Images are acquired of the radiopharmaceutical distribution in the patient using a PET scanner. The scanner consists of a ring of detectors around the patient that detect the photons emitted from the radiopharmaceutical which are reconstructed into three-dimensional (3D) images. This provides a non-invasive method to investigate biological or physiological processes within the body. There are a range of evidence-based indications for PET-CT [1], particularly in oncology, with the most widely used radiopharmaceutical being ^{18}Fluorine-labelled fluorodeoxyglucose (^{18}F-FDG).

Normal differentiated cells primarily generate energy through mitochondrial oxidative phosphorylation which requires oxygen. Whereas, even in the presence of ample oxygen, cancer cells predominantly generate energy through aerobic glycolysis whereby glucose is broken down to produce energy in the form of adenosine triphosphate. This increase in glucose demand in cancer cells compared to normal cells is known as the Warburg effect [31] and is exploited in ^{18}F-FDG-PET for imaging malignant tumours. ^{18}F-FDG is a radiolabelled glucose analog

and, like glucose, is transported into the cells via glucose transporters and phosphorylated by hexokinase. Unlike glucose however, the FDG forms FDG-6-phosphate which cannot be further metabolised and is trapped in the cancer cells. The accumulation of FDG-6-phosphate in the cancer cells therefore results in increased radioactivity compared to the surrounding tissues which is detected by the PET scanner.

^{18}Fluorine (^{18}F) is produced in a cyclotron by irradiating an ^{18}Oxygen target (usually ^{18}O enriched water) with a high energy beam of protons. The proton interacts with the ^{18}O nucleus creating ^{18}F. The physical and nuclear properties of ^{18}F are given in Table 2.1.

Property of ^{18}F	**Value**
Half-Life	109.77 min
Positron Fraction	0.967
Mean Positron Energy	0.250 MeV
Max Positron Energy	0.634 MeV
Mean Positron Range in Water	0.6 mm
Max Positron Range in Water	2.4 mm

Table 2.1 Physical and nuclear properties of ^{18}Fluorine [32].1 MeV = 1×10^6 electron volts (eV).

2.1.1 Positron Decay

The radionuclide contains a proton-rich nucleus making it unstable. To revert to its stable state, the radionuclide undergoes positron decay whereby a proton is converted to a neutron along with the emission of a positron (e^+) and an electron neutrino (ν_e). The emitted positron travels a short distance during which it loses kinetic energy as it undergoes inelastic collisions with orbital electrons or nuclei. Once almost all the kinetic energy is lost, the positron undergoes annihilation with an electron (e^-) resulting in the emission of two annihilation photons, as shown in Figure 2.1.

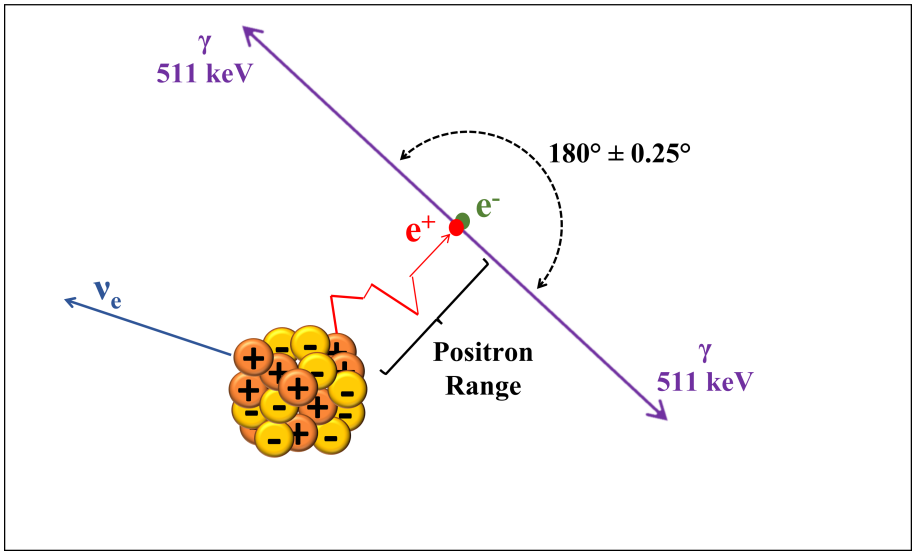

Fig. 2.1 The unstable radionuclide undergoes positron decay whereby a proton (+) is converted to a neutron (-) along with the emission of a positron (e^+) and an electron neutrino (ν_e). After travelling a short distance, the positron undergoes annihilation with an electron (e^-) producing two photons (γ) at approximately 180° to one another.

The distance from where the positron was emitted from the nucleus to the location where annihilation occurred is known as the positron range. The size of the positron range depends on the initial positron energy, which is dependent on the radionuclide, and the properties of the material it passes through [32]. The finite positron range results in blurring of the source distribution reducing the spatial resolution. The positron range for ^{18}F is very short (Table 2.1) resulting in a relatively small blurring effect with full-width at half-maximum (FWHM) of 0.54 mm [16] and therefore is not a significant limiting factor in the resolution for current PET systems.

Another fundamental factor limiting the spatial resolution of PET systems is photon non-collinearity. Both the electron and positron have rest masses of 511 keV therefore to conserve mass and energy, the total energy of the two resultant annihilation photons is 1.022 MeV. However, since the positron and electron are not normally at complete rest when the annihilation event occurs, the annihilation photons are not emitted at exactly 180° to one another, but 180° ± 0.25°. The amount of blurring due to non-collinearity is dependent on the diameter of the ring of detectors with a FWHM of 1.76 mm for a PET scanner with 800 mm diameter ring [33].

The largest limiting factor to the fundamental resolution of PET scanners is the size of the individual detector elements. As you move across the face of the detector element, the response increases linearly from the edges towards the maximum at the centre. The FWHM of the response is equal to half the width of the detector element. There are practical limitations on the manufacture of very small detector elements, with detector widths for current clinical PET systems of 4 mm to 6 mm and the latest Siemens system incorporating 3.2 mm crystal elements [34].

2.1.2 Coincidence Detection

A coincidence event is when two annihilation photons are detected by opposing detectors in the PET scanner within a short time interval, known as the coincidence timing window. The location of the originating annihilation event is then assumed to be somewhere along the line of response (LOR) joining the two detectors as shown in Figure 2.2.

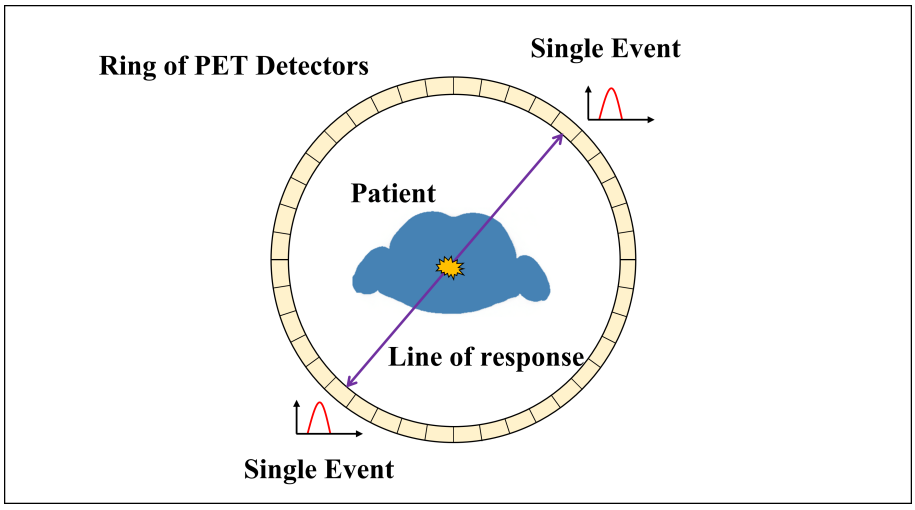

Fig. 2.2 Coincidence detection of annihilation photons by opposing detectors in a PET scanner.

The total number of coincidence events measured by opposing detectors represents the total radioactivity within the volume of patient tissue that lies along that LOR. The PET scanner is calibrated by measuring the scanner response in counts per second (cps) to a known amount of

radioactivity in Becquerels (Bq) in a known volume (ml) to provide quantitative measures of the radioactivity concentration (Bq ml^{-1}) in tissues.

2.1.2.1 True, Random and Scattered Coincidence Events

When two annihilation photons detected within the set coincidence timing window originate from a single annihilation event located along the LOR joining the two detectors it is known as a true event, Figure 2.3 a).

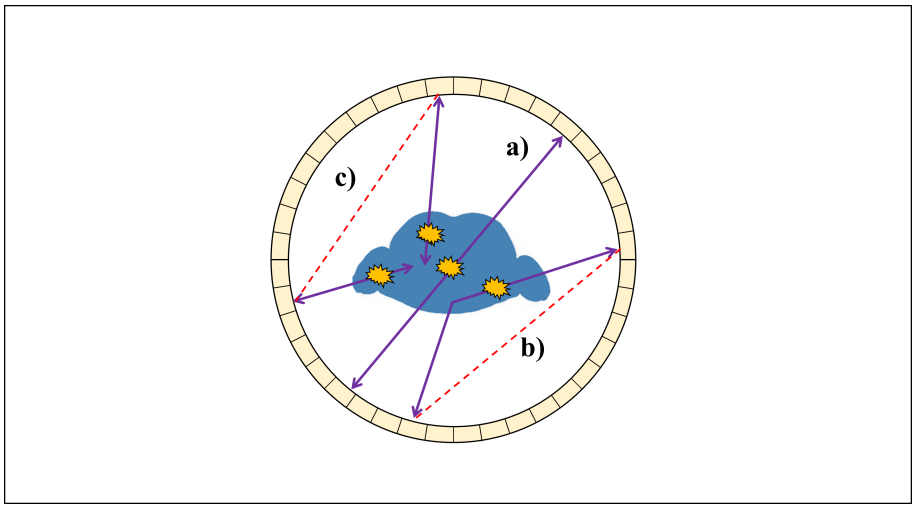

Fig. 2.3 Illustration of a) true, b) scattered and c) random coincidence events. If annihilation photons are scattered b) or photons from different events are detected within the same coincidence timing window c), the event is incorrectly located along the LOR joining the detectors as shown by the red dotted lines.

If one or both the annihilation photons from an annihilation event is scattered prior to detection, the event can be incorrectly located along the LOR joining the two detectors as seen in Figure 2.3 b). Scatter is mainly due to Compton scatter and can occur within the patient and the scanner components so the ratio of scattered to true events depends on patient and scanner geometry. Since scattered photons lose energy, the number of scattered photons detected is also dependent on the lower energy window threshold settings. Figure 2.4 shows the spectrum of photons detected for an ^{18}F source. The energy resolution of the PET scanner is defined as the FWHM of the 511 keV photopeak and the lower threshold on the energy window is set to include the photopeak whilst excluding as many of the scattered photons as possible.

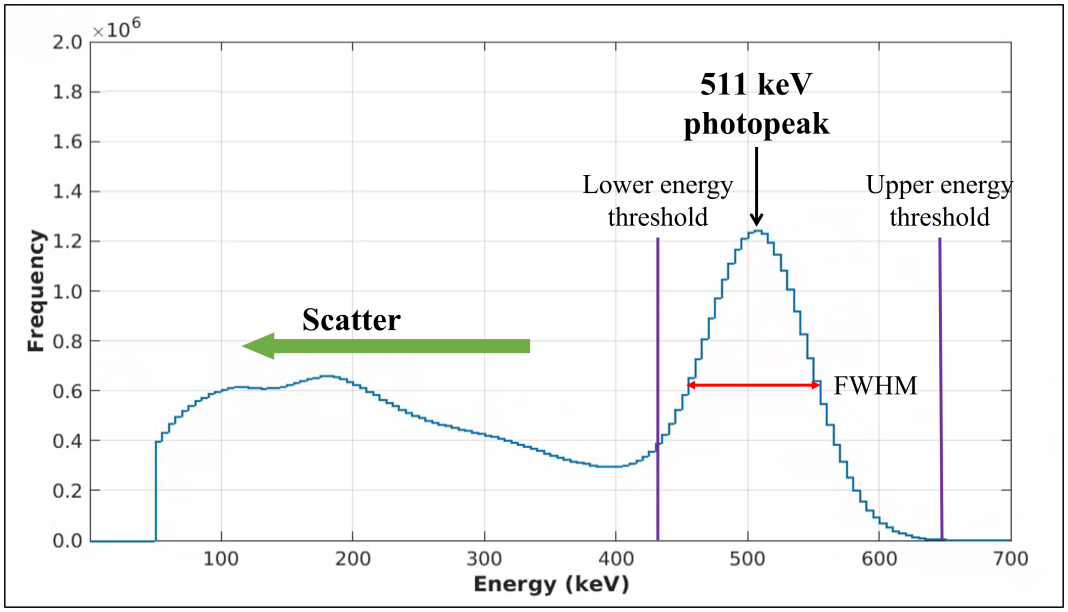

Fig. 2.4 Spectrum of coincident singles for an ^{18}F source showing the photopeak formed when the 511 keV annihilation photons deposit all their energy in the detector. A lower energy threshold is used to exclude a large proportion of the scattered events. The energy resolution of the scanner, is measured as the width of the photopeak at the point where it is half the maximum frequency (FWHM), as shown by the red arrow on the spectrum.

A random event is when the annihilation photons from two different annihilation events fall within the same timing window and the annihilation event is incorrectly assigned to the LOR joining the two detectors, Figure 2.3 c). The ratio of random to true coincidence events increases with radioactivity in the patient and decreases with decreasing width of the coincidence timing window.

2.1.3 PET Instrumentation

A PET detector or 'block detector' consists of a scintillator crystal segmented into elements with reflective material in the gaps, coupled to either an array of photomultiplier tubes (PMTs), as shown in Figure 2.5, or silicon photomultipliers (SiPMs).

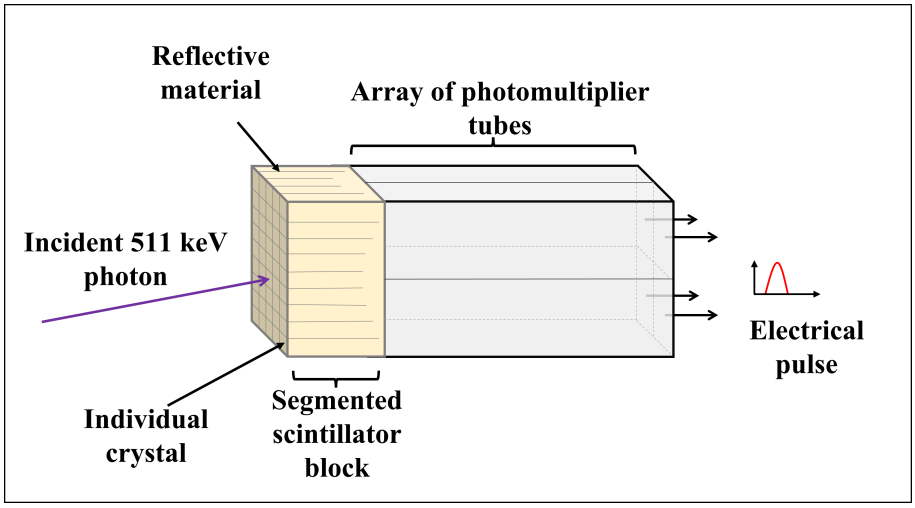

Fig. 2.5 Schematic of a PET detector showing the segmented scintillator crystal coupled to an array of PMTs.

The incident 511 keV annihilation photon deposits its energy in the crystal lattice exciting an electron from the valence band into the conduction band. On returning to the ground state, light photons are produced which are converted into an electrical pulse by the PMTs or SiPMs. The amplitude of the electrical pulse is proportional to the number of light photons produced in the crystal which in turn is proportional to the energy deposited by the incident annihilation photon. An energy window with upper and lower thresholds on the amplitude of the electrical pulse is used to reject scattered events with energies outside the expected 511 keV photopeak.

A good scintillation crystal is characterized by high light yield, fast decay time, high density and effective atomic number (Z_{eff}), and low self-absorption in the visible range of the electromagnetic spectrum. The two main scintillation materials used in clinical PET systems are bismuth germanate (BGO) and lutetium-based scintillator materials such as lutetium-yttrium-orthosilicate (LYSO). The key advantage of LYSO over BGO is the much faster decay time allowing the use of time-of-flight (TOF) information (as explained later in this section). Lutetium-based scintillator materials emit intrinsic radiation due to the presence of ^{176}Lu which decays by beta emission followed by gamma- and x-rays over a broad energy spectrum [35]. This gives rise to a background of true and random coincidence events originating from the crystals themselves. The scanner is unable to distinguish between these intrinsic events originating from the crystal

and the extrinsic events originating from the object being imaged. Bettinardi *et al* measured the total coincidence events originating from intrinsic radiation for the GE Discovery 690 as $\sim 1 \times 10^3$ counts per second (cps) [13]. The majority of these events are intrinsic randoms with an intrinsic trues rate of only $\sim$1 cps which can be considered negligible compared with typical coincidence rates of $\sim 258 \times 10^3$ cps.

PMTs have traditionally been the most widely used photodetector for PET systems. Due to their large physical size, one-to-one coupling with individual crystal elements is not possible with PMTs. Instead, a small number of PMTs (typically four) will cover all the elements in the segmented crystal block. To determine the individual element location, the signals from all PMTs are combined using Anger logic to provide the x and y positions [36]. Use of multiplexing in this way however results in a loss of intrinsic resolution. More recently PET systems using SiPMs have been available. This has been driven by the development of PET-MRI systems since PMT performance is degraded in the presence of a magnetic field. SiPMs can be tightly packed reducing light loss between individual photodetectors and they have fast single-photon timing response which leads to improved timing resolution for TOF systems.

2.1.3.1 PET Acquisition

Since positron emitting isotopes can be incorporated into compounds that follow fundamental processes in the human body, PET has the potential to measure physiological functions such as rate of glucose metabolism, regional blood flow or receptor concentrations in absolute terms. This requires sequential imaging of tissues over time (dynamic imaging) to obtain time-activity curves of the supply of radiopharmaceutical to the tissue (known as the input function) and the activity in the tissue (the output function). The data is then fitted to a kinetic model of the biological system which separates the tissues and blood into compartments with a series of linear equations that describe the exchange of materials between the compartments. Where the arterial blood pool is not within the imaging field of view, blood sampling is required to determine the input function. Due to the complexity of this type of study, absolute quantification is not feasible for

routine scanning, therefore clinical PET scans consist of a single static acquisition at a particular time after injection, known as the uptake time, to provide a 'snapshot' of the radiopharmaceutical distribution. The uptake time is typically chosen as 60 minutes post injection for ^{18}F-FDG-PET in oncology [37].

For oncology applications most clinical PET scans are acquired from mid-thigh to the base of the brain or vertex. The axial field of view (FOV) for current clinical PET scanners is 15 cm to 26 cm, meaning that only a section of the patient can be imaged at one time. Acquisitions are therefore performed with the patient moving through the scanner bore acquiring the PET in either step-and-shoot or continuous motion mode. During continuous motion, the patient couch travels through the scanner at a chosen speed whilst continuously acquiring the PET data. For step-and-shoot, the PET acquisition is performed as a series of short static acquisitions called 'bed positions' with the couch repositioning the patient between each acquisition as shown in Figure 2.6.

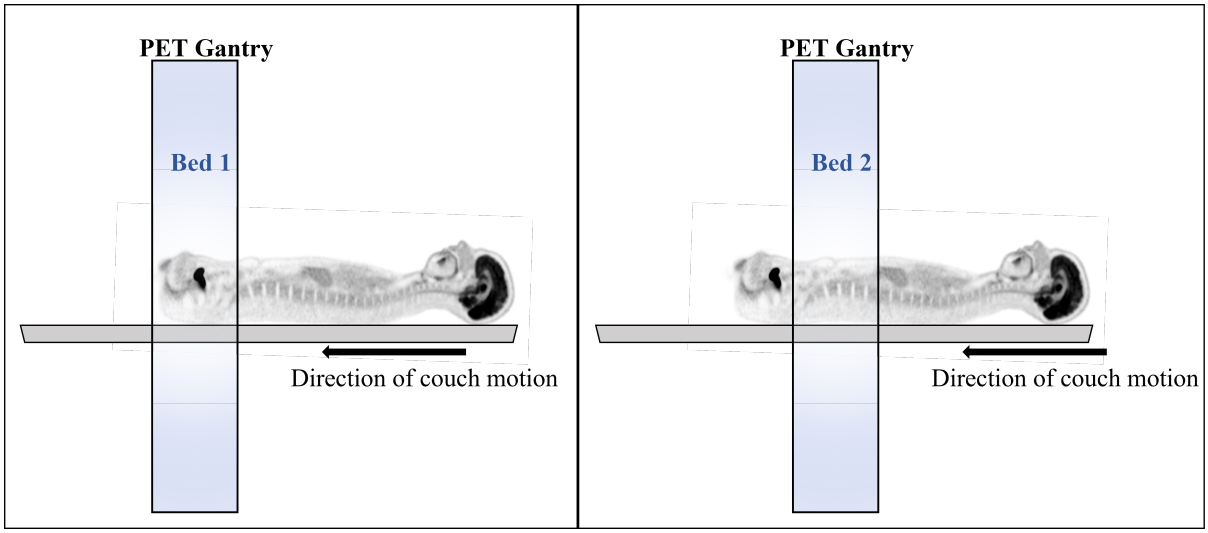

Fig. 2.6 PET scanners only cover a small section of the patient at once, so the patient is moved through the scanner bore acquiring short static scans known as bed positions.

The bed positions are typically overlapped by 23 % to 50 % to compensate for the loss in sensitivity at the axial extents of the FOV. To generate the final PET images, the data from individual bed positions is combined either after reconstruction by using a weighted average of the overlapped regions or by utilising the data from overlapped frames as input in the iterative

loop during reconstruction [38]. The acquisition time for each bed position is chosen depending on the scanner sensitivity, patient size and injected activity [37].

2.1.3.2 2D and 3D PET Systems

PET systems consist of multiple rings of detectors that make up the axial extent of the PET scanner. In older two-dimensional (2D) PET systems, lead or tungsten septa were inserted between the detector rings restricting the acquisition to LORs for detectors within the same detector ring (direct planes) or one or two rings either side (cross planes). Modern clinical PET systems do not have septa and acquire in 3D mode allowing detection of coincidence events from oblique planes as shown in Figure 2.7.

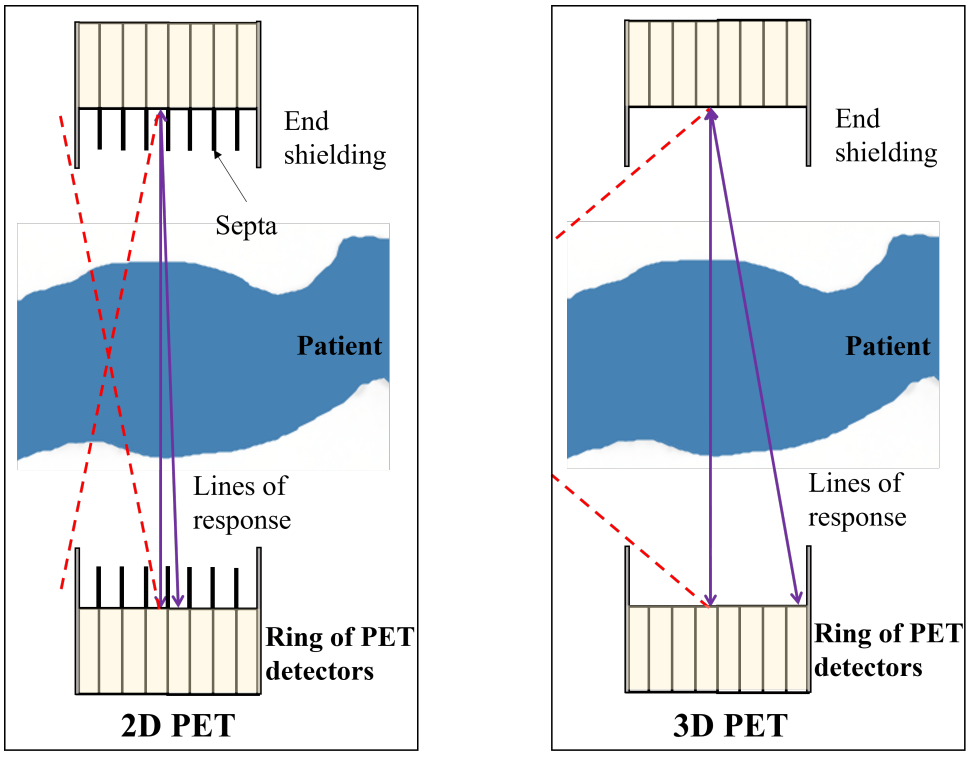

Fig. 2.7 In 2D PET systems (left), septa between detector rings prevent LORs from oblique planes. 3D PET (right) has no septa allowing LORs from oblique planes.

By including LORs from oblique planes, the sensitivity for 3D PET increases significantly compared to 2D but suffers from a larger proportion of random and scattered events due to events originating from outside the FOV (red dotted lines in Figure 2.7). Corrections for out of field

scatter and randoms are incorporated into modern 3D PET reconstructions as discussed later in this section.

2.1.3.3 Partial Volume Effect

The finite spatial resolution of PET systems results in blurring of small objects and spillover of radioactivity between regions in the image. In addition to this, discrete image sampling does not follow anatomical boundaries giving rise to voxel values that are the mean radioactivity concentration from all the tissues contained within the volume [39]. Both these phenomena are collectively known as the partial volume effect (PVE). The PVE results in small lesions $< 3 \times$ FWHM of the reconstructed image resolution appearing to have lower uptake and larger volume than reality. This can result in significant bias in quantitative measurements of uptake and volume.

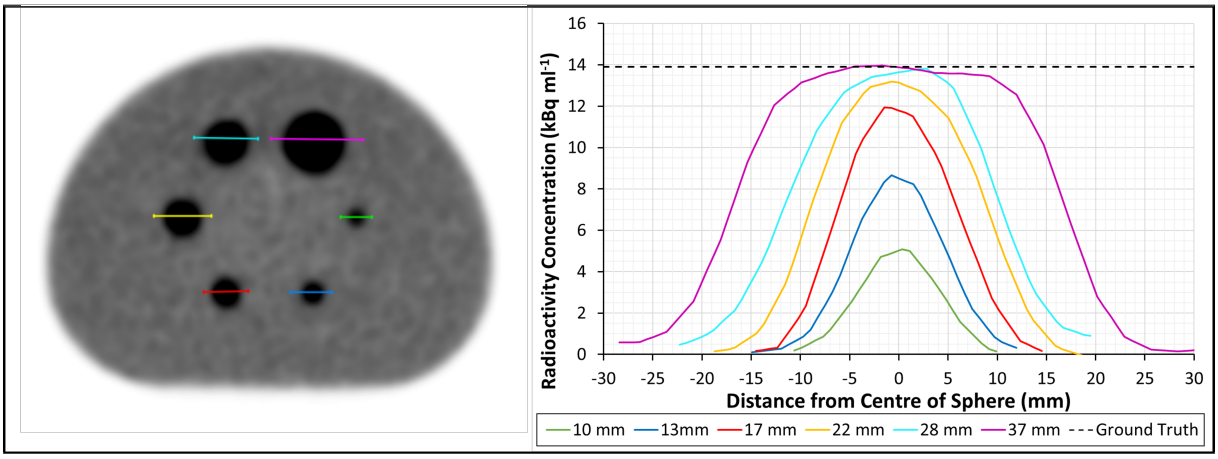

Fig. 2.8 Reconstructed PET image of a phantom containing spheres filled with identical radioactivity concentration. Profiles of the spheres are plotted showing the spill out of counts from the spheres into the background and spill in from the background into the spheres.

The impact of the PVE is dependent on the size of the objects being imaged and their radioactivity concentration relative to the surrounding tissues. Figure 2.8 shows a reconstructed PET image of the NEMA image quality phantom (Data Spectrum, Hillsborough, NC). The phantom contains six spheres, with diameters from 10 mm to 37 mm, filled with a solution of identical radioactivity concentration contained within a uniform background of dimensions as

described in the NEMA standard [40]. Profiles plotted through the spheres on the reconstructed image show that for all sphere sizes the counts are spread into the surrounding region causing blurring at the edges. Additionally, as the spheres get increasingly smaller, the maximum radioactivity concentration becomes increasingly lower than the ground truth. Various methods have been proposed to correct for the PVE, as reviewed by Erlandsson *et al* [41]. However, these methods have limitations and may worsen the accuracy and precision of quantification in small tumours [42]. Improvements in detector design leading to better spatial resolution and reconstruction techniques that incorporate resolution modelling and anatomical priors within the reconstruction would lead to a reduction in PVE and thus more accurate quantification for small lesions < 15-20 mm diameter.

2.1.3.4 Time-of-Flight

If the time of arrival for each annihilation photon from an event could be measured exactly, the difference in the arrival times could be used to derive the exact location of the annihilation event along the LOR. This principle is known as time-of-flight (TOF) PET. In reality, clinical scanners have a finite timing resolution which results in an uncertainty in the measurement of the arrival times. As discussed earlier, improvements in scintillator crystals and electronics have improved the temporal resolution in lutetium-based PET systems with the latest clinical systems achieving 210×10^{-12} seconds [34]. The difference in arrival times and the uncertainty from the timing resolution of the system are used within the reconstruction algorithm to constrain the location of the event to within a few cm along the LOR rather than the event being equally likely to be at any point along the LOR. Use of TOF better localises true events leading to better contrast recovery and lower noise i.e. a gain in signal-to-noise ratio (SNR) [43]. TOF also improves rejection of scattered and random events.

2.1.4 Reconstruction

The basic data output from a PET system consists of a chronological list of positional, timing and energy information for each detected coincidence event. This is known as list-mode data. For TOF systems the time of arrival for the two photons is also included. Since list-mode data can be particularly large, the data is usually binned in real-time into a smaller matrix known as a sinogram. This groups events that fulfil certain energy, positional and timing criteria together but at the expense of information about individual events. The construction of the sinogram is demonstrated in Figure 2.9. For LORs at a given angle of acquisition, θ, coincidence events from all the LORs are binned into a histogram based on the radial distance from the centre of the scanner gantry, r. These histograms, known as projections, are then stacked in the matrix according to the angle of acquisition (0° to 360°). For 3D PET, sinograms containing projections from all oblique planes are also constructed in addition to the projections from direct planes. These sinograms are then used to reconstruct 3D images of the object.

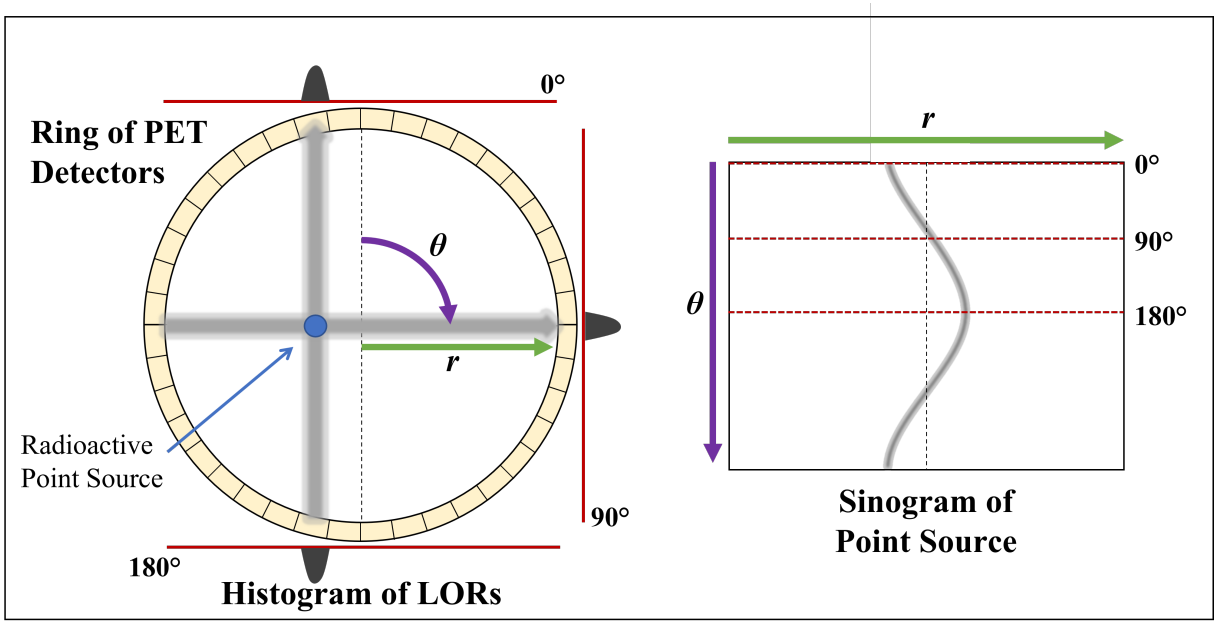

Fig. 2.9 The histogram of LORs from each angle around the radioactive point source (left) are stacked in a sinogram (right) according to the distance from centre of the the scanner field of view r and the acquisition angle θ.

2.1.4.1 Filtered Backprojection

Traditionally data acquired using 2D PET systems were reconstructed using filtered backprojection (FBP) which is an analytic inversion method [44]. As seen previously in Section 2.1.3, the number of coincident annihilation photons detected by two opposed detectors is approximately proportional to the integral of the total radioactivity concentration along the LOR. The simplest method of image reconstruction is backprojection whereby the counts from each projection in the sinogram are 'projected' across the image matrix, with the counts evenly distributed between voxels that fall along the projection path. Counts from all projections are added to generate an image approximating the radiation distribution. However, noise from the projections is also propagated across the image creating a star artefact, therefore to reduce the noise a filter is applied before backprojection. 2D PET data are reconstructed as a series of transaxial slices which are then stacked to produce a 3D dataset. With 3D PET data the situation is complicated by the fact that only a subset of the oblique planes are collected. This is discussed in more detail by Colsher *et al* [45].

Analytic reconstruction techniques, such as FBP, use a simplified imaging model which is easy to implement and fast to compute. However, this assumes PET data is noise-free when in reality there are a number of stochastic uncertainties associated with PET imaging which cannot be modelled using FBP reconstruction techniques thus limiting the accuracy of the final reconstructed images [46].

2.1.4.2 Iterative Reconstruction

Many iterative reconstruction techniques were originally proposed in the '80s and '90s, but due to their complexity, long reconstruction times and limitations in available computational power, many did not translate into clinical practice. With cheaper and faster computational technology and developments in PET scanner hardware and more efficient iterative algorithms, PET has seen the introduction of new and modified iterative reconstruction techniques in clinical PET systems. One key advantage of using iterative reconstruction methods is that they can incorporate

modelling of all aspects of the system response including randoms, scatter and attenuation improving the quantitative accuracy of the reconstructed images.

There are several iterative reconstruction algorithms described in the literature, as reviewed by Qi and Leahy [44], most of which are based on the maximum likelihood-expectation maximisation (ML-EM) algorithm [47, 48]. The ML-EM algorithm starts with an initial guess of the image, usually a uniform image with a constant value. This estimation image is forward projected to generate projections matching the polar co-ordinates of the PET scanner. Correction factors for each projection are derived by comparing the projections from this estimated image to the measured projections. The correction factors are then back projected to generate a correction image. The algorithm then multiplies the original estimation image by this correction image and divides by weighting factors according to the model of the system response. This forms a new estimated image which is used as input for the next iteration. This process repeats until the estimated and measured data converge, with the corresponding image representing the maximum likelihood (ML) solution. ML-EM reconstruction algorithms are generally slow and require a lot of iterations to converge. To combat this, Hudson and Larkin developed an accelerated method of ML-EM, known as ordered subsets-expectation maximisation (OS-EM), that groups the projections into ordered subsets within each iteration [49]. This method speeds up convergence by a factor proportional to the number of subsets. OS-EM is currently the most widely used reconstruction in clinical applications.

Both ML-EM and OS-EM reconstruction algorithms converge to an image that best fits the data, however if the data is noisy, as it is in PET imaging, inevitably the ML solution will also be noisy [50]. One way to control the noise is to incorporate a regularisation function in the algorithm based on prior knowledge of the imaging system or imaged object. Alternatively, since the noise in the projections is amplified on each iteration, the algorithm can be stopped before reaching convergence. Too few iterations however, will bias the result towards the initial guess which is typically a uniform image so some level of noise must be allowed to achieve a reasonable solution. Restricting the number of iterations in this way is widely used in clinical applications. To further suppress noise in the final images, a Gaussian smoothing filter can

be applied to the reconstructed image, but at the cost of reducing the image resolution and increasing quantitative bias. Optimal selection of the number of iterations, subsets and FWHM of the smoothing filter in OS-EM reconstructions is object- and task-dependent. Therefore, on clinical PET systems the user is usually able to select the number of iterations, subsets and the level of post reconstruction smoothing applied allowing them to decide the trade-off between bias and noise for the intended application.

2.1.4.3 Bayesian Penalised Likelihood Reconstruction

Bayesian penalised likelihood or maximum *a posteriori* reconstruction methods include *a priori* information about the imaging system or image formation process in the statistical model [51, 52]. Q.Clear is a proprietary BPL reconstruction algorithm implemented on newer GE PET systems [53]. In this case the algorithm incorporates prior knowledge about the limited spatial resolution of the scanner (point spread function modelling) and a regularising prior that penalises the relative differences between neighbouring voxels [54]. The penalisation is greater for small voxel differences, which are assumed to be noise, than for larger voxel differences which are assumed to be signal. This allows each voxel to be run to convergence without introducing a high level of noise in the images and removes the need to select iterations and subsets or apply a post filter. The user does however have the option to choose the strength (β) of the regularising term relative to the data statistics which controls the image noise. Since it is allowed to run to convergence, the reconstructed images should be closer to the truth and less noisy compared to OS-EM if an appropriate β value is chosen.

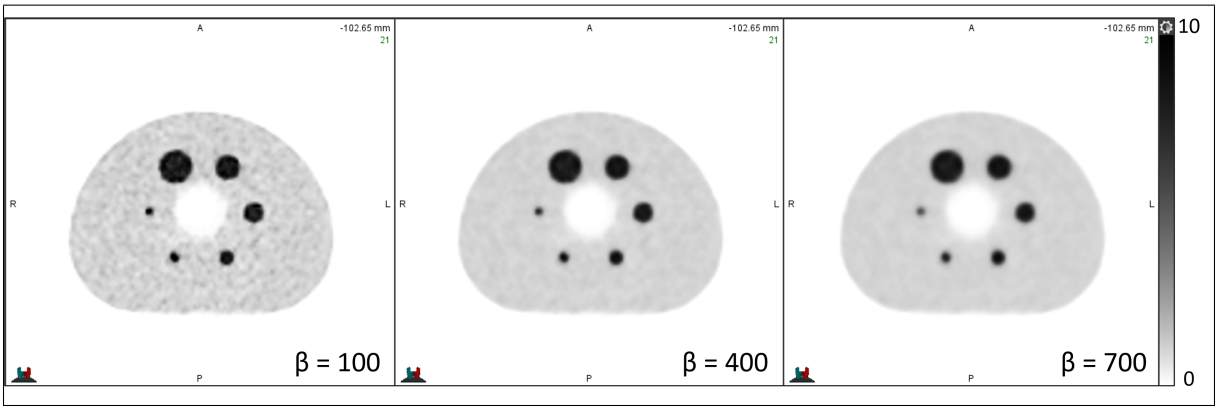

Fig. 2.10 PET images of the NEMA image quality phantom reconstructed using increasing β values.

Reconstructed PET images of the NEMA image quality phantom using Q.Clear with increasing β values are shown in Figure 2.10. As can be seen for a β value of 100, the different sized spheres appear well defined, but the images are noisier. As the regularisation term is increased, the images become increasingly smooth, but with the edges of the spheres becoming less well defined. Measurements of radioactivity concentration for the spheres also change depending on the β value used as demonstrated in the plot of maximum activity concentration versus sphere diameter in Figure 2.11.

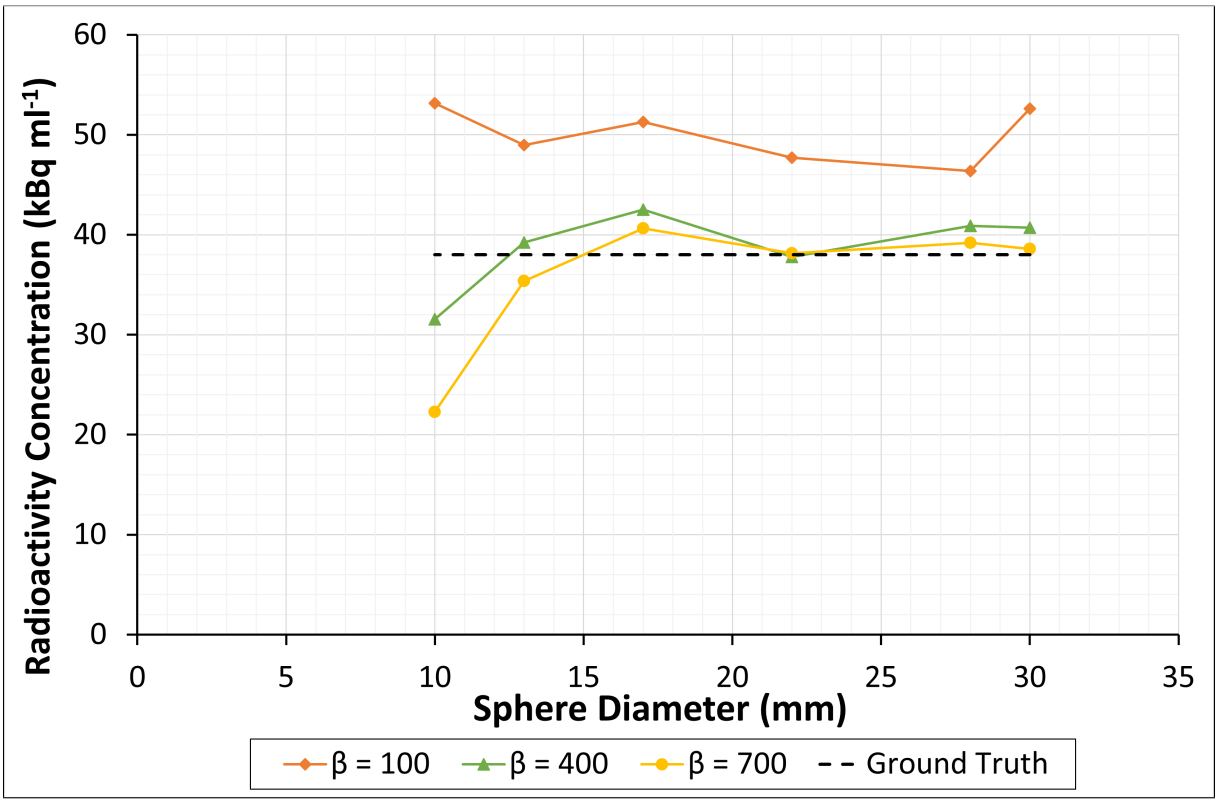

Fig. 2.11 Maximum radioactivity concentration measured for spheres of increasing diameter in PET images reconstructed using Q.Clear with β values of 100, 400 and 700.

For the $\beta = 100$ reconstruction, the maximum voxel values for all sphere sizes are much higher than the ground truth (dotted line). For the $\beta = 400$ and $\beta = 700$ reconstructions, the larger spheres are closer to the ground truth, but as the strength of the regularisation is increased, the resolution of the image is reduced and the radioactivity concentration measured for the smallest spheres is underestimated. A β value of 400 is typically chosen as a compromise between image noise and bias for ^{18}F-FDG-PET in oncology applications [55].

2.1.5 Corrections

To produce more accurate quantification in the reconstructed images, the PET data need to be corrected for decay, dead time, attenuation, scatter, randoms and variations in detector response. These corrections are either applied to the raw PET data as a series of multiplicative factors prior to image reconstruction [46] or they can be included within the system model.

2.1.5.1 Attenuation

As photons from an annihilation event travel through the patient they may interact with atoms within the tissues. At 511 keV, the photoelectric effect is negligible, and the most likely interaction is Compton scattering whereby the photon transfers some of its energy to an outer shell electron ejecting it from the atom. As a result, the photon is scattered at an angle dependent on the amount of energy transferred. If one or both annihilation photons are absorbed or scattered outside the FOV, the coincidence event is not recorded, and the total number of coincidence events detected along the LOR is decreased. This is known as attenuation. The degree of attenuation is dependent on the thickness and atomic number (Z) of the tissue traversed and so photons originating from the centre of the patient are attenuated more than those at the surface.

Before the development of hybrid PET-CT systems, attenuation correction was performed using a transmission source consisting of a line source of ^{68}Ge that rotated around the patient. The attenuation factors were determined from the difference in transmission of the 511 keV photons from the line source with the patient in the FOV compared to the transmission without anything in the FOV. Modern PET-CT systems now use CT-based attenuation correction. The CT data acquired as part of the PET-CT, are by design co-registered to the PET and as such can be used to generate a scan-specific map of linear attenuation coefficients (μ-map) to correct the PET data for attenuation during reconstruction.

Voxel values in CT images use a linear density scale of Hounsfield units (HUs), known as the CT number. The CT numbers for different tissues are computed from the linear attenuation coefficients of the tissues (μ_{tissue}) at CT energies relative to the linear attenuation coefficient of water (μ_w):

$$CT_{number} = \frac{\mu_{tissue} - \mu_w}{\mu_w}.1000$$

To generate the μ-map the CT images are first interpolated to match the PET spatial resolution and then a calibration curve, such as that shown in Figure 2.12, is used to convert the HU for photons at CT energies (40 keV to 140 keV) to linear attenuation coefficients for 511 keV PET

photons [56]. Figure 2.13 shows the μ-map generated for a CT scan of the NEMA image quality phantom.

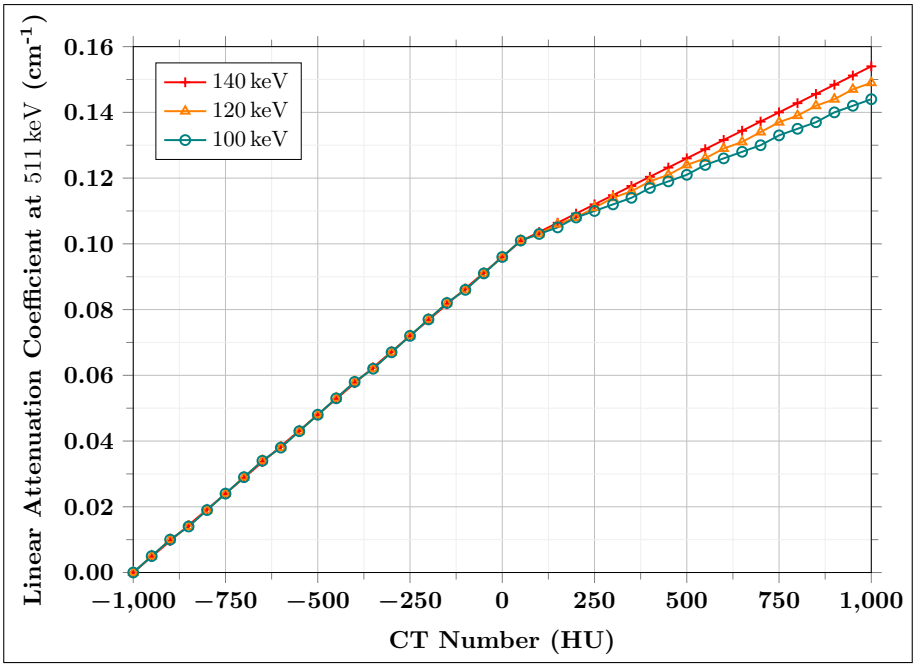

Fig. 2.12 Example of a bi-linear calibration curve used to convert CT numbers in HU acquired at different keV to linear attenuation coefficients at 511 keV.

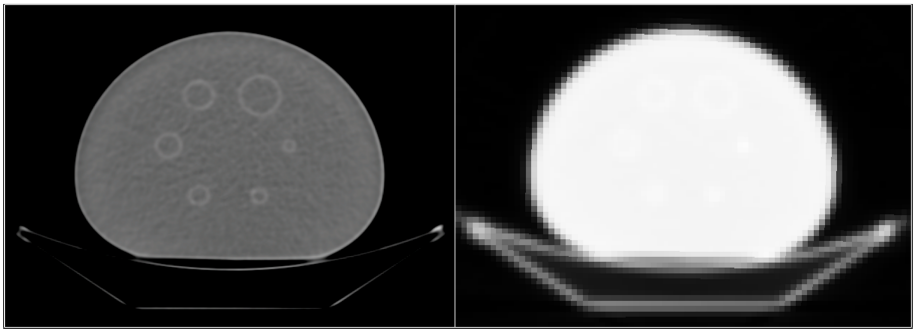

Fig. 2.13 Axial CT slice of the NEMA image quality phantom on the patient bed showing the six spheres (left) and the derived μ-map (right).

The derived μ-maps are forward-projected to generate 3D sinograms matching the PET scanner geometry. The attenuation correction factors (ACFs) for each projection in the sinogram are calculated as the exponential of the integral along the LOR. The projections in the PET sinogram are then multiplied by the corresponding ACF.

2.1.5.2 Normalisation

The response measured for each LOR in a PET scanner varies due to differences in detector geometry across the field of view (FOV) and random and systematic variations in the intrinsic detector efficiencies. The normalisation correction is used to correct for these variations by applying normalisation coefficients to each LOR during reconstruction. The exact procedure for normalisation is vendor specific, but generally involves the acquisition of a 'blank' scan where all detectors are exposed to a uniform radioactive source to measure the response along each LOR [57]. The inclusion of oblique LORs in 3D PET resulted in a much larger number of active LORs compared to 2D PET [58]. As such, direct normalisation using a routine acquisition of a blank scan is impractical due to the long time required to acquire enough counts in each LOR without excessive scatter and dead time effects. To overcome this, component-based normalisation was developed whereby the total count rate variability is divided into two components: the component due to detector geometry and the component due to differences in the intrinsic crystal efficiencies [59]. As the scanner geometry does not change over time, only one high statistics acquisition is required to measure the geometric correction factors for a particular scanner design. The intrinsic crystal efficiencies however do change with time and so the correction factors are calculated by performing a shorter daily scan as part of the routine scanner quality control (QC) schedule [60]. The separate components of the normalisation correction are saved along with the raw PET data acquired during an individual scan so can be stored and used for retrospective reconstruction or processed using offline research tools.

2.1.5.3 Randoms

As previously discussed, the rate of random coincidence events is proportional to the rate of singles events on the detector squared, which in turn is dependent on the activity in the FOV, and the width of the coincidence timing window. The contribution of randoms to the total coincidences can be significant particularly for scans using high activities and 3D acquisition mode. Use of a narrow coincidence timing window helps to minimise the randoms fraction, but

the window must remain wide enough to allow for the time taken for the annihilation photons from true events to reach the detectors and the uncertainty in the coincidence timing [61].

There are two main techniques for correcting for randoms in PET systems. The first being the estimation from the singles rate measured on the detectors [62, 63]. For a detector pair the randoms rate R is:

$$R = 2t_w.C_m.C_n$$

where $2t_w$ is the coincidence timing window width and C_m and C_n are the rate of singles for each channel from the two detectors. Integrating the product of the singles rates over time therefore gives the total number of randoms for that LOR.

The second method is to use a delayed timing window offset from the main coincidence timing window that directly measures the contribution from just the random events. This can then be subtracted from the total coincidences acquired in the main window to leave the true events. The advantage of using this method over estimation from the singles rate is that, unlike the singles events, the coincidences detected in the delayed window are subject to the same dead-time effects in the coincidence timing circuitry as the true events.

2.1.5.4 Scatter

The upper and lower thresholds for the energy window on a PET scanner are generally set quite broad (typically 425 to 650 keV) which means many scattered coincidences are accepted and cannot be easily separated from the true coincidences. This can cause significant degradation in the image quality and quantitative accuracy [61]. Since all attenuation in PET is due to Compton scatter, the CT image can be used to derive an estimation of the scattered events which is then subtracted from the projection data. An alternative method uses the projections from outside the imaged object which, after correction for randoms, should only contain misplaced events due to scatter. The tails in the projections outside the imaged object are then extrapolated to estimate the scatter contribution for the entire projection.

2.1.5.5 Point Spread Function Modelling

As already described in Section 2.1.3.3, there are several physical effects that degrade the resolution of a PET system leading to the PVE. Point spread function (PSF) modelling uses simulations or experimental measurements of a point source at different locations in the FOV to generate a model of the scanner response as a result of these effects for inclusion within the system model. Use of PSF modelling can improve spatial resolution and contrast recovery however, it can modify the structure of noise resulting in a 'lumpy' texture [64].

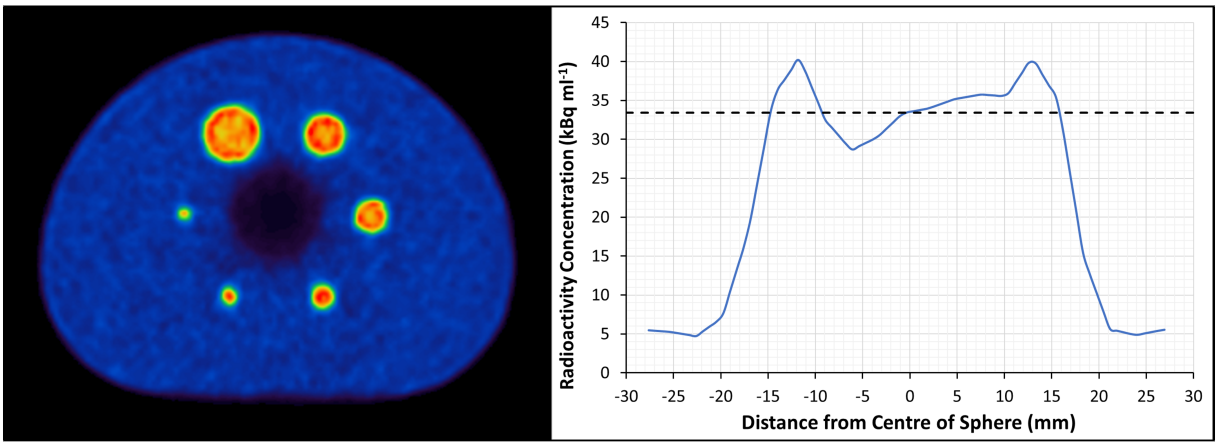

Fig. 2.14 PET image of the NEMA phantom reconstructed with PSF modelling in the system model. A profile of the largest sphere shows the over- then under-shoot of the radioactivity concentration at the edges compared to the ground truth (dotted line).

Additionally, edge and ringing artefacts can also be observed at the boundaries of structures [65]. An example of this can be seen in the PET image of the NEMA phantom in Figure 2.14 that has been reconstructed using PSF modelling. The profile of the radioactivity concentration in the largest sphere (blue line) shows how these artefacts could lead to overestimation of quantitative measurements compared to the ground truth (dotted line) if the maximum value were chosen. Therefore, while inclusion of PSF in the system model has been shown to improve lesion detection [66, 67], it needs to be used with care to avoid misinterpretation of small lesions for monitoring response assessment [68, 69].

2.2 Image Quality Assessment in Medical Imaging

The intrinsic properties of the imaging system introduce noise and blurring in the signal of the imaged object during acquisition and reconstruction. Image quality is a measure of the perceived degradation in the reconstructed images determined by the reviewer. In addition to the properties of the imaging system, the perceived image quality can also be linked to the level of reviewer experience and the viewing conditions [70]. Many objective measures of varying complexity have been used to assess image quality. These metrics do not reflect overall image quality and have differing correlation to perceived image quality by reviewers [71, 72], but can be useful as surrogates for certain features of image quality such as perceived image sharpness, contrast and noise.

In diagnostic imaging there is usually no gold standard or ground truth to compare to. Where histopathology or follow up imaging is available as a reference, it is often subject to selection bias or lacks independence from the imaging procedure under test [19]. Trial designs for technology assessment therefore evaluate the diagnostic accuracy of a new technology compared to existing practice [71, 73]. Very few studies link perceived improvements in image quality due to new technologies to clinical outcomes [74] and statistical significance may not translate to clinical significance [75]. The results will also depend strongly on the intended clinical task and the method of analysis used e.g. lesion detection, measurements of uptake or full kinetic analysis using dynamic imaging.

2.2.1 Technical Image-based Metrics

The performance of new PET technology and methodology, such as a new image reconstruction algorithm, is often measured in terms of technical image-based metrics such as signal-to-noise ratio (SNR), contrast-to-noise ratio (CNR) and noise equivalent count rate (NECR). These are measured using physical phantoms or simulations, where the ground truth is known, and can provide objective measurements for comparison of different systems or for optimising parameters of the imaging system. However, these assessments do not provide a task-specific evaluation, and

do not include many relevant variables that are likely to affect the properties of the final image, including patient-related factors such as body habitus, variation in tracer biodistribution related to medication and/or sub-optimal pre-scan preparation and variation in attenuation properties e.g. as a result of patient movement during the acquisition.

2.2.1.1 Noise Metrics

The standard deviation (SD) of reconstructed voxel values within a region of constant radioactivity concentration is a simple measurement of noise in an image. Since the significance of the SD is dependent on the size of the mean value, the coefficient of variation (COV) normalises the SD to the mean allowing the comparison of variation in datasets with different means:

$$CoV(\%) = \frac{SD}{Mean} \times 100$$

Neither the SD nor CoV fully characterise the image noise, so measurements using multiple ROIs in a single image are used to derive the image roughness (IR) and the background variability (BV) [40, 76]. The IR is the average of the CoVs for multiple ROIs and provides a measure of the voxel-to-voxel variations across an image rather than for a single ROI. The IR is related to the perceived noise in a single image [76]. BV is calculated by dividing the standard deviation in mean values for all the ROIs, by the average of the mean values for all the ROIs. The BV is therefore a measure of the region-to-region variation across the image which is useful in the detection of non-uniformities.

Both IR and BV are measures of noise that are applicable to a single image, or realisation. If multiple realisations are available, additional measures of noise can be made that more accurately represent uncertainties in the images. Ensemble noise (EN) is the CoV in the mean for ROIs across independent realisations and is inversely proportional to detection task performance [76]. Like BV, the EN provides a measure of the variation across regions in the image except across multiple realisations. The standard deviation image can also be derived by calculating the SD across realisations for each voxel. Often multiple realisations are not available for analysis

making BV and IR more widely reported for assessment of regional and voxel-wise variations respectively. Since there is a linear correlation between BV and EN, BV can be used as a surrogate for EN to quantify spatial variation where multiple realisations are not available [76].

2.2.1.2 Signal Metrics

The signal-to-noise ratio (SNR) is the ratio of the signal intensity in the imaged object (S_{object}) to the standard deviation (noise) in the background region (SD_{bgd}):

$$SNR = \frac{S_{object}}{SD_{bgd}}$$

Misplaced events due to scattered and random coincidences result in noise in the reconstructed image. Therefore for a fixed signal intensity, the SNR decreases as the number of detected scatter and random coincidences increases. The SNR doesn't account for the signal intensity in the background (S_{bgd}), therefore the contrast-to-noise ratio (CNR) is defined as the difference between the signal intensity in the object and the background (contrast) divided by the standard deviation in the background region:

$$CNR = \frac{S_{object} - S_{bgd}}{SD_{bgd}}$$

The noise in a PET image is related to the total number of detected coincidences which is made up of true, scattered and random events. The noise equivalent count rate (NECR) incorporates the effect of noise due to subtraction of the randoms and scatter contributions and provides a direct link to the image SNR and the coincidence count rates [77]. The NECR is calculated as:

$$NECR = \frac{T^2}{(T + S + kR)}$$

where T, S and R are the count rates for true, scattered and random coincidences, and k is equal to 1 for randoms correction from singles or 2 for randoms correction using a delayed coincidence window to account for the noise component. The NECR is defined as the minimum number

of coincident events required to obtain a statistically equivalent noise level in the presence of measured scatter and random events if all incident events were considered as true events. NECR is usually measured using phantoms as described in the NEMA standard [40], but an NECR derived from patient data has been shown to have moderate correlation with clinically perceived image quality [78].

2.2.1.3 Recovery Coefficients

The recovery coefficient (RC) is a widely used index of system performance in PET that is measured using phantoms, such as the NEMA image quality phantom, that contain spheres of different sizes filled with known radioactivity. The RC for a sphere is calculated as the ratio of the radioactivity concentration measured in the image ($C_{measured}$) to the true concentration in the sphere (C_{true}) [79]:

$$RC = \frac{C_{measured}}{C_{true}}$$

RCs are measured for all the sphere sizes and plotted against sphere diameter to generate a recovery curve. The maximum voxel value is most often used for determination of $C_{measured}$ as it is closest to the true value. Additionally, RCs are frequently calculated calculated using the mean activity concentration for a VOI delineated using a relative threshold of the maximum voxel value in the sphere. A threshold of 50 % corrected for local background is often chosen as it is theoretically closest to the real volume and has been shown to have the best balance for successfully defining the lesion edges for different tumour-to-background ratios and for repeatability of volume measurements in phantom and patient test-retest experiments [80]. The RCs depend on the system PSF, the size of the imaged object and noise, particularly when using measurements based on the maximum voxel value [81].

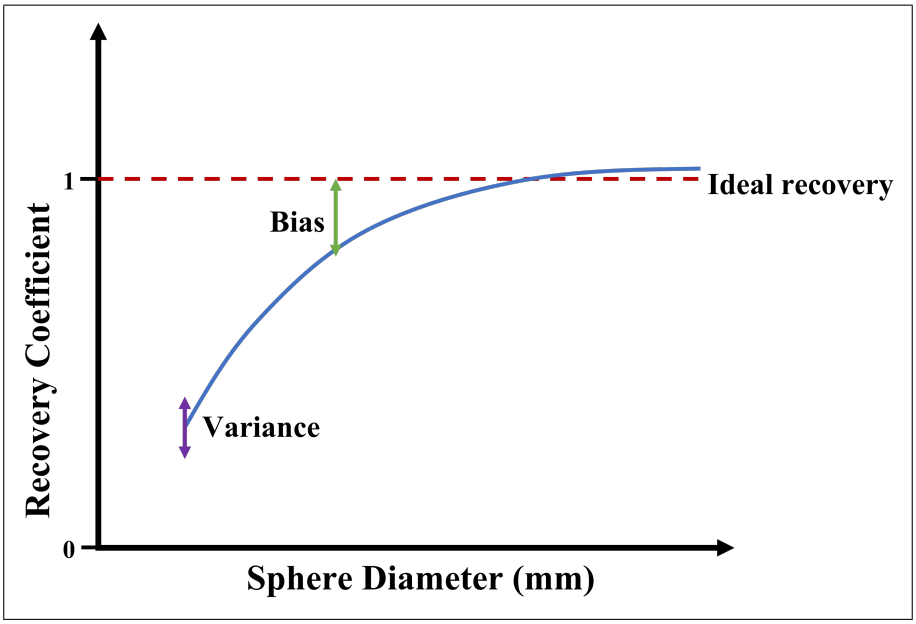

Fig. 2.15 Schematic plot of a recovery curve for an idealised PET system (red dotted line) and a typical PET system without PSF modelling (blue line).

For an ideal system where the true radioactivity concentration is fully recovered, the RCs would be 1 regardless of object size as shown in Figure 2.15 (red dotted line). The blue line demonstrates the typical shape of the recovery curve using the maximum voxel values for a real PET scanner. As previously discussed, the PVE causes the recovery of the radioactivity concentration in the spheres to become increasingly poor as the sphere size decreases $< 3 \times$ FWHM, which is equivalent to around 18 mm to 20 mm for clinical PET systems. The bias is the percentage difference between the measured RCs and the ideal recovery of 1.

$$Bias\ (\%) = \left(\frac{C_{measured}}{C_{true}} - 1 \right) \times 100$$

The variance in the RCs is calculated as the sample standard deviation for repeated measurements and may be normalised to the mean to give the coefficient of variation (CoV). Recovery curves are useful in the evaluation of new technology as they can characterise the response of PET scanners under standard conditions.

2.2.2 Task-based Image Assessment

In clinical practice, PET is usually performed for specific diagnostic tasks such as lesion detection, quantification at a single time point or measuring response to therapy from a series of measurements at different times. These tasks are broadly divided into two types – classification and estimation tasks. Classification tasks would include lesion detection and/or localisation whilst estimation tasks include measurements of parameters that help to characterise the lesion such as degree of radiopharmaceutical uptake or volume [22]. The characterisation of task-specific performance indices in terms of bias and variability is in general very hard to determine as the ground truth is often unknown and is not necessarily related in a simple way to the basic image-based indices. Often there is also a trade-off between bias and variance depending on the intended application.

2.2.2.1 Classification Tasks

In medical imaging, classification tasks extract features from the images and use published models or clinician experience to categorise patients to aid in diagnosis or management. Examples of binary classification tasks in oncology include lesion present vs lesion absent or malignant vs benign. In an imaging context, measures could be objective measurements based on a characteristic of the tumour such as volume or uptake or could be subjective based on the observers confidence rating.

A confusion matrix is a useful tool for determining the predictive quality of a diagnostic test in performing a classification task using a chosen fixed threshold value. For example, to assess how well a new diagnostic test can determine whether a person has a disease or not, predictions from the diagnostic test are compared to the reference standard (this may be the ground truth or an existing gold standard test). This gives four possible combinations for the outcome of the diagnostic test as shown in the confusion matrix in Figure 2.16:

- True positive (TP) - the test correctly predicts the person has the disease.
- False negative (FN) - the test incorrectly predicts the person does not have the disease.

- False positive (FP) - the test incorrectly predicts the person has the disease.
- True negative (TN) - the test correctly predicts the person does not have the disease.

		Diagnostic Test	
		Positive	**Negative**
Reference Standard	**Positive**	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Fig. 2.16 Binary confusion matrix showing the classification of the two groups according to the diagnostic test versus the actual condition from the reference standard.

From this data, the sensitivity and specificity of the diagnostic test can be calculated. The sensitivity or true positive rate (TPR) is the fraction of people who have the disease that were correctly identified as positive by the diagnostic test:

$$TPR = \frac{TP}{TP+FN}$$

A high sensitivity indicates the test predictions result in a small number of false negatives. The specificity or true negative rate (TNR) is the fraction of people without the disease that were correctly identified as negative by the diagnostic test:

$$TNR = \frac{TN}{TN+FP}$$

A high specificity indicates the test predictions result in a small number of false positives. In addition to the TPR and TNR, the false positive rate (FPR) and false negative rate (FNR) can also be calculated as follows:

$$FPR = (1 - TNR)$$

$$FNR = (1 - TPR)$$

where the FPR is the fraction of people without the disease that were incorrectly identified as having the disease and the FNR is the fraction of people with the disease that were incorrectly identified as being negative. Ideally a diagnostic test should have both a high sensitivity and high specificity to ensure people with the disease are correctly identified whilst maintaining a low false positive rate. The trade-off between sensitivity and specificity for a particular diagnostic test is dependent on the severity of the disease and options for patient management, for example the clinician will need to weigh the consequences of missing a patient with the disease who then does not get treatment versus the consequences of a false positive leading to treating a patient without the disease.

The sensitivity and specificity are useful metrics in the context of comparing the performance of the diagnostic test to the reference standard, however they do not give an indication of the likelihood that the diagnostic test can successfully identify whether people do or do not have the disease, based on their test results [82]. Therefore, it is helpful to also provide the positive and negative predictive values where the positive predictive values (PPV) is the probability that people with a positive diagnostic test result actually do have the disease:

$$PPV = \frac{TP}{TP+FP}$$

and the negative predictive values (NPV) is the probability that people with a negative diagnostic test result actually do not have the disease:

$$NPV = \frac{TN}{TN+FN}$$

Receiver operating characteristic (ROC) curve analysis is a statistical tool commonly used for characterising diagnostic accuracy in medical imaging [83]. ROC curve analysis is used to determine the ability of the diagnostic test to perform binary classification tasks compared to a reference standard such as histology or long term follow up and enables the overall performance of tests to be compared, independently of specific threshold values. The first stage in ROC

analysis requires determination of the sensitivity and specificity of the diagnostic test for a given threshold. The TPR is then plotted as a function of the FPR as the threshold is varied as shown in Figure 2.17.

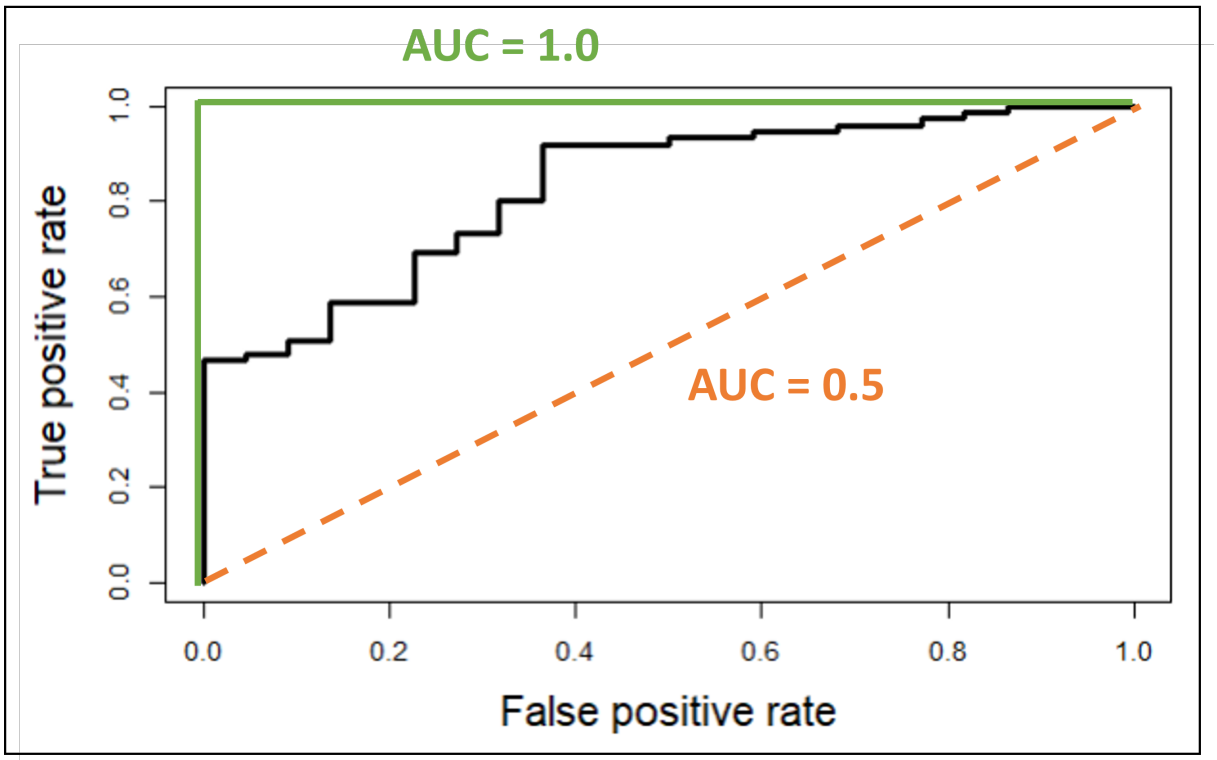

Fig. 2.17 Example of a receiver operating characteristic curve.

ROC curve analysis can be used to compare thresholds for distinguishing between positive and negative cases in the classification task or for comparing performance of different diagnostic tests for the same classification task. The advantage of the ROC curve is that it demonstrates the trade-off in TPR and FPR which can help in choosing the threshold or test most suitable for the clinical scenario. The figure of merit for an ROC curve is the area under the ROC curve (AUC). With reference to Figure 2.17, the orange dashed line would give an AUC of 0.5 which would mean the diagnostic test is not able to distinguish between the two groups and is no better than random guessing. The green line with an AUC of 1.0 demonstrates perfect discrimination between the two groups i.e. 100 % TPR and 0 % FPR. A typical ROC curve for a diagnostic test

would fall somewhere between these two (black line), the higher the AUC value the better the test performs for a given task.

Since the AUC assesses the performance across all thresholds rather than just clinically relevant thresholds and it doesn't provide information on losses or gains for individual patients, interpretation in a clinical context can be difficult. Alternatively, net benefit analysis can be used for determining the clinical value of a test [84]. Net benefit analysis incorporates disease prevalence and weighting for the consequences of a false positive or false negative result thus providing a more representative measure of the clinical usefulness of including or excluding a diagnostic test. The disease prevalence is relatively straight-forward to determine, however the assessment of the clinical consequences, such as the benefit of early detection or the harm of unnecessary treatment, can be subjective. To overcome this, either expert consensus can be used to determine the weighting used, or the net benefit can be plotted for a range of appropriate weights to derive a decision curve [85].

2.2.2.2 Estimation Tasks

Estimation tasks involve the measurement of parameters that describe certain characteristics of the imaged object such its size, density or uptake. Quantitative measurements of metabolic uptake in ^{18}F-FDG PET images can provide additional information for staging, response assessment, detection of recurrence or surveillance in oncology applications.

2.2.3 Quantification in PET

There are several metrics used to quantify PET uptake in clinical and research applications. The standardised uptake value (SUV) is a simple semi-quantitative measure that normalises the measured radioactivity concentration to the injected activity and distribution volume (DV). If the radiopharamaceutical were distributed evenly throughout the volume, this would result in an SUV of 1 everywhere. Use of SUV allows the comparison of uptake across different patients or for different time points within the same patient. Body weight is most often used as a surrogate

for DV in clinical practice [86] and will be used throughout this thesis, but lean body mass (LBM) is also recommended as ^{18}F-FDG doesn't accumulate in fat tissues [87]. The SUV is unitless assuming that 1.0 g of tissue is equivalent to 1.0 ml:

$$SUV = \frac{C_t\,(kBq.ml^{-1})}{A_t\,(MBq)/W\,(kg)}$$

where C_t is the concentration of radioactivity in tissue at time t, A_t is the injected activity decay corrected to time t and W is the patient weight measured on the day of scanning.

Reporting of the maximum standardised uptake value (SUV_{max}) is recommended for staging and response assessment in oncological PET [37]. This takes the maximum SUV in a single voxel within an ROI. The SUV_{max} has gained popularity in clinical use as it is easy to measure and highly reproducible between observers within a single scan [88]. However, SUV_{max} is highly susceptible to changes in the image noise [89] or selection of reconstruction parameters [90].

More recently the SUV_{peak} was defined and has been recommended for reporting solid tumours [2]. This uses a sphere with a diameter of approximately 1.2 cm to produce a 1.0 cm^3 volume spherical VOI centred around the hottest region in the tumour (not necessarily the hottest voxel). One issue with using this definition for SUV_{peak} is that some tumours, particularly after treatment, are often smaller than the 1.0 cm^3 volume defined for the SUV_{peak} making it difficult to assess treatment response [91]. Another potential downside for both SUV_{max} and SUV_{peak} is that for larger lesions both are limited to a small part of the tumour and information about the whole tumour volume is not utilised. For example, the SUV_{max} or SUV_{peak} alone cannot differentiate between a large tumour with a small region of high uptake from a small tumour with high uptake. It could however be argued that the region with the greatest ^{18}F-FDG uptake, and thus the highest metabolic activity, represents the most aggressive part of the tumour which may be most indicative of prognosis or treatment response. A larger region could also include non-tumour tissue, such as necrotic tumour or stomal tissue, and the average SUV for the whole region rather than the maximal value will not reflect the actual tumour behaviour [92].

Other potential measures include SUV_{mean}, metabolic tumour volume (MTV) and total lesion glycolysis (TLG) which is the product of the SUV_{mean} and MTV. These measures all require manual or automated segmentation of the tumour and can provide additional information about the metabolism for the whole lesion which have been shown to have predictive or prognostic value [93, 94]. Segmentation of PET images is challenging due to the relatively poor image resolution, noise and the large variation in pathologies [95]. To overcome these challenges, several automated segmentation algorithms of varying complexity have been developed as summarised in AAPM Report 221 [96]. The most widely used techniques however are relatively simple and involve the use of a fixed threshold whereby all voxels above either an absolute value or a percentage of the maximum voxel value are included in the volume. Thresholds of 41 % and 50 % are recommended based on phantom measurements [37], however optimal thresholds are dependent on the image characteristics and will tend to overestimate volumes for small lesions [95]. Figure 2.18 shows the difference in volumes for a lung lesion segmented using thresholds of 41 % and 50 % of the maximum voxel value and an advanced gradient-based algorithm (PETedge, Medical Image Merge (MIM) Encore v7.1.5, MIM Software Inc, Cleveland, OH). The location of the SUV_{max} (green cross) and the SUV_{peak} volume (purple sphere) are also shown.

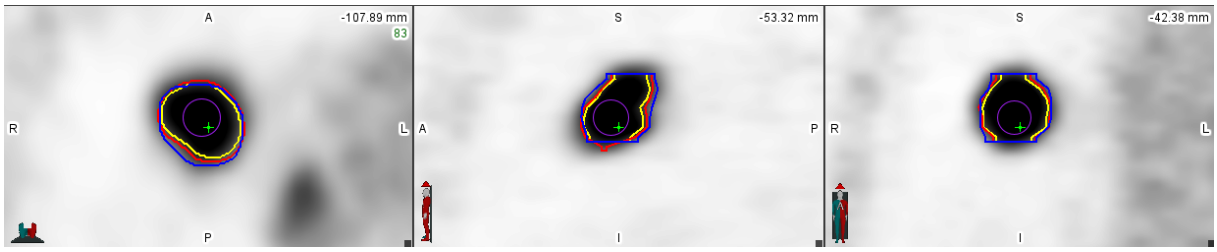

Fig. 2.18 Axial, coronal and sagittal slices showing a lung lesion segmented using the gradient-based algorithm (blue), 41 % threshold (red) and 50 % threshold (yellow). The location of the SUV_{max} is denoted by the green cross and the SUV_{peak} volume by the purple sphere.

The quantitative values derived from these different volumes are also highly variable. As seen in Table 2.2, the SUV_{mean}, MTV and TLG values for the different volumes have differences up to 15 %, 53 % and 30 % respectively.

	Segmentation Method		
	41 % Threshold	**50 % Threshold**	**Gradient Based**
SUV_{max}	14.3	14.3	14.3
SUV_{peak}	12.2	12.2	12.2
SUV_{mean}	9.2	10.0	8.4
MTV (cm^3)	7.3	5.7	8.7
TLG (cm^3)	66.4	56.4	73.2

Table 2.2 Quantitative values measured for the lung lesion in Figure 2.18 using different segmentation techniques.

2.2.4 Phantoms for Image Quality Assessment

Physical and computational phantoms can be used for optimisation of the imaging process where imaging real subjects is not practical or ethical due to the radiation exposure. In PET applications, physical phantoms often consist of plastic containers of varying size and shape filled with a radioactive solution. Standardised phantom designs are useful as reference objects and provide a means to compare the performance of a particular PET scanner against other systems or published values [97] and have been used successfully to harmonise imaging protocols for multicentre clinical trials [98].

To overcome some of the limitations of physical phantoms, computational phantoms have been designed that can be used with Monte Carlo techniques to model the physics of radiation interactions within the body to produce simulated data for dosimetry and imaging applications. This allows a range of variables to be investigated where the exact anatomy and activity concentration in the organs and tissues can be defined. Computational phantoms consist of virtual 3D anatomical models, the complexity of which have evolved significantly over the years [99, 100]. The latest generation of computational phantoms provide complex anatomical geometries including detail about tissue densities providing more accurate simulations of radiation transport

in the body. Compared to physical phantoms, computational phantoms provide greater versatility, efficiency, precision and safety [100].

The 3D model in a computational phantom can either be mathematically defined or based on segmented tomographic image data from real humans or objects (voxelised phantoms). Voxelised phantoms, such as the International Commission on Radiological Protection (ICRP) reference male and female voxelised phantoms [101], are anatomically accurate as they are based on real CT, MRI and cross-sectional cryosection images, but only represent one specific patient's anatomy. It is difficult to model anatomical variations or motion in voxelised phantoms [102] and the phantoms are limited to the resolution of the original dataset used for segmentation limiting their accuracy.

Alternatively, mathematical models have been generated using computer-aided design (CAD) to define the anatomical shapes from equations or geometric primitives. There are two mathematical methods of CAD for generating the surfaces of organs and tissues in computational phantoms [100]. The first method uses constructive solid geometry (CSG) and combines simple shapes (cylinders, spheres etc.) that can be described using quadratic equations. CSG based phantoms such as the MIRD-5 phantom (Medical Internal Radiation Dose Committee Pamphlet No.5) are easy to set up and, where the modelled geometry is relatively simple, can provide good results for dosimetry applications. However, for the human body this results in simplistic stylised phantoms that are not very accurate for simulating imaging data.

The latest generation of computerised phantoms use boundary representation (BREP) to define the organ and tissue surfaces. BREP methods use topological and geometric information providing the relationships between vertices, edges and faces as well as orientation. The surfaces of the organs and tissues are defined as non-uniform rational B-spline (NURBS) with very smooth faces or as a polygon mesh. This allows BREP-based structures to be much more flexible and can describe much more complex anatomical structures. The 4D extended cardiac-torso (XCAT) phantom uses BREP based CAD and the exterior of the object can be defined as NURBS [102]. It is classed as a hybrid computational phantom as it is based on segmented human data from the Visible Human Project (https://www.nlm.nih.gov/research/visible/visible_human.html) but

uses NURBS to define each object providing a more anatomically realistic phantom that can be easily manipulated for a range of applications in biomedical imaging research [103]. The XCAT phantom can therefore model anatomical variations and motion to generate a range of detailed PET images. The range of anatomical variation however is limited as each new model is derived from a template adult phantom based on the Visible Human data. The organs and tissues are also modelled as being homogeneous which does not reflect the true tissue structure.

2.2.5 Monte Carlo Modelling

There are several software packages available for performing PET simulations, ranging from more complex particle-tracking Monte Carlo based methods to simpler dedicated analytic methods as reviewed by Buvat and Lazaro [104]. The choice of software depends on the level of accuracy required versus the available computational resources and the time to create the number of datasets required [105].

Monte Carlo modelling is widely used in nuclear medicine research and there are several Monte Carlo simulation platforms available [106]. For geometry and tracking (Geant4) is a well validated and maintained simulation toolkit developed for modelling of the passage of particles through matter [107, 108, 109]. The Geant4 application for tomographic emission (GATE) simulation platform [110, 111] is open-source software that incorporates the libraries from Geant4 and is designed specifically to run simulations of SPECT and PET systems. GATE has been widely used and previously validated for a number of clinical PET systems including the GE Advance/LS PET scanners, Siemens ECAT and Philips Allegro [112, 113, 114]. Another advantage of GATE is it allows the use of analytical phantoms defined in a macro as a series of geometric shapes with assigned materials, or it can read in image data files, such as patient CTs, to use as voxelised phantoms [115].

Use of well validated physical models along with detailed geometry descriptions of the PET system produces accurate results, however simulations in GATE require a lot of computational resources and time particularly where high activities and large distributions are simulated. One

option to reduce the computing time is to make the simulation parallelisable allowing the simulation to run on multiple processors at once and then merging the results. This can be run on a desktop PC with multiple processors or distributed on a computer cluster or Grid infrastructure where available [116, 117, 118].

The method proposed in this thesis to further reduce the resources and time required to generate many PET images is to perform Monte Carlo simulations of lesions using a validated model of the clinical scanner which are then inserted into PET data from real patients. Use of accurate physics-based approaches for the lesion insertion in projection space should result in images with accurate contrast, blur and noise characteristics making the simulated lesions indistinguishable from real lesions. This methodology will create PET datasets with the anatomical and physiological variation from real patients containing simulated lesions where the characteristics can be controlled. Once validated, these datasets will serve as an alternative to using real patients in clinical evaluations of new PET technology in virtual imaging trials (VITs). This method requires validation for a particular imaging system, but then can be used to rapidly simulate new geometries or different tumour types for assessment.

2.3 Virtual Imaging Trials

Modelling and simulation of biological systems and investigation of treatment effects in a virtual clinical trial (VCT) framework has become an established part of the drug development process and is a key factor in improving efficacy [119]. To ensure PET is being used appropriately in clinical trials and clinical practice, we require new and innovative ways to evaluate developments in PET technology or methodology. One way to address this is through the use of virtual imaging trials (VITs), also referred to as *in silico* imaging trials [120, 121]. For VITs, computational phantoms are used to generate a virtual population of patients. Models of the imaging system are used to simulate the imaging process and produce images of the population where the ground truth is known. This provides much more flexibility than traditional clinical trials, as the same population can be used to investigate multiple aspects of the imaging system, or the same

imaging system can be used to simulate images of different populations. It is also much more cost effective and doesn't require real patients thus avoiding any associated radiation exposure. The data from the VIT can then be assessed using observer studies or quantitatively using task-based measures.

Results from VITs can be used in real clinical trials, either prospectively to inform the study design through identifying sources of uncertainty, determining optimal imaging parameters or for identifying aspects of the imaging that need to be standardised and, in some cases, if they should be included in the exclusion criteria. Alternatively, the data could be used for informing the data analysis prospectively or retrospectively by identifying appropriate task-based measures that are sensitive enough to detect the effect under study or by determining whether to include PET data that does not conform to the original study design.

VITs for breast screening applications are well established and there are two notable projects that have developed and validated *in silico* tools for performing virtual clinical trials: the VICTRE (Virtual Clinical Trial for Regulatory Evaluation) project [122, 123] and the OPTIMAM VCT Toolbox [124]. Both project teams have developed a range of tools for performing *in silico* imaging trials as part of a computational pipeline to compare digital mammography and breast tomosynthesis. The key task in mammography is lesion detection and with the recent introduction of digital breast tomosynthesis, detection rates have increased, however this comes at the cost of doubling the radiation dose which is associated with an increased lifetime risk of cancer induction [125]. VITs using these tools therefore are focused on optimization of radiation dose and lesion detection. Both projects have similar principles whereby mathematical models of breast tissue and breast masses are used to create voxelised phantoms. 2D and 3D mammography images are then generated using ray-tracing to create projections of the breast phantom and using Monte Carlo modelling to model the scatter-to-primary ratio. Noise and resolution characteristics are then applied to match to a specific imaging system and x-ray parameters. Assessments of lesion detection used human and model observers.

In CT and photon counting CT applications, Abadi *et al* have performed several VITs using simulated acquisitions of the XCAT phantom to investigate the impact of image acquisition

parameters on image quality [126, 127]. The authors used a validated scanner-specific model that incorporates ray-tracing and Monte Carlo techniques to model the primary and scattered signals. Assessment of the image quality consisted of technical image-based quality metrics and the bias and variability in radiomics features for task-based assessment.

There are only a few published VITs in PET imaging and all have investigated the impact of technical or biological factors on variability in quantitative measures or lesion detectability rather than assessment of new technology. Harrison *et al* performed a virtual imaging trial to assess the impact of variability in SUV on response assessment in longitudinal studies [25]. The PET analytic simulation package (ASIM) [128, 129] was used to generate simulated images of an anthropomorphic test object containing a spherical lesion approximating a breast lesion. ASIM is an open-source package that generates projection data with defined noise and resolution properties rather than fully modelling the physical processes in the imaged object and acquisition hardware. This makes it much faster to produce a large number of sinograms using different parameters and with multiple realisations than Monte Carlo packages such as GATE. However, this comes at the expense of accuracy. Sinograms for a range of different tumour to background ratios, tumour sizes and noise levels were generated and reconstructed using the open-source software for tomographic image reconstruction (STIR) package [130]. The authors investigated the impact of technical factors, patient preparation, acquisition and reconstruction parameters and analysis methods on the bias and variability of SUV_{max} and SUV_{mean} measurements. ROC curve analysis was used to assess the impact of the variability on detecting a true change in SUV. The anthropomorphic phantom used in the study was of a fixed size and didn't contain realistic heterogeneous uptake or realistic tumour shapes which limits the applicability of the results. TOF-PET and advanced reconstructions with PSF modelling were also not investigated as part of the study as these were not implemented in the open-source software at the time.

Kurland *et al* used a VIT framework to investigate the effect of uptake time on the required sample size for multicentre PET trials measuring metabolic response using SUV [28]. Rather than generation of simulated images, this study used simulated time activity curves using kinetic parameters from a real population of patients with locally advanced breast cancer. The authors

defined the sensitivity based on a 40 % decrease in SUV as a strong response to chemotherapy and specificity as the probability of observing a change < 30 % given no true change in tumour metabolism. The sensitivity was then used to determine the required sample size for different distributions of uptake times compared to one with strict adherence to 60 minutes. Wangerin *et al* also performed VITs using data from a real population of patients with breast cancer to define the ground truth from kinetic modelling [26, 27]. Synthetic time activity curves were generated to include biological variation and differences in uptake time. To replicate the noise characteristics, multiple sinograms were generated in ASIM using the NEMA image quality phantom. ROC curve analysis was performed using measurements of SUV and K_i. To investigate lesion detectability a model observer was used. All these studies were based on a particular tumour type meaning results may not be applicable to other tumour types. The PET images were of the NEMA image quality phantom which again was of a fixed size and didn't contain realistic heterogeneous uptake or realistic tumour shapes.

The purpose of the work in this thesis is to generate PET images with the physiological uptake from real patients and realistic lesions simulated using Monte Carlo modelling. Once validated, these datasets will then be used in place of real clinical subjects to assess the impact of TOF-PET and advanced reconstructions with PSF modelling on quantitative measures in a virtual imaging trial.

Chapter 3

Design and Validation of Monte Carlo Simulations of the GE Discovery 710 and Simulated Lesion Insertion Technique

3.1 Introduction

Virtual imaging trials have been presented as a way to investigate the impact of advances in medical imaging technology in place of clinical trials which can be costly, and often unethical where radiation is concerned [131]. This approach requires a method of generating synthetic images that are representative of the patient population and provide meaningful figures of merit for the intended task [132]. The main issues reported with use of synthetic images are that they are either inaccurate or there is a large computational burden [133], therefore the methodology for generating PET datasets in this work needs to address both these points.

Several methods for generating populations of PET data have been published. The use of computational anthropomorphic phantoms with realistic anatomy such as the XCAT [102] or Zubal [134] phantoms have been combined with Monte Carlo modelling to simulate whole PET scans incorporating lesions [135, 136]. To generate more realistic variability in the PET images, the computational phantoms can be adjusted to patient-specific anatomy and uptake [137]. Use of

well validated Monte Carlo simulations in this way produces PET data based on accurate physics of the acquisition system and the ground truth is known from the input simulation parameters. However, as simulation time is proportional to the amount of activity being simulated, use of Monte Carlo techniques to generate entire patients or phantoms is slow and computationally demanding making it an impractical option to generate large datasets. Analytical simulation tools such as SMART-PET [138] or PETSTEP [139] can be used for generating PET images of entire patients or phantoms in a fraction of the time of Monte Carlo simulations. These tools take an image representing the 'true' activity distribution, smooth it according to the PSF of the PET system and forward project to generate sinograms which are then reconstructed into PET images. To generate patient images, the user assigns the physiological uptake in the 'true' activity images based on the segmented organs and tissues which results in unrealistic homogeneous physiological uptake. The user must also define the noise level and the scatter and randoms fractions which requires advance knowledge of the response of the imaging system to different imaging conditions. The contribution of scatter, attenuation and randoms will vary depending on the activity distribution, and the size and shape of the imaged objects or patients, therefore generalisations are made which causes a loss of accuracy in the reconstructed images.

As discussed in Chapter 2 Section 2.2, image quality and quantitative accuracy are dependent on the activity, size and shape of the lesions as well as the physiological uptake in surrounding tissues and patient habitus. The technique proposed in this work combines the accuracy of Monte Carlo modelling with the physiological uptake and habitus of real patients by inserting simulated lesions into real patient scans in the projection-domain. Compared to insertion in the image-domain, projection-domain insertion provides superior modelling of the acquisition physics. This leads to more accurate knowledge of the impact of acquisition parameters, patient factors and scanner properties on scatter, randoms and attenuation and in turn how these relate to the final image quality and quantification of uptake. Simulating just the lesions using Monte Carlo modelling means the ground truth of the simulated lesions is known while being less computationally demanding than simulating a whole patient.

One method of lesion insertion is to use forward-projections of real reconstructed patient PET images to generate sinograms for the insertion of simulated lesions [140]. The sinograms from the simulated lesions are then added to those of the patient and reconstructed. As the patient PET images are already reconstructed using non-linear iterative techniques, the forward-projections are not true representations of the original raw data from the scanner and cannot accurately portray the influence of different acquisition and reconstruction parameters on the lesion characteristics. The proposed technique uses the original sinogram data from the real patient acquisition exported directly from the PET scanner. After insertion, the generated sinograms will be reconstructed and corrections applied in the same way as the original sinogram data using the manufacturer software. This will require an accurate and reliable method for insertion of the lesions.

This chapter describes the design and validation of the model of the clinical PET system used to perform Monte Carlo simulations throughout this thesis [$D710_{simu}$]. The methodology used to generate and insert simulated lesions into real patient PET data in the projection-domain is also described. This technique can be used to generate populations with different lesion characteristics to evaluate a wide range of realistic imaging situations. In this thesis, the data are used to investigate the impact of advanced reconstruction methods on clinical measures in virtual imaging trials.

3.2 Methods

The first stage of the work was to set up the pipeline for performing Monte Carlo simulations of the GE Discovery 710 PET scanner (General Electric Medical Systems, GEMS, Milwaukee, WI) and process the output data to generate reconstructed PET images that are accurate and visually realistic. The accuracy of the simulated data was validated by comparing the output of simulated phantom acquisitions to real experimental data acquired on the clinical GE Discovery 710 PET scanner at the King's College London & Guy's and St Thomas' PET Centre [$D710_{real}$]. The methodology for insertion of the simulated objects into real PET data in projection space was then developed and validated using phantom and clinical data.

3.2.1 Monte Carlo PET Simulations

Monte Carlo simulations were set up using the Geant4 application for tomographic emission (GATE) toolkit (version 8.2) incorporating the Geant4 libraries (version 10.5.p01) [110, 111, 141]. Simulations were performed on a Linux-based (Ubuntu 18.04) PC with an Intel® Core™ i9-9900K processor with 8 cores (16 threads) at 3.6 GHz base frequency and 64 GB RAM. The basic steps in running a simulation in GATE are shown in Figure 3.1. Macro files are used to define the physical properties of the objects to be imaged and the imaging system. The output consists of detailed information for each detected event, including the location and timing, which requires further processing to produce reconstructed PET images.

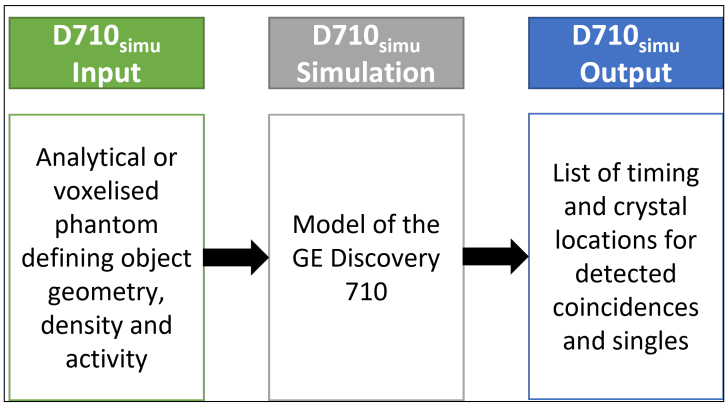

Fig. 3.1 Steps in performing a Monte Carlo simulation in GATE: geometry, density and activity of imaged objects are provided as input to the model which outputs energy, position and timing information for each detected event.

3.2.1.1 Modelling the GE Discovery 710 PET scanner

A series of macro files were written to define the GE Discovery 710 PET scanner geometry, detector electronics and the physics processes for the Monte Carlo simulations. Figure 3.2 demonstrates the steps used to build the PET scanner geometry in GATE. The first step creates the dimensions of the world within which the whole scanner geometry is contained. Next, the components of the scanner are defined as volumes at different levels, each being a daughter of the previous level: $world \rightarrow cylindricalPET \rightarrow rsector \rightarrow module \rightarrow crystal$. The dimensions

(x, y, z) of each volume and their locations relative to the mother volume are specified in the macro.

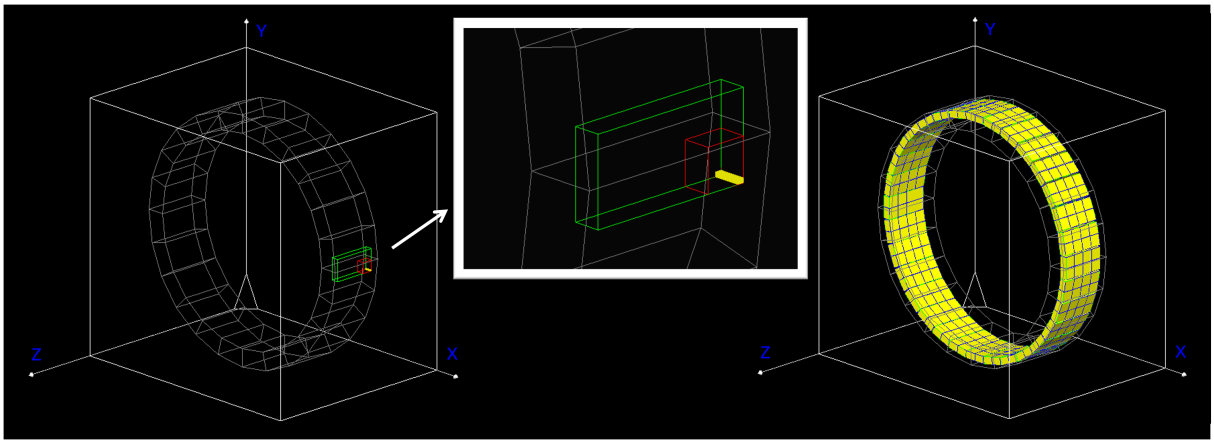

Fig. 3.2 GATE visualisation showing the addition of modules (green), blocks (red) and the crystals (yellow) to construct the final $D710_{simu}$ model.

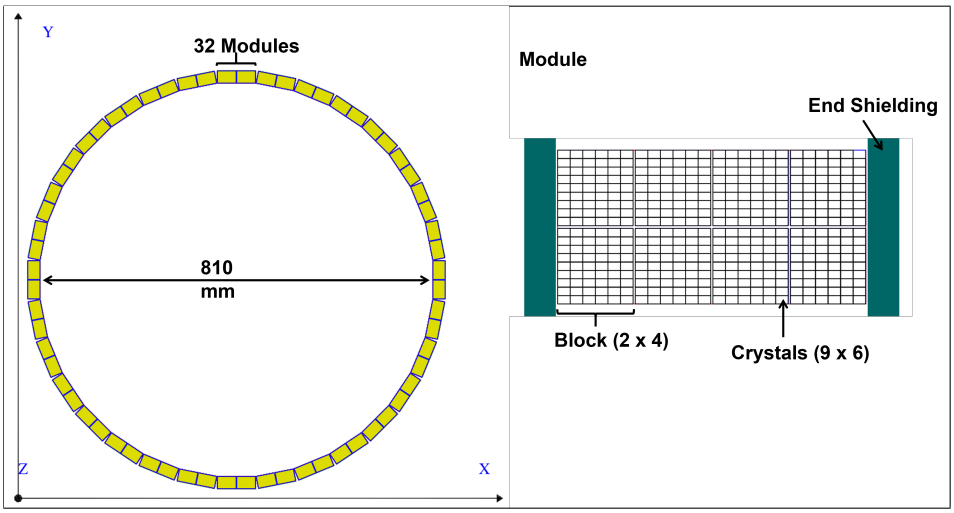

Fig. 3.3 GATE visualisation showing the construction of the individual modules and their arrangement in the $D710_{simu}$.

Appendix A shows a shortened version of the macro written to produce the $D710_{simu}$ geometry. The *cylindricalPET* constructor was used and the detector configuration was based on published information for the GE Discovery 690 [13] [1]. In total, the scanner contained 13 824 crystals arranged in 32 modules around the scanner circumference giving an inner diameter of 810 mm as illustrated in Figure 3.3. Each detector module (defined as an *rsector* in GATE)

[1]The GE Discovery 690 and 710 models have the same PET hardware configuration

consisted of 2 × 4 blocks (defined as *modules* in GATE) which in turn contained 9 × 6 crystals. Individual crystal dimensions were 4.2 mm × 6.3 mm and 25 mm depth. Additionally, the scanner casing, end-shielding and detector backing were defined in the full macro.

The chemical composition of the materials used for the different components of the scanner were defined in a materials database describing the elements and their mass fractions. The composition of the lutetium-based scintillator material in the $D710_{real}$ was unavailable, therefore the material used for the crystals in the $D710_{simu}$ was based on published data for the GE Discovery RX [142]. The scintillator material was defined as lutetium-yttrium-orthosilicate (LYSO) with a relative yttrium content of 6 % resulting in a density of 7.23 g cm^{-3}.

To model the detector response and signal processing of the PET system, GATE uses a digitizer module whereby the energy deposited in the crystal is processed via a series of filters to generate digital pulses. The key parameters for the $D710_{simu}$ digitizer module were based on published performance data [13] and known scanner settings as summarised in Table 3.1.

Digitizer Parameter	**Value**
Mean Energy Resolution	12.4 %
at 511 kilo-electron volt (keV)	
Energy Window Settings:	
Lower Threshold	425 keV
Upper Threshold	650 keV
Coincidence Time Window Width	5.4 nanoseconds
Mean Coincidence Timing Resolution	544 picoseconds

Table 3.1 Parameters used in the digitizer modules for the $D710_{simu}$ model in GATE.

The crystal quantum efficiency, which sets the efficiency of the system to convert energy deposited in the crystal into an electrical signal, was adjusted until the $D710_{simu}$ produced matching coincidence and singles count rates for an identical acquisition of a uniform cylinder on the $D710_{real}$. A final value of 0.90 was chosen for the efficiency. The difference in measured

prompts, randoms, scatter and singles for the cylinder scan were −3.7 %, −1.1 %, 6.7 % and 1.4 % respectively. The event data recorded by GATE provides exact crystal locations for each detected photon which would result in a spatial resolution exceeding that of the real scanner. Therefore, to obtain a spatial resolution analogous to that of the $D710_{real}$, Gaussian blurring was added to the radial and axial crystal locations before binning into sinograms. The degree of blurring was adjusted iteratively to match the spatial resolution of the $D710_{real}$ measured using point sources resulting in full-width at half-maximum (FWHM) values of 0.78 mm and 0.82 mm.

3.2.1.2 Defining the Phantom Input

A combination of analytical and voxelised phantoms were used as input for the simulations in GATE. Analytical phantoms were defined in macros in the same way as the model of the imaging system whereby the geometry and materials for the components of the phantom were specified as volumes positioned within the world relative to the scanner volume. For voxelised phantoms, CT images of the phantoms or patients were imported into GATE and the material properties or radioactivity assigned according to the HU on a voxel-by-voxel basis using lookup tables. For example, a voxel value of -1000 HU would be assigned as air with no radioactivity present and a value of 0 HU assigned water with a certain radioactivity in Bq. The materials database in GATE further defines the properties of the designated materials in terms of their density, atomic number and mass to accurately simulate the physics of the particle interactions within the different materials. Half-life and type of decay for the radioisotope were specified. To account for the fact that ^{18}F is not a pure positron emitter, the desired activity was multiplied by the positron fraction of 0.9670.

3.2.1.3 Processing the Output PET Data from the Simulation

Positional and timing information for the detected singles and coincidence events were output in ROOT files (ROOT Data Analysis Framework, CERN, version 6.20). Data from the ROOT files were extracted to text files for further processing in MATLAB (The MathWorks, Inc, version

2020b). Custom MATLAB scripts were written to pre-process the simulated PET data prior to reconstruction as outlined in Figure 3.4.

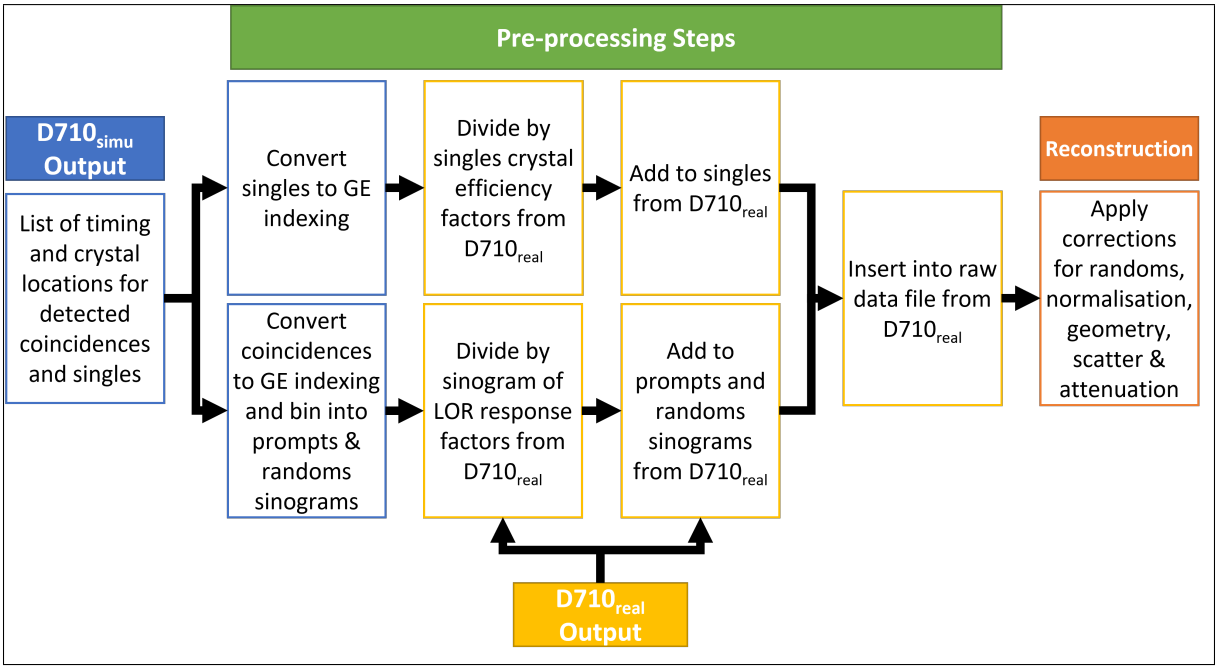

Fig. 3.4 Outline of the pre-processing steps for the simulated PET data before reconstruction.

First the GATE crystal indexing was converted to the GE axial and circumferential indexing. Coincidence events for each line of response (LOR) were then sorted into projections for all direct- and cross-planes according to the $D710_{real}$ geometry and binned into 3D sinograms. For 3D time-of-flight (TOF) volumes, the difference in arrival times for coincidence events was also used to sort the LOR data into separate time bins.

To obtain artefact-free results from the $D710_{simu}$, it is important to correct for non-uniform detection efficiencies as would be done for a real PET scanner [143]. On commercial scanners non-uniformities in detector response are corrected for using component-based normalisation [59] consisting of separate components relating to geometrical effects and individual crystal efficiencies (see Section 2.1.5). The former is fixed for a specific scanner geometry; therefore it was assumed the geometry for the $D710_{simu}$ was equivalent to the $D710_{real}$ and the geometric component of the normalisation from the $D710_{real}$ was applied to the simulated data during reconstruction as normal. Unlike the $D710_{real}$ however, crystals in the $D710_{simu}$ model had

uniform response. To prevent over-or under-correction of the simulated data when inserted into real scans, the simulated singles were divided by the individual crystal efficiencies for the $D710_{real}$ and the simulated prompts sinograms were divided by the response for each LOR derived from the crystal efficiencies.

For phantom acquisitions, the intrinsic radiation in the LYSO crystals was accounted for by adding the individual crystal singles and prompts sinograms from the $D710_{simu}$ acquisitions to those from scans of equivalent duration performed on the $D710_{real}$ without any activity in the field of view (FOV). The contribution of intrinsic radiation to the coincidences count rate was measured on the $D710_{real}$ as approximately 1000 cps based on a 1 hour scan with no activity in the field of view.

To generate equivalent output to the $D710_{real}$, all reconstructions and corrections (randoms, normalisation, geometry, attenuation and scatter) were performed in MATLAB using the GE proprietary PET Toolbox (Duetto, version 2.07) with parameters matching those used on the $D710_{real}$. The randoms correction method on the $D710_{real}$ is derived from the singles [63] rather than using a delayed coincidence window technique. To retain all the required DICOM header information and for compatibility with the toolbox, the prompts sinograms and individual crystal singles data from the simulation were inserted into existing GE raw data files of matching duration exported from the $D710_{real}$.

3.2.1.4 Simulating Lesions

To generate PET images consisting of real patient scans with inserted simulated lesions, the sinograms from simulated lesions performed on the $D710_{simu}$ were added to the sinograms from real PET scans acquired on the $D710_{real}$.

The voxelised activity and density phantoms for defining the lesions in GATE were generated in MIM Encore (MIM Software Inc, version 7.1.5) from the CT component of real PET-CT acquisitions on the $D710_{real}$. The CT was resampled to 1.0 mm^3 isotropic voxels and 3D volumes of interest (VOI) used to define the size and shape of the lesions for insertion. For the density

phantom, the voxels within the VOIs were set to the desired HU for the lesion. Conversion of the CT HU to physical densities for tissues used a lookup table based on the data published by Schneider *et al* [144]. To generate the activity phantom, all voxels outside the VOIs were set to zero and those inside to 1. A text file was then used to specify the activity (in Bq) contained in each voxel of the binary image equal to one.

To account for the fact that there was already tissue with uptake in the real data prior to insertion, the activity set in GATE was scaled by subtracting the mean activity concentration measured in the real PET images at the lesion location. The density phantom was also used for CT-based attenuation correction of the PET data during reconstruction. The total coincidence events in the region where the lesion is inserted in the healthy patient is made up of events from the patient lung that are attenuated by the lung tissue (HU_{lung}) and events from the lesions attenuated by the lesion tissue (HU_{lesion}). The HU in the CT used to derive the attenuation correction factors (ACFs) for attenuation correction (HU_{ctac}) were adjusted to compensate for the different proportions using a method analogous to that proposed by Stute *et al* [145]:

$$HU_{ctac} = \left(\frac{C_{lesion}}{C_{lesion}+C_{lung}}\right) * HU_{lesion} + \left(\frac{C_{lung}}{C_{lesion}+C_{lung}}\right) * HU_{lung}$$

where the mean activity concentrations in the lung (C_{lung}) and lesion (C_{lesion}) were used as surrogates for the number of coincidence events. It was assumed the region of the lung where the lesion was inserted had no structure and the mean HU value for the lung was used.

3.2.2 Validation of the Scanner Model using Phantoms

To validate the GATE D710$_{\text{simu}}$ model, simulations of the performance tests for spatial resolution and system sensitivity were performed according to the National Electrical Manufacturers Association (NEMA) NU 2-2012 Standard [146]. The simulated point and line sources consisted of analytical phantoms defined in GATE using simple geometric shapes in a macro. Simulated phantoms were filled with activities matched to those used in the real phantoms after accounting for decay and residual syringe activities. The simulations were repeated three times and mean

results for the $D710_{simu}$ were compared against identical performance tests conducted as part of acceptance testing for the $D710_{real}$ at the King's College London & Guy's and St Thomas' PET Centre.

3.2.2.1 Spatial Resolution

Spatial resolution is often used to assess PET scanner performance and is a measure of the ability of a system to resolve small objects. The resolution is dependent on the radionuclide and the geometry of the scanner (see Section 2.1.1). Usually this is assessed using images reconstructed using filtered backprojection (FBP) as it is a linear algorithm and easily comparable between different PET systems [147]. In addition, it is useful to measure the spatial resolution for reconstruction algorithms used clinically such as ordered subsets-expectation maximisation (OS-EM).

Spatial resolution was measured for three 1 mm diameter ^{18}F point sources created at the ends of glass capillary tubes and suspended in air. Each source contained approximately 200 MBq ml^{-1} and was positioned at (x, y) locations of (0, -10), (0, -100) and (100, 0) mm respectively. 60 s acquisitions were performed with the three sources located at the centre of the axial FOV and at a quarter of the FOV (z = 0.0 mm and 39.3 mm).

PET data were reconstructed to give (x, y) pixel sizes of 0.8 mm × 0.8 mm and slice thickness of 3.3 mm using 3D Fourier rebinning followed by 2D FBP (FORE-FBP) and 3D OS-EM (VPHD). The VPHD reconstruction used 2 iterations, 24 subsets and a Gaussian post filter of 6.4 mm FWHM to match the parameters used clinically. Profiles were plotted in all three orthogonal directions on the reconstructed images of the point sources and the resolution measured as the full-width at half-maximum (FWHM) and the full width at tenth maximum (FWTM) of the peaks.

3.2.2.2 System Sensitivity

The sensitivity of a PET system is defined as the number of counts measured per unit time per unit radioactivity (cps kBq^{-1}). Positron emitters require a significant amount of material around the source to annihilate with an electron and produce the 511 keV coincident photons. However any material surrounding the source also attenuates the photons so a direct measurement of absolute sensitivity in air independent of attenuation and scatter is not possible [148].

To overcome this, the test described by the NEMA standard uses repeated measurements of an ^{18}F line source surrounded by increasing thicknesses of attenuating material, in this case aluminium tubes. The thickness of the wall for the first aluminium tube (1.25 mm) is chosen to match the maximum range of ^{18}F positrons in aluminium (E_{max} = 0.634 MeV) [32]. It is therefore assumed that all positrons are annihilated in the first tube and any decrease in count rate observed with the additional tubes is purely due to attenuation of the photons [148]. To derive the attenuation-free sensitivity, the data is extrapolated to give the sensitivity with no attenuating material [149]. The activity in the line source should be sufficiently low to produce a randoms coincidence rate of less than 5 % of the total event rate [146], which based on the manufacturer's recommendations, is < 10 MBq for the $D710_{real}$ [150].

System sensitivity was measured using a 70 cm line source filled with an ^{18}F solution giving 4.83 MBq at the start of the first acquisition. The line source was inserted into the first aluminium tube and suspended in air along the centre of the transaxial FOV. Five sequential 60 s acquisitions of the line source were performed each with an additional aluminium tube to attenuate the annihilation photons. The acquisitions were repeated with the line source suspended at 10 cm radial offset. The PET data was processed in MATLAB using single-slice rebinning (SSRB) following the procedure described in the NEMA NU 2-2012 Standard [146] and extrapolated to zero attenuation to give the system sensitivity in air.

3.2.2.3 Validation of Corrections

Acquisitions of a uniform cylinder are sensitive to errors in scanner calibration, normalisation, scatter and attenuation corrections [151]. Therefore to check the accuracy of the corrections applied during reconstruction of the $D710_{simu}$ data, matching acquisitions of a uniform cylinder containing ^{18}F were acquired on the $D710_{real}$ and $D710_{simu}$ according to the Quantitative Imaging Biomarkers Alliance (QIBA) FDG PET-CT Profile [152]. The analytical phantom used as input for the simulations in GATE consisted of a cylinder of water with dimensions of 9.8 cm radius and 17.6 cm length surrounded by a 4 mm thick plastic shell. CT-based attenuation correction of the $D710_{simu}$ PET data was performed using a synthetic CT of the cylinder created in MATLAB. The physical densities of air, water and plastic (0.00 g cm^{-3}, 1.00 g cm^{-3} and 1.18 g cm^{-3}) were transformed to Hounsfield units (HUs) for 140 kV p and the voxel values in the CT were set to these values. A target activity concentration of 5 kBq ml^{-1} was chosen for the cylinder based on an injected activity of 350 MBq assuming uniform distribution in a 70 kg patient. This is also within the range of 3.7 kBq ml^{-1} to 7.4 kBq ml^{-1} recommended by QIBA [152].

The cylinder was filled with an ^{18}F solution giving 4.78 kBq ml^{-1} at the start of the acquisition. The phantom was suspended in air (Figure 3.5) and scanned for a total of 300 s. To speed up the simulation time, the acquisition was performed as 5 $\times$ 60 s simulations and the data merged before binning into sinograms to give the final 300 s dataset. PET data were reconstructed using 3D OS-EM (VPHD) with 2 iterations, 24 subsets, a Gaussian post filter of 6.4 mm FWHM and voxel size of 2.7 mm $\times$ 2.7 mm $\times$ 3.3 mm.

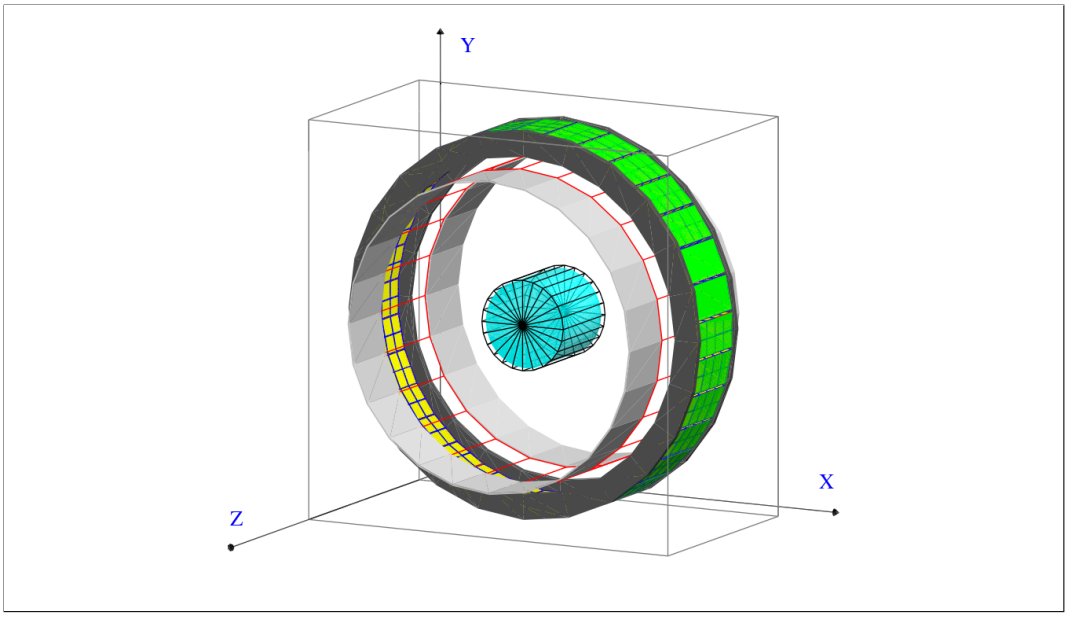

Fig. 3.5 Visualisation of the $D710_{simu}$ model in GATE showing the uniform cylinder suspended in the centre of the gantry.

The reconstructed PET images were analysed in MIM Encore (version 7.1.5, MIM Software Inc.) using the procedure described by the QIBA FDG profile [152]. A circular region of interest (ROI) of approximately 200 cm^2 (16 cm diameter) was placed on the central axial slice and copied to all other slices. To evaluate any quantification bias in the simulations, the mean activity concentration measured for the reconstructed images was compared to the activity concentration set in the simulation. The noise was measured as the standard deviation (SD) within each ROI and the slice-by-slice variation as the SD in the mean activity concentration measured for the individual slices. Values for the simulated cylinder were compared to the noise and variation measured for a uniform cylinder acquired under the same conditions on the $D710_{real}$ with matching activity.

3.2.3 Phantom Validation of the Lesion Insertion Technique

The physical NEMA IEC body phantom (Data Spectrum Corporation) was used for validation of the lesion insertion technique. The NEMA phantom consists of a fillable 'torso' background and 6 fillable spheres with inner diameters of 10, 13, 17, 22, 28 and 37 mm, as shown in Figure 3.6.

Simulated spheres of water without walls were inserted into real acquisitions of the phantom background and compared to acquisitions of the physical phantom with the spheres and the background together.

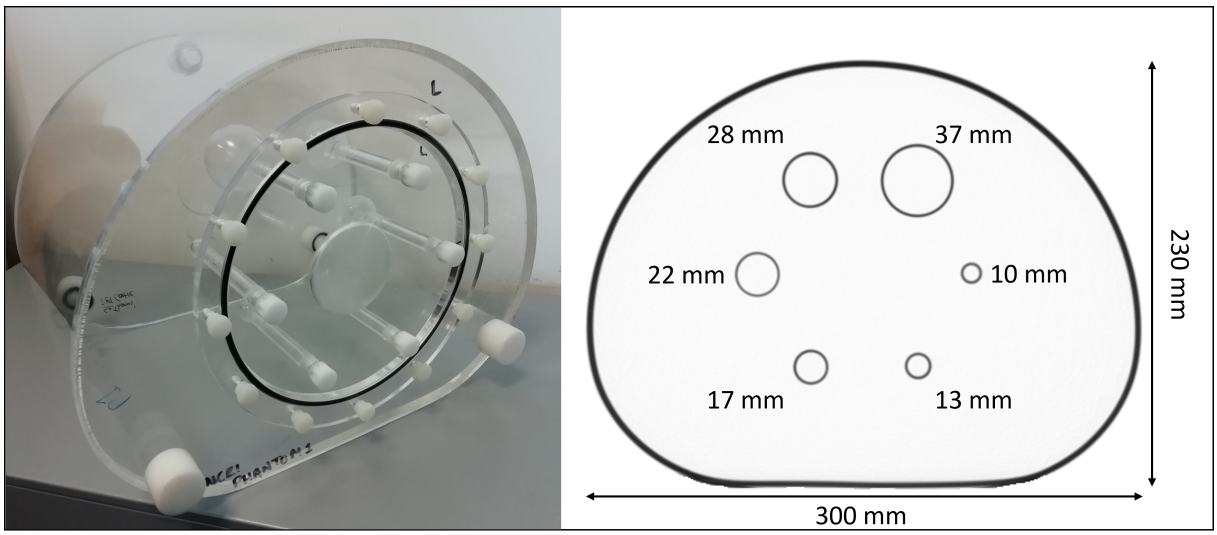

Fig. 3.6 Photo of the physical NEMA IEC body phantom (left) alongside a schematic diagram (right) showing inner diameters of the fillable spheres.

Separate acquisitions (n = 3) of the physical NEMA phantom on the $D710_{real}$ with both background and spheres filled with ^{18}F were used as the reference data [$NEMA_{real}$]. For each acquisition, the phantom was placed on the patient couch and centralised in the gantry. The mean ^{18}F activity concentrations in the background and spheres at the time of scanning were $5.0 \pm 0.6\,kBq\,ml^{-1}$ and $26.5 \pm 6.5\,kBq\,ml^{-1}$. For all acquisitions a CT was acquired (140 kV p, variable mA s, 0.5 s rotation time and 1.375 pitch) followed by a 3 min PET acquisition.

A single 10 minute list mode acquisition of the NEMA phantom background filled with ^{18}F without spheres was performed on the $D710_{real}$. The activity concentration in the phantom at the scan start was $7.8\,kBq\,ml^{-1}$. The list mode data was binned into shorter scan durations of 95 s, 107 s and 123 s to achieve equivalent count statistics as the three real phantom acquisitions with different background activity concentrations. The sinogram data was used for the insertion of the spheres simulated using the same scan durations [$NEMA_{simu}$].

Using the method described in Section 3.2.1, the CT of the $NEMA_{real}$ was resampled to 1.0 mm^3 isotropic voxels and the spheres defined using spherical VOIs with diameters matching those of the real phantom, Figure 3.7a). For the density phantom, voxel values within the VOIs were set to a fixed HU equivalent to the density of water (1.0 $g\,cm^{-3}$), Figure 3.7b). Since the background in the real phantom was also water, the HU was not adjusted in the CT used for attenuation correction.

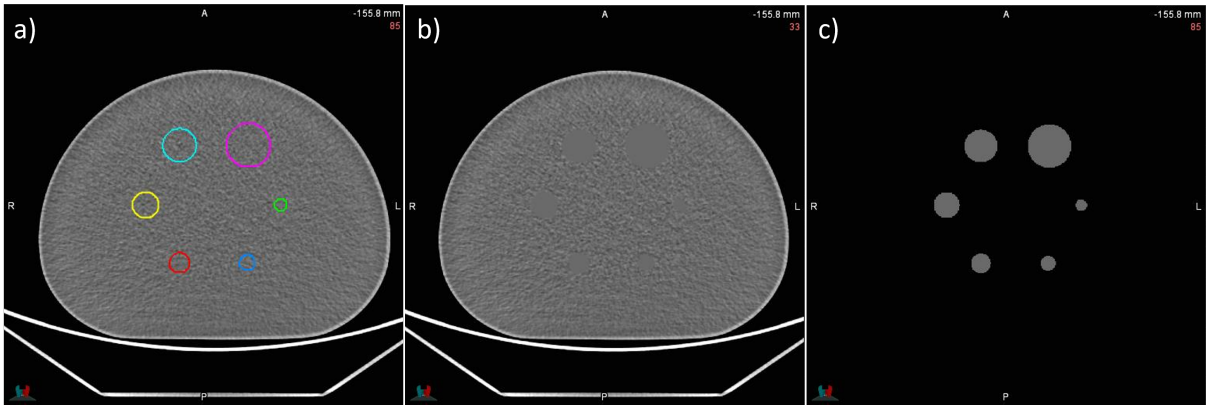

Fig. 3.7 Voxelised phantoms for defining the density and activity are generated using a real CT scan of the NEMA phantom. a) Spherical VOIs are drawn on the resampled CT. b) HU within the VOIs are set to match the density of water. c) Voxels outside the VOIs are set to zero.

For the activity phantom, Figure 3.7c), the background voxels were set to zero for no activity and those within the VOIs set to 1. The activity for the simulated spheres was chosen to match the spheres:background ratio of 5.3 ± 1.2 times the background in the $NEMA_{real}$. To account for the fact that there was already activity in the phantom, the activity within the spheres was reduced by the mean activity measured in the phantom background.

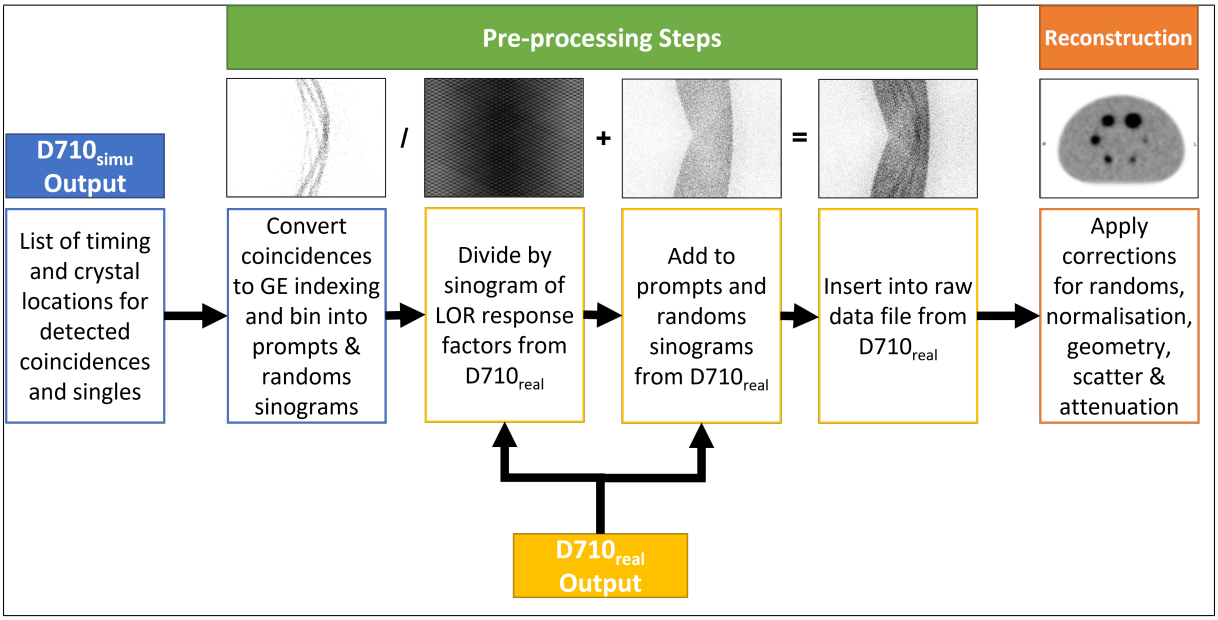

Fig. 3.8 Steps involved in the insertion of the simulated spheres into the real NEMA background in the projection-domain.

Figure 3.8 illustrates the steps involved in insertion of the simulated spheres into the real phantom background using the technique described in Section 3.2.1. PET data were reconstructed using 3D OS-EM (VPHD) and 3D OS-EM with TOF (VPFX) with 2 iterations, 24 subsets and a Gaussian post filter of 6.4 mm FWHM.

The activity concentrations for the spheres and background in both the $NEMA_{real}$ and $NEMA_{simu}$ were measured using MIM Encore. The activity concentration in the spheres was measured as the maximum voxel value and the mean for a 3D VOI with a threshold of 50 % of the maximum voxel value. Recovery coefficients (RCs) for the spheres were calculated as the ratio of the measured to true activity concentration. Results for the three individual acquisitions of the $NEMA_{real}$ and $NEMA_{simu}$ were averaged and plotted to generate the recovery coefficient (RC) curves for comparison.

3.2.4 Insertion into Clinical Data

Anonymized PET-CT datasets from ten clinical patients with a measurable pulmonary nodule were used for insertion of the simulated lesions. The PET-CT scans were acquired using 3 min

per bed position with an 11 slice (23 %) overlap and reconstructed using VPFX (2 iterations, 24 subsets and 6.4 mm FWHM Gaussian post filter). The CT scans were acquired using 140 kV p, variable mA s, 0.5 s rotation time, 1.375 pitch and reconstructed using the standard kernel and adaptive statistical iterative reconstruction (ASiR) blending of 40 %. The mean and standard deviation (SD) for injected activity, uptake time and body mass index (BMI) were 338.3 ± 35.0 MBq, 82.2 ± 9.3 min and 24.9 ± 4.6 kg m^{-2} respectively.

The reconstructed PET and CT images for the ten patients were resampled to 1 mm × 1 mm × 1 mm voxels in MIM Encore. The real lesions were segmented on the resampled PET images using a 3D VOI with a threshold of 50 % of the maximum voxel value as this most closely matched the anatomical extent of the lesions on the resampled CT and required the least manual adjustment. Using the co-registered CT images, the VOIs were manually adjusted to exclude any non-tumour tissues such as the heart or rib bone. For each PET scan the VOI was mirrored to the opposite lung and, if necessary, shifted in the x-direction to ensure the lateral location was matched in both lungs, as shown in Figure 3.9.

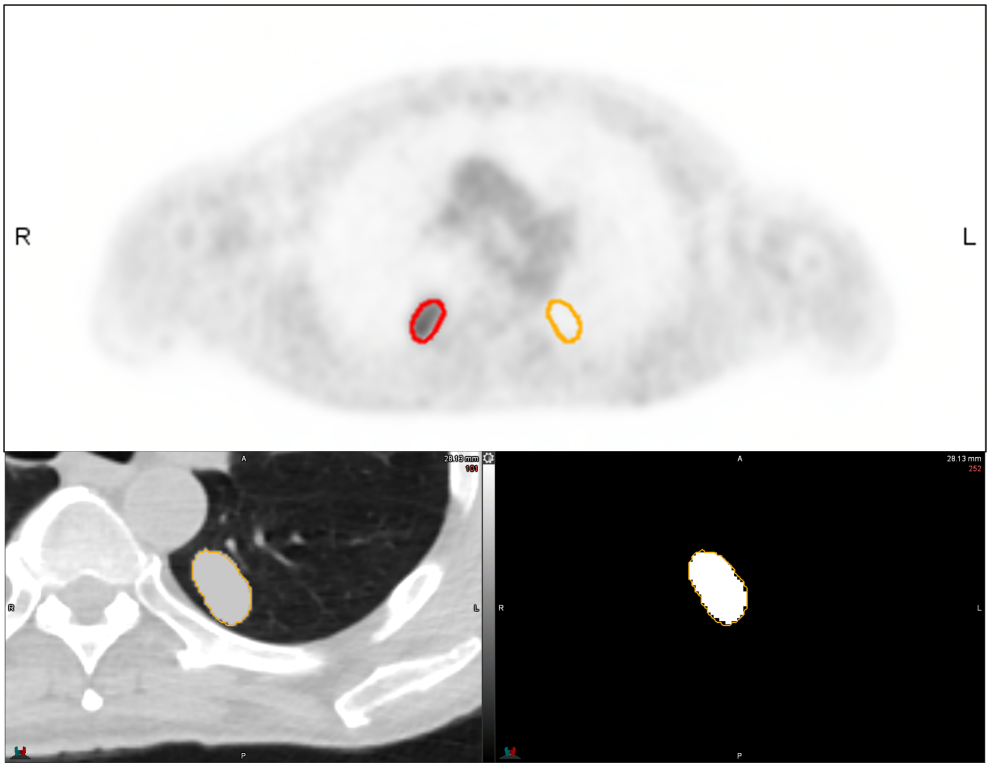

Fig. 3.9 Example of a real lesion outlined on the PET image and copied to the contra-lateral lung (top) to generate the voxelised density (bottom left) and activity (bottom right) phantoms used for the simulation input.

To generate the density input image, the mirrored VOI was copied to the CT images and the voxels within the VOI were set to 16 HU, this value was chosen based on the median unenhanced HU measured by Swenson *et al* for malignant and benign solitary pulmonary nodules (SPNs) [153]. The target activity set in the simulated lesion was derived from the maximum activity concentration measured in the real lesion in the contralateral lung. To estimate the true activity in the real lesion, partial volume correction (PVC) was applied using the real lesion diameter and the recovery coefficient (RC) measured for the clinical reconstruction, 3D OS-EM with TOF (VPFX), on the $D710_{real}$ using the technique described by Srinivas *et al* [154]. It was assumed the uptake and density in the area of the lung where the lesion was inserted was uniform. If the lesion was located in the region where two bed positions overlapped in the real scan, separate activity and density phantoms were created for each bed position with the lesion located accordingly.

The simulations were run for 3 min per bed position to match the real PET scans. Once inserted, the PET scans were reconstructed using the same parameters as the original clinical PET data. Two bed positions, covering the thorax, were reconstructed for all patients using an 11 slice bed overlap.

3.2.4.1 Quantitative Assessment

To assess the ability of the technique to faithfully reproduce the real lesions in the contra-lateral lung, standardised uptake values (SUVs) were measured for the real and simulated lesions in the reconstructed PET images. 3D VOIs using 50 % threshold of the maximum voxel value were used to outline the lesions. If necessary, the VOIs were manually edited to exclude non-lesion uptake and the SUV_{max}, SUV_{mean} and SUV_{peak} values measured. The 1 cm^3 sphere to define the SUV_{peak} was allowed to extend beyond the VOI and was not restricted to the maximum voxel value.

3.2.4.2 Clinical Observer Study

An experienced PET Physician then performed a blinded two-alternative forced choice (2AFC) test to determine whether the simulated lesions were indistinguishable from the real lesions. The 10 datasets were displayed and analysed using MIM Encore. The clinician was asked to review the PET images as they would for a clinical PET study except without using the CT as the synthetic lesions inserted in the contralateral lung of the CT were clearly artificial and would bias the result. The PET windowing used settings of SUV 0 to 10 with an inverse linear colour scale. They were asked to choose which lung they thought contained the simulated lesion and the confidence in their choice using the following scale:

- definitely left
- probably left
- unsure
- probably right

- definitely right

Additionally, they were asked to record any observations such as artefacts or clinical appearances that informed their decision.

A weighted score was assigned to the answers as shown in Table 3.2, where a negative score was assigned if the clinician correctly identified the simulated lesion. The score was weighted according to their confidence in identifying the lesions meaning a score of −10 would indicate the clinician was able to confidently identify all the simulated lesions.

	Unsure	**Probably**	**Definitely**
Simulated lesion selected	0.0	−0.5	−1.0
Real lesion selected	0.0	0.5	1.0

Table 3.2 Weighted scoring assigned to the 2AFC clinician review. Negative values indicate the clinician identified the simulated lesion.

3.3 Results

3.3.1 Validation of the Scanner Model using Phantoms

3.3.1.1 Spatial Resolution

Tables 3.3 and 3.4 show the axial and transverse spatial resolution measured for the $D710_{simu}$ alongside results from the $D710_{real}$ using FORE-FBP and VPFX reconstructions. As shown, the FWHM and FWTM results for the $D710_{simu}$ were similar to those for the the $D710_{real}$ with maximum absolute differences of 1.03 mm for the FORE-FBP reconstruction and 1.00 mm for the VPFX reconstruction.

Spatial Resolution	**FWHM**		**FWTM**	
(mm)	**D710$_{real}$**	**D710$_{simu}$**	**D710$_{real}$**	**D710$_{simu}$**
At 1 cm radius:				
Transverse	4.69 ± 0.06	4.71 ± 0.01	9.18 ± 0.09	9.80 ± 0.02
Axial	5.54 ± 0.06	5.11 ± 0.02	11.23 ± 0.12	12.26 ± 0.01
At 10 cm radius:				
Transverse radial	5.46 ± 0.06	5.57 ± 0.02	10.33 ± 0.05	10.99 ± 0.08
Transverse tangential	4.91 ± 0.04	4.93 ± 0.02	9.42 ± 0.02	10.41 ± 0.03
Axial	5.99 ± 0.05	6.00 ± 0.02	12.11 ± 0.07	12.97 ± 0.01

Table 3.3 Spatial resolution measured as full-width at half-maximum (FWHM) and full width at tenth maximum (FWTM) for the D710$_{real}$ and D710$_{simu}$ scanners using the Fourier rebinning followed by 2D FBP (FORE-FBP) reconstruction.

Spatial Resolution	**FWHM**		**FWTM**	
(mm)	**D710$_{real}$**	**D710$_{simu}$**	**D710$_{real}$**	**D710$_{simu}$**
At 1 cm radius:				
Transverse	7.44 ± 0.01	7.58 ± 0.01	13.85 ± 0.01	14.39 ± 0.02
Axial	6.48 ± 0.03	6.39 ± 0.05	13.22 ± 0.02	13.55 ± 0.08
At 10 cm radius:				
Transverse radial	7.91 ± 0.01	8.06 ± 0.01	14.67 ± 0.03	15.10 ± 0.03
Transverse tangential	7.52 ± 0.02	7.71 ± 0.02	14.01 ± 0.02	14.58 ± 0.01
Axial	6.03 ± 0.01	6.37 ± 0.01	12.57 ± 0.03	13.57 ± 0.02

Table 3.4 Spatial resolution measured as full-width at half-maximum (FWHM) and full width at tenth maximum (FWTM) for the D710$_{real}$ and D710$_{simu}$ scanners using the 3D OS-EM with TOF (VPFX) reconstruction.

3.3.1.2 System Sensitivity

The mean system sensitivity measured with increasing number of tubes surrounding the line source is plotted in Figure 3.10. Extrapolation of the data to 0 tubes gave mean ± SD system sensitivities for the $D710_{simu}$ and $D710_{real}$ of $6.83 \pm 0.01 \text{cps kBq}^{-1}$ and $6.83 \pm 0.21 \text{ cps kBq}^{-1}$ respectively.

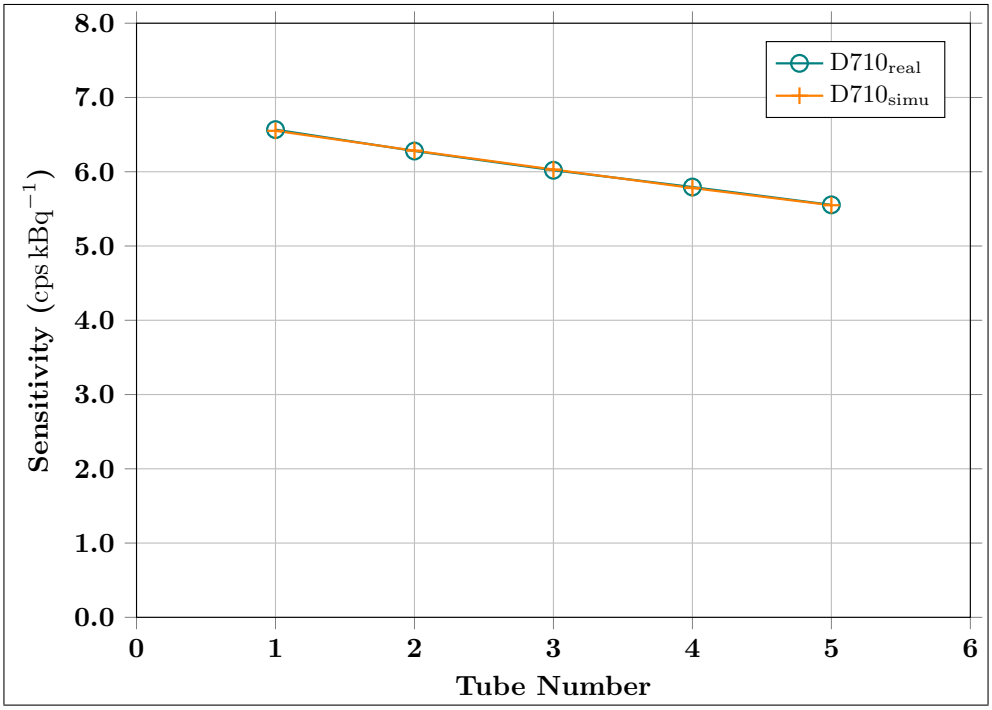

Fig. 3.10 Plot of mean system sensitivity for increasing number of tubes for the $D710_{simu}$ and $D710_{real}$ acquisitions.

Figure 3.11 shows plots of sensitivity for each axial slice measured using the smallest diameter tube at 0 cm radial offset. The mean difference between the slice sensitivity for the $D710_{simu}$ and the $D710_{real}$ was 1.6 ± 2.3 %.

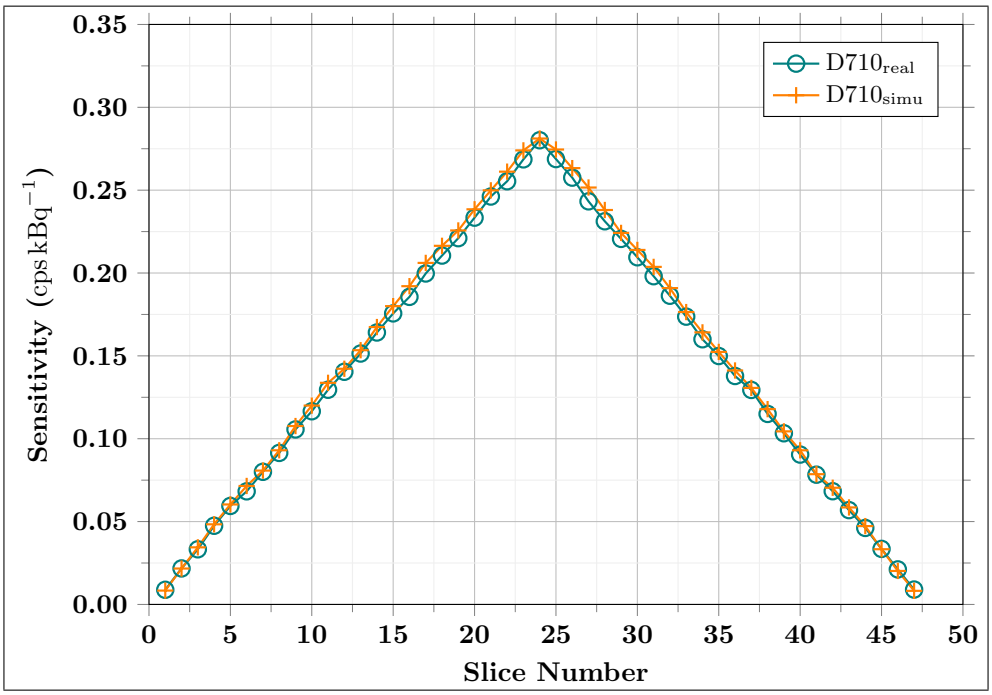

Fig. 3.11 Axial sensitivity profiles for the D710$_{\text{simu}}$ and D710$_{\text{real}}$ acquisitions.

3.3.1.3 Validation of Corrections

Axial, sagittal and coronal images of the real and simulated uniform cylinders reconstructed using VPHD are shown in Figure 3.12. No visual artefacts were observed in the reconstructed images of the simulated phantom.

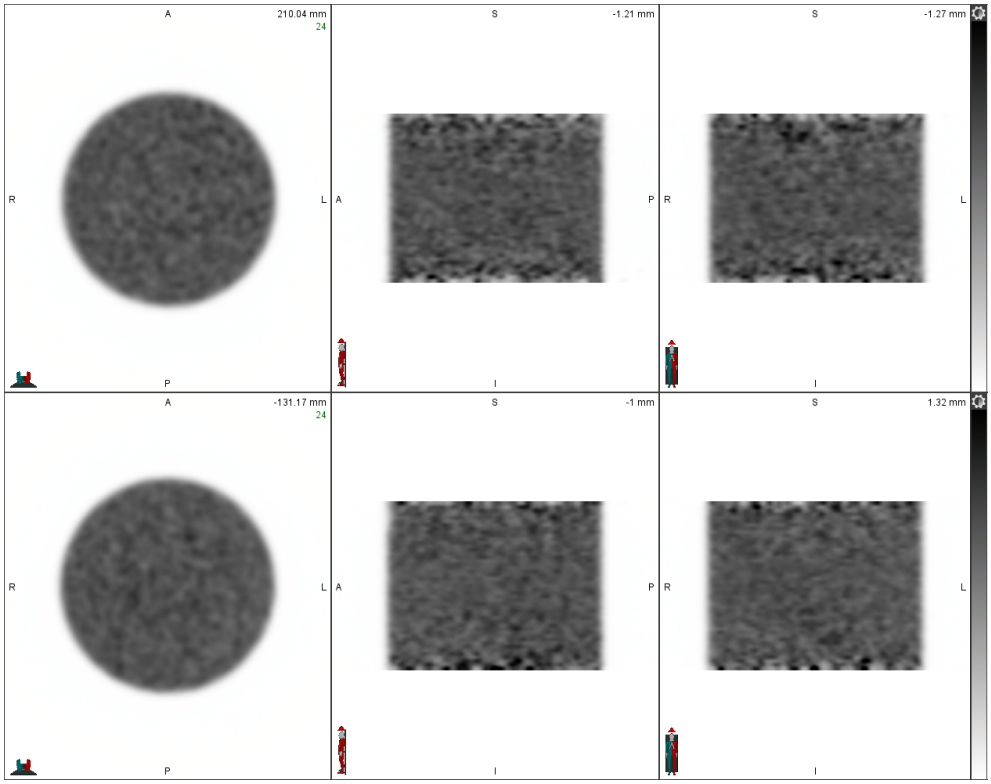

Fig. 3.12 Reconstructed PET images of the simulated (top row) and real (bottom row) uniform phantom acquisitions.

The mean activity concentration measured for each axial slice, after excluding results for the 2 noisy slices at each end of the phantom, is plotted in Figure 3.13. The error bars show ± 1 standard deviation (SD) for the voxel values within the ROI on the individual axial slice. The average difference between the slice mean for the real and simulated cylinders was 0.23 ± 1.52 % (range −3.9 % to 2.5 %).

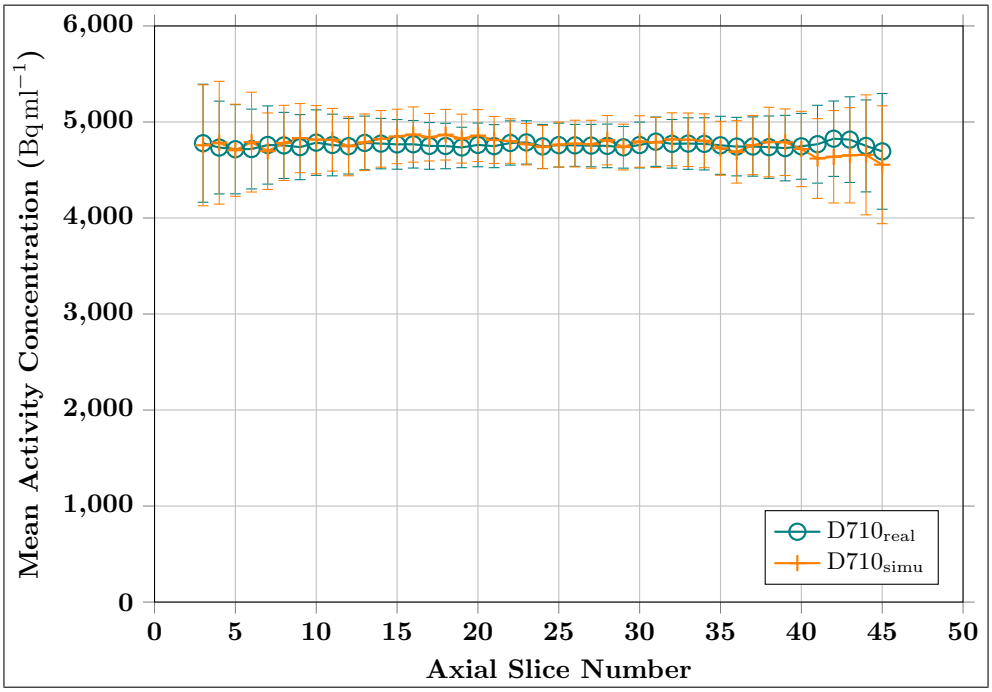

Fig. 3.13 Plot of the mean activity concentration measured for each axial slice of the reconstructed PET images for the D710$_{simu}$ and D710$_{real}$ acquisitions. The error bars represent the noise within each slice ± 1 standard deviation (SD).

The mean activity concentration measured for the simulated cylinder was 4.77 ± 0.07 kBq ml^{-1}, whilst the mean activity concentration for the real cylinder was 4.76 ± 0.02 kBq ml^{-1}. Using the true activity concentration set in the simulation for the phantom, this gave a bias of −0.24 %. This compares to a bias of −0.46 % in the mean activity concentration measurements for the real cylinder using the measurement of activity injected into the physical phantom.

3.3.2 Phantom Validation of the Lesion Insertion Technique

Figure 3.14 shows reconstructed PET images of the NEMA$_{real}$ (top) and NEMA$_{simu}$ (bottom). In the images for the NEMA$_{real}$ the plastic filling tubes can be seen below the spheres in the coronal and sagittal slices. These are absent in the NEMA$_{simu}$ images as they were not included as part of the simulation. No other visual differences or artefacts were seen.

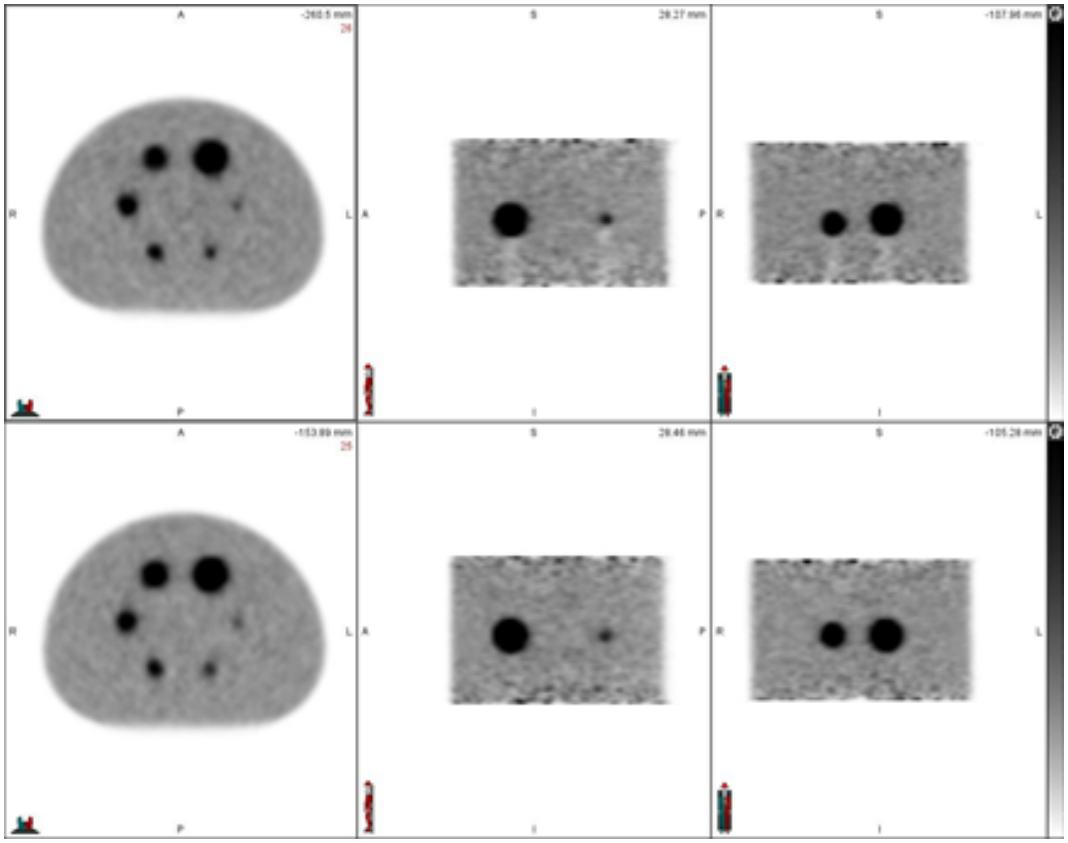

Fig. 3.14 Axial, sagittal and coronal views of the reconstructed PET images. The top row shows the $NEMA_{real}$ and the bottom row the $NEMA_{simu}$.

Figures 3.15 and 3.16 show the recovery coefficient (RC) curves for the maximum and mean activity concentration measured in the different sphere sizes for the $NEMA_{real}$ and $NEMA_{simu}$ when reconstructed using 3D OS-EM (VPHD) and 3D OS-EM with TOF (VPFX). The average ± standard deviation (SD) differences in RCs for the $NEMA_{simu}$ using the VPHD reconstruction were 2.8 ± 5.9 % (range −6.5 % to 9.9 %) for the max and 2.6 ± 4.3 % (range −1.7 % to 9.9 %) for the mean. For the VPFX reconstruction the average differences in RCs for the $NEMA_{simu}$ were −3.2 ± 4.6 % (range −8.1 % to 2.9 %) for the max and −3.6 ± 5.8 % (range −12.6 % to 2.4 %) for the mean.

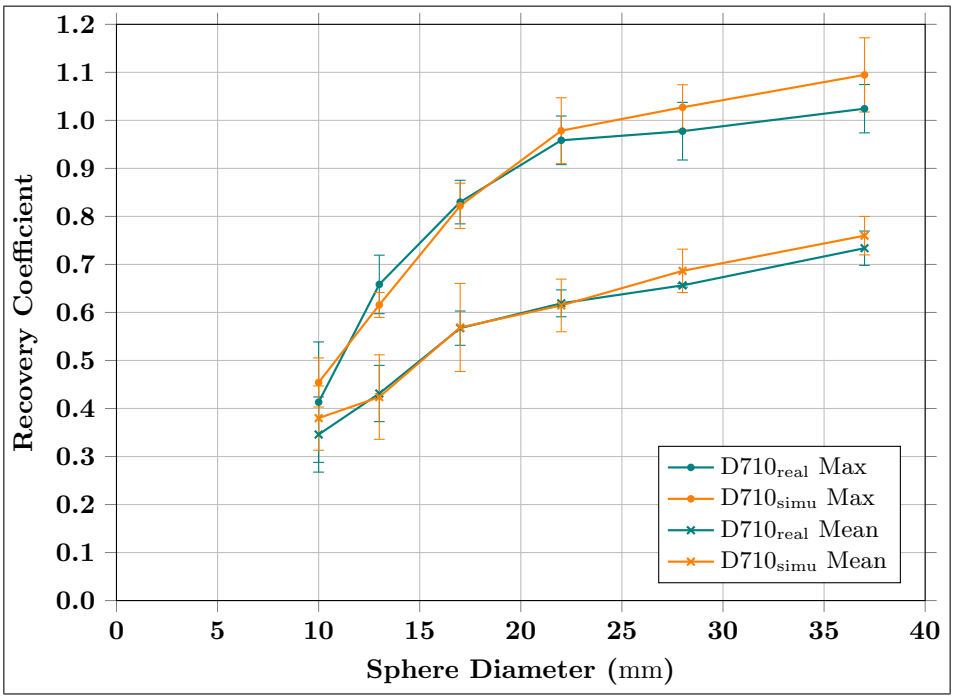

Fig. 3.15 Plots of the average recovery coefficients for the different sphere diameters measured using the 3D OS-EM (VPHD) reconstruction. Results are shown for the max and mean activity concentration in the $NEMA_{real}$ and the $NEMA_{simu}$. The error bars represent ± 2×SD of the mean for the three acquisitions.

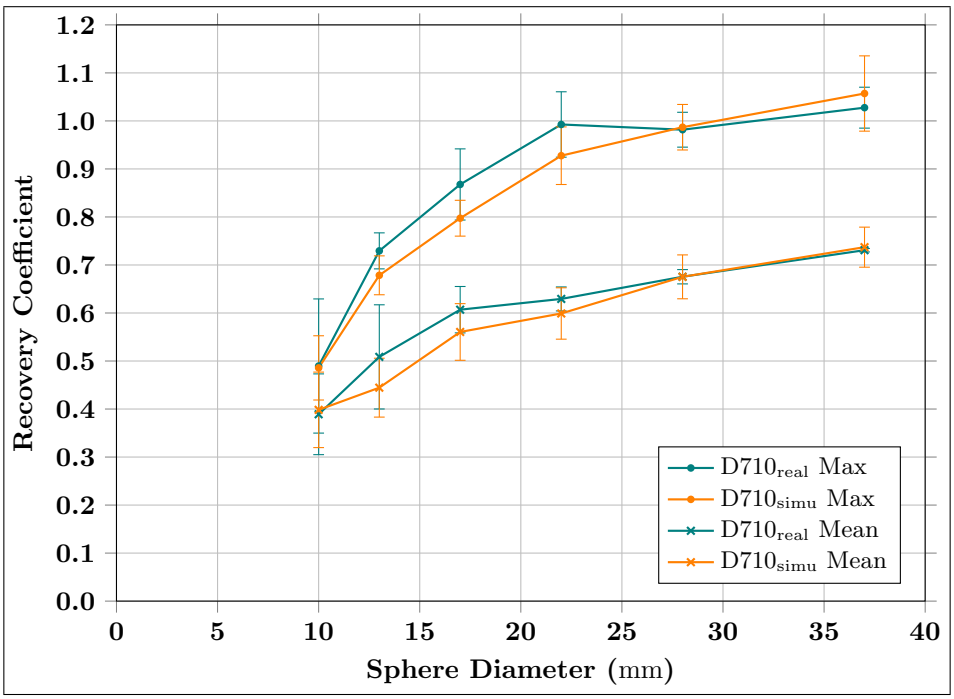

Fig. 3.16 Plots of the average recovery coefficients for the different sphere diameters measured using the 3D OS-EM with TOF (VPFX) reconstruction. Results are shown for the max and mean activity concentration in the $NEMA_{real}$ and the $NEMA_{simu}$. The error bars represent ± 2×SD of the mean for the three acquisitions.

3.3.3 Insertion into Clinical Data

3.3.3.1 Quantitative Assessment

The mean volume of the voxelised phantoms used to simulate the lesions was 3.2 ml (range 0.7 ml to 6.6 ml) with mean activity concentration of 13.7 kBq ml^{-1} (4.2 kBq ml^{-1} to 30.4 kBq ml^{-1}). The mean simulation time for the lesions was 1.3 ± 0.2 min kBq^{-1}. Figure 3.17 shows one of the PET scans with a real pulmonary lesion in the right lung and a simulated lesion in the left lung.

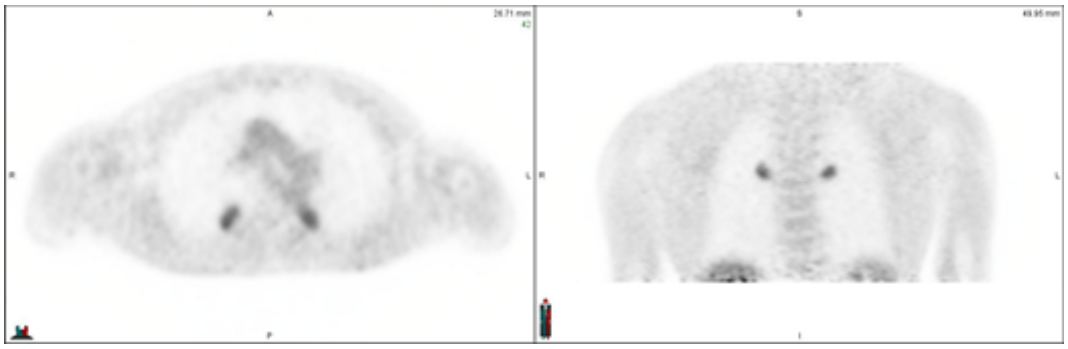

Fig. 3.17 Axial and coronal slices of a real patient scan showing the real lesion (right lung) and inserted simulated lesion (left lung).

The absolute differences in uptake measured for the real and simulated lesions for the 10 patients are plotted in Figure 3.18. Results for the SUV_{max}, SUV_{mean} and SUV_{peak} are shown with error bars representing the estimated standard deviation derived from the repeated acquisitions/simulations of the NEMA phantom in Section 3.3.2. As can be seen on this plot, the SUV measures for the simulated lesions were well matched with the real lesions except for patient 8. In this case, the simulated lesion was located very close to the cardiac tissue which had significant uptake compared to the lung background. As such, spill-in caused the simulated lesion activity to be higher than the real lesion in the contralateral lung which was not located near to cardiac tissue. After excluding the data from this patient, the average ± standard deviation (SD) differences in the SUV_{max}, SUV_{mean} and SUV_{peak} measurements were -0.20 ± 0.29, 0.00 ± 0.19 and -0.10 ± 0.26 respectively.

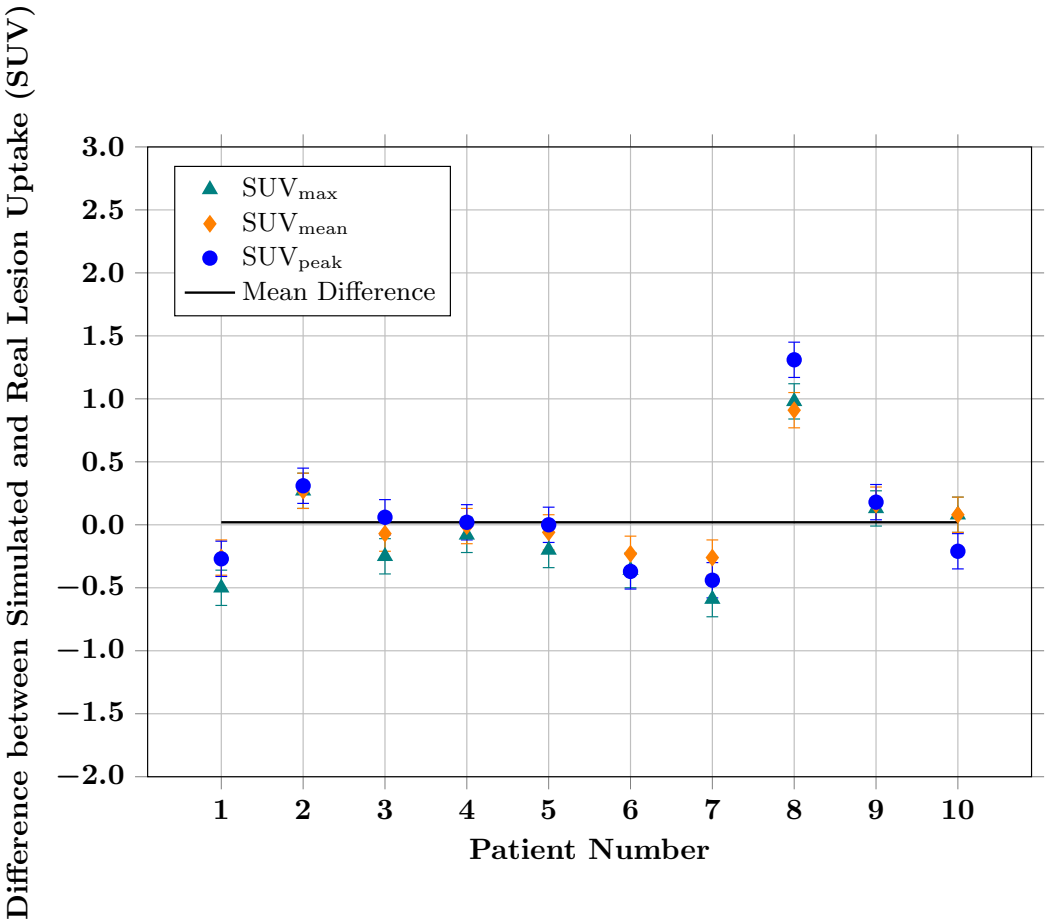

Fig. 3.18 Absolute differences in SUV_{max}, SUV_{mean} and SUV_{peak} measured for the simulated and real lesions in the 10 patients. The error bars represent the estimated standard deviation for the absolute differences derived from the variance in measurements for repeated phantom acquisitions/simulations.

3.3.3.2 Clinical Observer Study

Results for the clinical observer study are presented in Table 3.5. Overall, the clinician identified the simulated lesion location for 4 patients, selected the real lesion for 4 patients and was unsure in 2 patients giving an overall score of 0 (on a scale of -10 to 10). The clinician didn't score any of the lesions as 'definitely' which would suggest that the simulated lesions were not obviously fake compared to the real lesions.

Patient No.	Location of Simulated Lesion	Clinical Review	Score
1	Left upper lobe	Probably left	−0.5
2	Left lower lobe	Probably right	0.5
3	Left upper lobe	Unsure	0.0
4	Left lower lobe	Probably left	−0.5
5	Right upper lobe	Probably right	−0.5
6	Right lower lobe	Probably left	0.5
7	Left upper lobe	Unsure	0.0
8	Left lower lobe	Probably right	0.5
9	Right upper lobe	Probably left	0.5
10	Right lower lobe	Probably right	−0.5

Table 3.5 Location of the simulated lesions for the 10 anonymised datasets and the results of the clinician review.

For patients 1 and 8 the clinician commented on the sharpness of the lesions and chose the sharper of the two which correctly identified the simulated lesion in patient 1 but not 8. The real lesion for patient 4 was bilobed, but this wasn't replicated in the simulated lesion as the assigned activity was uniform making it easier to locate the simulated lesion. Whilst there were not enough cases to perform a statistical test, the results show the described methodology for insertion of simulated lesions into real patient scans can produce clinically realistic results.

3.4 Discussion

This chapter describes the methodology to generate reconstructed PET images consisting of real patient physiology with simulated lesions of known characteristics. Simulations were performed in GATE using a model of the GE Discovery 710 PET scanner that was validated against the real PET system to ensure accuracy of the output data. Lesions were simulated after accounting for existing uptake and density in the clinical images and then inserted into the real patient

scans in projection space. The combined raw PET data were then reconstructed and corrections applied using the GE research toolbox. Phantom assessment of the lesion technique showed the methodology produced images of the simulated objects of similar appearance and uptake to those of the physical objects acquired on the real scanner. Finally, a series of 10 simulated lesions were inserted into anonymised patient scans and assessed using quantitative measures and a 2AFC clinical observer study. The quantitative assessment showed SUV_{max}, SUV_{mean} and SUV_{peak} values were all closely matched to those of the real lesions they were modelled on and the clinical observer study showed the technique could generate lesions that were indistinguishable from the real lesions.

Unlike the physical phantom, the simulated spheres inserted into the NEMA phantom used for validation of the $D710_{simu}$ were wall-less spheres of water. The spheres in the physical NEMA phantom had vacuum formed plastic walls with a mean diameter of 1.05 ± 0.09 mm as derived from weight measurements. Since there is radioactivity in the background compartment of the NEMA phantom, the walls will create a gap with no radioactivity between the sphere and background. This results in a reduction of the radioactivity measured in the sphere which becomes greater as the sphere diameter is decreased. This is because the wall thickness becomes more significant compared to the diameter of the sphere and the partial volume effect causes averaging of the radioactivity in the sphere and the non-active space. To estimate the extent of the systematic bias due to the presence of the walls the method described by Hofheinz *et al* can be used [155]. Assuming an isotropic PSF with FWHM of 7.1 mm (from the average spatial resolution measured at 10 cm for the VPFX reconstruction) and a sphere to background ratio of 5.3, the percentage difference in predicted recovery curves for the maximum voxel value can be estimated as −0.6 % for the 17 mm sphere, −2.4 % for the 13 mm sphere and −5.0 % for the 10 mm sphere. The larger spheres show no bias. A similar pattern would be expected for the mean activity concentration measurements as the thresholds are based on the maximum voxel value.

Stute *et al* also used a projection-domain method to insert 14 large lung lesions simulated using a model of the Philips GEMINI GXL into a real patient scan [145]. The purpose of their study was to generate a small number of PET images for assessment of segmentation

techniques where the true volumes of the lesions were known. Instead, the scanner model used for simulations in this thesis were of a modern PET system using LYSO crystals allowing for assessment of the latest acquisition and reconstruction methods including TOF. The voxel size used to define the lesions is much smaller ($1\,mm^3$, compared to $64\,mm^3$) which keeps the finer detail present in the lesions and doesn't introduce unrealistic stepped edges in the reconstructed images. This also allows much smaller lesions to be generated which are of particular interest and clinical relevance in assessing new PET technologies. The methodology proposed here for generating the voxelised density phantom for input into GATE incorporates a method to automatically convert CT numbers into mass density and elemental weights as derived by Schneider *et al* [144]. This avoids the need for segmentation of healthy tissues in the real patient scan(s), as was performed in the work by by Stute *et al* [145], which would be too time consuming for this work where individual lesions are inserted into many different real patient scans rather than a single scan.

Clinical PET scans are acquired over 3 minutes and so real lesions are subject to respiratory motion and the PET uptake may be blurred [156]. This is not accounted for in the lesions simulated in this work. GATE has the ability to split simulations into time slices so data could be simulated for a specified time at each position of the respiratory cycle. Therefore, the described lesion simulation technique using GATE could be extended to include respiratory motion models or real patient respiratory signals [157], but this would require the generation of voxelised activity and density phantoms with the lesion positioned for the different points in the respiratory cycle.

The current methodology was designed to be used with the manufacturer software in order to evaluate the effects of commercially available reconstructions. However, the advantage of using GATE for the simulations is that the output data can also be used in open-source software, such as CASToR [158], in situations where a more generalised solution is required that is not limited to the specific manufacturer software.

Only a single reviewer performed the clinical observer study due to limited access to the MIM Encore software. Results would be more robust if 2 or more observers independently reviewed the datasets allowing an assessment of consensus between observer choice and confidence in

identifying the simulated lesions for the ten patients. The review also only included the PET as the synthetic lesions in the CT were clearly unrealistic, whereas in clinical practice the reviewer would assess both the lesion morphology on CT and the ^{18}F-FDG uptake on the PET.

The voxelised phantoms used for defining the simulated lesions in this work were derived from real lesions in patient scans and used homogeneous uptake. It would be impractical to use this method for generating larger sets of simulated lesions, therefore the next stage of this work aims to develop an automated method to create synthetic lesions based on typical characteristics seen in real lesions and incorporate heterogeneous uptake.

3.5 Conclusions

A model of the clinical GE Discovery 710 PET scanner has been designed and validated against the real PET system for performing accurate simulations of phantoms and lesions. Further, a technique for insertion of the simulated lesions into real patient PET data in projection space has been developed. This allows for the generation of PET datasets combining the physiological and anatomical variation observed in real patients with realistic simulated tumours where the size, shape and uptake can be specified. These datasets will be used to investigate the impact of new PET developments on image quality and quantitative measures in this thesis through the use of virtual imaging trials.

Chapter 4

Generation and Validation of PET Images of Patients with Realistic Simulated Solitary Pulmonary Nodules

4.1 Introduction

Chapter 3 describes the validation of the PET scanner model used for simulation of lesions and the process of insertion into real patient PET data in projection space. This technique was shown to be able to generate PET images with visual and quantitative characteristics similar to real pulmonary lesions. To ensure that PET images created using this methodology can reliably be used for performing clinically relevant tasks in place of real PET data and a real PET imaging system in the context of virtual imaging trials, the data needs to be carefully validated [131]. In this thesis, the chosen method to perform this validation was to generate PET images for a virtual population of patients consisting of simulated lesions with characteristics replicating a population of real patients with known solitary pulmonary nodules (SPNs).

^{18}F-FDG-PET imaging is used as a non-invasive way to help differentiate between malignant and benign SPNs. SPNs are focal areas of abnormal tissue (lesions) ≤ 3 cm diameter in the lung and are often found incidentally on chest x-ray or CT or through lung cancer screening

programs [159]. These nodules can arise from several conditions that may be malignant or benign such as neoplasms, infection, inflammation and, vascular and congenital abnormalities [159]. Cells in malignant nodules grow unregulated invading nearby healthy tissues and can metastasize to distant sites via the blood and lymph system. Benign nodules are non-cancerous, tend to grow slowly, and unlike malignant lesions, they do not metastasize or invade surrounding tissues. In the management of patients with SPNs, it is important to determine as early as possible whether the nodule is malignant or benign as if it is lung cancer one of the main factors determining prognosis will be the stage at presentation.

In the UK, radiologically indeterminate SPNs are managed according to the British Thoracic Society (BTS) guidelines which recommend the use of clinical prediction models [7]. Initial CT-based assessment for patients with an SPN involves determining the malignancy risk using the Brock risk prediction tool [8]. Patients with a risk $\geq$ 10 % undergo ^{18}F-FDG-PET-CT imaging to help further categorise the risk of malignancy using the Herder clinical prediction model [9]. The resultant risk is then used to help guide the patient management as shown in Table 4.1.

Risk of Malignancy	**Patient Management**
< 10 %	CT surveillance
10 % to 70 %	Image-guided biopsy or, depending on individual risk and patient preference, excision biopsy or CT surveillance
> 70 %	Surgical resection where appropriate or non-surgical treatment

Table 4.1 Risk Stratification of SPNs as determined from the Herder clinical prediction model [9].

In this context, location of the SPN is already known from the previous CT or x-ray and PET imaging is used to help in determining the risk of malignancy and potential spread. The risk of malignancy is based on whether the ^{18}F-FDG uptake within the SPN is greater than the lung background (detection), liver or mediastinal blood pool (MBP) activity (estimation) making it a joint detection-estimation task. In a meta-analysis Gould *et al* reported ^{18}F-FDG-PET to be 93.9 % sensitive and 85.8 % specific at identifying malignant SPNs [11], however an increase in

glucose metabolism is also seen in inflammatory processes resulting in false-positive results for benign lesions [160].

Novel PET reconstruction algorithms that incorporate point spread function (PSF) modelling to recover uptake due to loss of resolution, (Section 2.1.5), have been shown to improve signal-to-noise ratio (SNR) and thus lesion detection in oncology applications [161, 66]. However, there is also a significant impact on quantitative measures of tumour uptake, such as standardised uptake value (SUV), and this effect is dependent on lesion size and tumour-to-background ratio [162, 163]. Additionally, inappropriate selection of reconstruction parameters could lead to underestimation of lesion uptake due to over-smoothing or overestimation from the appearance of edge artefacts [68]. Use of these novel algorithms has been shown to have minimal effect on uptake measurements for reference tissues such as the liver or mediastinal blood pool (MBP) [163, 164] so clinical assessments involving comparison of tumour uptake to these reference tissues will also be impacted. This is the case for the characterisation of SPNs and so moving to these novel algorithms could potentially increase the number of true-positive results for malignant lesions but at the same time increase the number of false-positives i.e., making PET more sensitive, but less specific for identifying malignant lesions. This could lead to inappropriate changes in patient management if following guidance based on published data from patients who underwent PET with older technology. It is important therefore to have a proper understanding of the effect of novel reconstruction algorithms on measurements of lesion uptake particularly in small lesions.

In the original study performed by Herder *et al*, derivation of the clinical prediction model used PET scans acquired between 1997 and 2001 on an older PET-only bismuth germanate (BGO) system acquiring in 2D mode. Since its publication, PET technology has evolved significantly, firstly with the wide adoption of hybrid PET-CT systems whereby the CT is used for attenuation correction and then use of 3D acquisition and reconstruction techniques. Subsequent improvements in hardware and software have seen the introduction of time-of-flight (TOF)-PET in systems with lutetium-based scintillation crystals and the use of novel reconstruction algorithms incorporating PSF modelling. A later study by Al-Ameri *et al* retrospectively applied

the Herder clinical prediction model to patients who were identified through the lung cancer multidisciplinary team between 2008 and 2013 and had ^{18}F-FDG-PET-CT to assess an SPN [10]. Histopathology and follow up imaging were used to confirm diagnosis. The authors found the Herder model had a high accuracy for discriminating between malignant and benign lesions with an area under the ROC curve (AUC) of 0.92, 95 % CI [0.88, 0.97] which was almost identical to the original study by Herder *et al*, with an AUC of 0.92, 95 % CI [0.87, 0.97] [9]. Details of the PET imaging parameters were not provided in the published paper but, based on the date range of the study data, the PET would have been a combination of 2D and 3D PET-CT and from 2010 onwards TOF-PET (data from UK PET Core Lab). This pre-dated the wide-spread availability of novel reconstruction algorithms incorporating PSF and in particular the proprietary Q.Clear reconstruction algorithm which was released by GE in 2014. Murphy *et al* investigated the effect of the novel Q.Clear reconstruction algorithm on the assessment of malignancy risk in SPNs for a cohort of patients scanned between 2013 and 2017 at King's College London & Guy's and St Thomas' PET Centre [165]. 97 patients who underwent ^{18}F-FDG-PET-CT for assessment of an SPN were identified retrospectively from the local Hospital database. 75 (77 %) of the SPNs were malignant, and 22 (23 %) benign as determined by histological confirmation (biopsy or surgical resection) or follow up imaging. The raw PET data for the 97 patients were retrospectively reconstructed using the Q.Clear algorithm ($\beta = 400$) which includes PSF modelling and a regularisation term to control noise in the reconstructed image (Section 2.1.4). The accuracy in risk of malignancy determined using the novel reconstruction was then compared to the risk using the standard clinical OS-EM reconstruction with TOF (VPFX with 2 iterations, 24 subsets and 6.4 mm Gaussian filter).

This chapter describes the generation of PET images for a virtual population of patients with SPNs designed to replicate the real population from the study by Murphy *et al* [165]. The methodology described in Chapter 3 was used to develop and insert simulated SPNs with the same characteristics as those from the original study into raw PET data from 'healthy' patients in the projection-domain. Analysis of the PET data generated from the virtual population followed that used for the real population. Validation of the virtual population was then performed through

comparison of quantitative measures and outcomes in terms of receiver operating characteristic (ROC) curves and area under the ROC curve (AUC) for both populations. An automated method of producing the synthetic lesions for input into the Monte Carlo model is also described.

4.2 Methods

Anonymised raw PET data and reconstructed ^{18}F-FDG-PET-CT scans from 97 patients were randomly collected from clinical scans performed at the King's College London & Guy's and St Thomas' PET Centre (IRAS project ID: 251611). To ensure anonymity of the data, the raw PET and CT data were anonymised by clinical staff on the scanner prior to use in this work. The only inclusion criteria were that PET scans were performed on a GE Discovery 710 scanner at St Thomas' Hospital following the standard clinical protocol, and that the raw PET data was available. Since there are two GE Discovery 710 scanners, data was allowed from both systems. It should be noted that selection of patient scans with physiological uptake of ^{18}F-FDG on a half body PET-CT scan was blinded to all clinical information so the patient condition was unknown, however for the purposes of this thesis, the patient scans will be described as 'healthy' to differentiate from the PET-CT scans from patients in the study by Murphy *et al* with a known SPN. Any PET-CT scans that didn't cover at least base-of-brain to mid-thigh or contained significant artefacts that could affect quantification, such as excessive motion or extravasation, were excluded. These scans were then used for insertion of the simulated lesions to produce the PET scans for the virtual population.

All PET-CT scans for the SPN and healthy cohorts were scanned at the King's College London & Guy's and St Thomas' PET Centre on the GE Discovery 710 PET systems . Patient preparation and scan acquisition followed the standard clinical protocol for both patient cohorts. Patients were fasted for 6 hours and blood glucose was checked prior to injection ensuring blood glucose was < 11 mmol/L. The patient weight was measured on the day of the scan and used for normalisation of SUVs. Injected activities were fixed rather than weight based with mean ± standard deviation (SD) injected activities of 328.6 ± 25.7 MBq and 323.2 ± 23.4 MBq

for the SPN and healthy cohorts respectively. The mean uptake time for the SPN cohort was longer than that for the healthy cohort (83.6 ± 10.0 min versus 70.2 ± 12.7 min) due to a change in the local protocol from 90 minutes to 60 minutes during the time the data was acquired. In the SPN cohort 11 patients were scanned at 60 minutes (60.6 ± 0.9 min, range 59 to 62) and 86 patients were scanned at 90 minutes (86.5 ± 6.0 min, range 75 to 99). This compares to 69 patients scanned at 60 minutes (62.9 ± 3.4 min, range 57 to 72) and 28 patients scanned at 90 minutes (88.0 ± 8.8 min, range 75 to 106) for the healthy cohort.

Both PET and CT images were acquired from mid-thigh or toes to base-of-brain or vertex under shallow inspiration. The CT used for attenuation correction and localisation was acquired first using 140 kV p, variable mA s, 0.5 s rotation time and 1.375 pitch and reconstructed using the standard kernel with ASiR blending of 40 %. The PET was acquired using 3 min per bed position and 11-slice (23 %) overlap. PET data were reconstructed using the standard clinical reconstruction (VPFX) and the novel Q.Clear reconstruction (QCFX) with the parameters shown in Table 4.2.

Parameter	Reconstruction Name	
	VPFX	**QCFX**
Algorithm Used	OS-EM	BPL
Matrix Size	256×256	256×256
Pixel Size (mm)	2.73	2.73
Iterations	2	N/A
Subsets	24	N/A
Post Filter	Gaussian	None
	FWHM = 6.4 mm	
z-axis Filter	4 mm	None
Strength (β) of the Penalising Factor	N/A	400
Time-of-Flight (TOF) Included	Yes	Yes
Point Spread Function (PSF)	No	Yes
Modelling Included		

Table 4.2 Reconstruction algorithm and parameters used for the standard clinical reconstruction (VPFX) and the novel Q.Clear reconstruction (QCFX). OS-EM = ordered subsets-expectation maximisation, BPL = Bayesian penalised likelihood. Parameters that are not applicable for the particular reconstruction are denoted as N/A.

4.2.1 Simulation of the Solitary Pulmonary Nodules

To generate a large set of PET images for the virtual population, simulated lesions with characteristics similar to the original study were required for insertion into the healthy PET scans. To achieve this, each patient in the healthy cohort was randomly assigned the radiological and clinical characteristics from a patient in the SPN cohort. The designated radiological and clinical features for the simulated lesions included the diameter measured on CT, the maximum ^{18}F-FDG uptake, lobar location and whether the SPN had a spiculated appearance. The simulated lesions were designated as being malignant or benign based on the clinical outcomes measured for the

real SPNs in the assigned patients from the SPN cohort. Real malignant SPNs were confirmed by histopathology, either biopsy or surgical resection, and benign nodules by follow up imaging.

4.2.1.1 Automatic Generation of Synthetic Lesions

An automated method for creating the voxelised activity and density phantoms containing the synthetic SPNs to use as input for the GATE simulation was developed in MATLAB (The MathWorks, Inc, version R2020b). The synthetic lesions were generated using a 3D tumour growth model based on the model described by Poleszczuk *et al* [166] and modified for the purpose of this project.

The tumour growth model starts with a logical 3D array where all elements are set to 'false' to indicate they are unoccupied. Elements in the array are then set to 'true' when occupied by a cell. Starting with an initial element in the array occupied by a cancer stem cell (CSC), the model locates the unoccupied neighbouring elements and any existing cells undergo symmetric or asymmetric division to produce either another CSC or a non-stem cancer cell (CC) on each iteration of the algorithm. As part of the model, probabilities for symmetric cell division, cell proliferation capacity, migration and spontaneous cell death can be adjusted [167]. Each element in the array was chosen to be equivalent to 50 µm × 50 µm × 50 µm which is approximately 16 times the size of an individual cell assuming a cell diameter of 20 µm and cell density of 10^8 cells cm^{-3} [168]. This size was chosen partly due to limitations on the available memory on the PC used to generate the lesions, but also to keep the time to generate each synthetic lesion reasonable (median time 28 minutes). For a lobular or spiculated appearance the model was adapted to include initialisation with up to eight occupied elements, benign lesions had 1 or 2, spiculated 3 to 8 and malignant 1 to 8. Placement of the initialisation points used random x, y and z co-ordinates drawn from a normal distribution with mean = central array co-ordinate and standard deviation = 0.1 × lesion diameter. The model was also modified so the simulation stopped once the occupied elements reached the assigned lesion diameter in either the x or y direction.

For the purposes of this project the size and shape of the lesions was of interest rather than growth time or cell composition. Features of malignant and benign SPNs can overlap, but malignant tend to be larger and have irregular or spiculated margins whereas benign lesions tend to have smoother well-defined borders and are spherical in shape [169, 170]. To investigate the impact of varying the probabilities of symmetric cell division, spontaneous cell death, cell migration and cell proliferation capacity on the lesion appearance, 256 $\times$ 10 mm diameter lesions were generated using a single initialisation point with probabilities from 20 % to 80 % in 20 % increments. Figure 4.1 shows four of the 10 mm diameter lesions created with an 80 % probability for a) symmetric cell division, b) spontaneous cell death, c) cell migration and d) proliferation. All other probabilities were fixed at 20 %. It should be noted that the probabilities were not independent of one another.

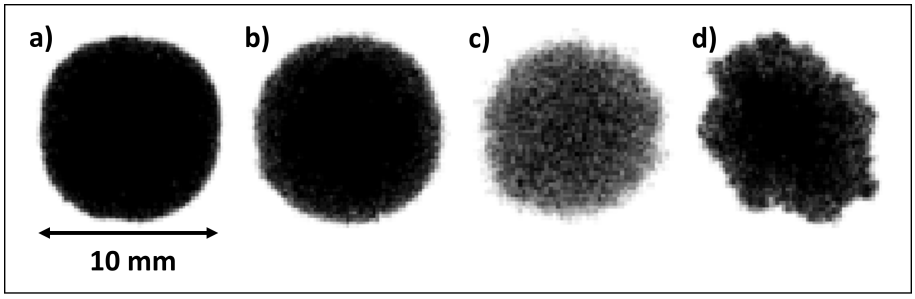

Fig. 4.1 Examples of 10 mm diameter synthetic lesions generated using the tumour growth model with a single initialisation point. In the above examples, the probability under investigation was set at 80 % while all others were fixed at 20 %. Probabilities under investigation were a) symmetric cell division, b) spontaneous cell death c) cell migration c) and d) cell proliferation.

Increased probabilities for symmetric cell division and spontaneous cell death were associated with denser lesions with smooth margins Figure 4.1 a) and b). High probabilities in both created unrealistic lesion appearances so probabilities were restricted to $\leq$ 50 % for all lesions. The probability of cell migration was inversely related to lesion density Figure 4.1 c). Increased probability for cell proliferation generated lesions with increasingly irregular shapes and rough margins more typical of malignant SPNs Figure 4.1 d). Using this data, probabilities were randomly assigned from the ranges given in Table 4.3 to generate the 97 lesions for insertion.

Probability Description	Spiculated Range	Malignant Range	Benign Range
Symmetric cell division	10 % to 20 %	20 % to 40 %	30 % to 50 %
Spontaneous cell death	10 % to 20 %	20 % to 40 %	30 % to 50 %
Cell migration	70 % to 80 %	60 % to 80 %	40 % to 60 %
Cell proliferation	70 % to 80 %	50 % to 70 %	30 % to 50 %

Table 4.3 Probability ranges used in the tumour growth model to generate lesions with benign and malignant characteristics.

For the voxelised phantoms, the logical array was converted to numerical values and resampled to 1 mm^3 voxels using linear interpolation to match the voxel size for the resampled CT from the healthy patients. The resultant array contained values from 0 to 1 equivalent to the fraction of occupied elements within that volume. The steps below were then used to derive the radioactivity and material properties to assign to the voxels.

The target uptake for the simulated lesion was derived from the SUV_{max} for the real lesion in the SPN patient measured using the clinical reconstruction (SUV_{SPN}). Since the uptake in the lung for the SPN and healthy patients were potentially different, the SUV_{mean} measured in the lung for the SPN patient (SUV_{lung}) was subtracted from the SUV_{SPN} to give the relative SUV ($SUV_{(SPN-lung)}$). It was assumed that the uptake in the lung was uniform. Estimation of the true uptake in the real lesion (SUV_{pvc}) was then performed by applying a partial volume correction (PVC) to the $SUV_{(SPN-lung)}$ using the lesion diameter and a recovery coefficient (RC) curve measured for the clinical reconstruction (VPFX) on the $D710_{real}$ using the technique described by Srinivas *et al* [154]. Finally, the $SUV_{(pvc)}$ was converted into $kBq.ml^{-1}$ to give the target activity concentration of the lesion to be inserted into the healthy patient ($C_{Healthy}$):

$$C_{Healthy}\,(kBq.ml^{-1}) = SUV_{(pvc)} \cdot \left(\frac{A_t\,(MBq)}{W\,(kg)} \right)$$

where A_t was the injected activity decay corrected to the acquisition time t, and W the healthy patient weight from the DICOM header. The activity concentration was converted from $kBq.ml^{-1}$

to $Bq.mm^{-3}$ and each voxel in the synthetic lesion array was multiplied by this value to generate the voxelised activity phantom. Since the voxels in the synthetic lesion array contained values from 0 to 1 depending on the density of cells within that volume, activities towards the edges of the lesion with lower cell densities contained lower activities as would be seen in a real lesion.

The HU for the synthetic lesions were assigned randomly from the ranges for the pre-enhancement HU for malignant (16.5 ± 14.4) and benign (19.1 ± 21.2) lung nodules measured by Swensen *et al* [153] assuming a normal distribution and each voxel in the synthetic lesion array was multiplied by this value. To create the final voxelised density phantom, the synthetic lesion array was inserted into the resampled CT for the healthy patient by replacing the existing voxel values at the location of the lesion. Metadata from the SPN study included a record of the nodule location according to the lobes of the lungs as shown in Figure 4.2. This was used to guide the manual selection of the central point for placement of the simulated lesion in the healthy patient CT using MIM Encore (MIM Software Inc, version 7.1.5).

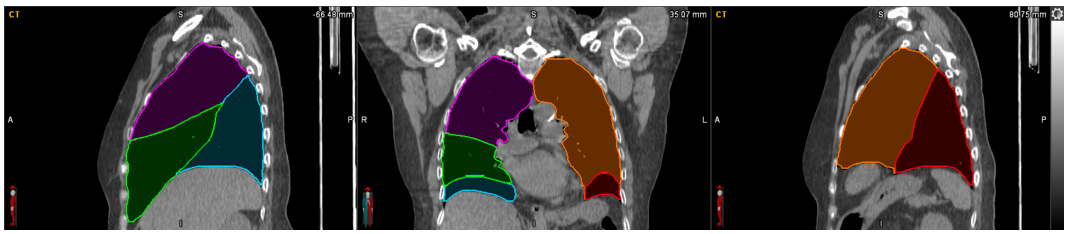

Fig. 4.2 Sagittal, coronal and axial CT slices segmented to show the lobes of the lungs: purple = right upper lobe (RUL), green = right middle lobe (RML), blue = right lower lobe (RLL), orange = left upper lobe (LUL), red = left lower lobe (LLL).

The CT slice number and x, y slice locations of the central point of the lesion were then used in the MATLAB script to derive the offsets for automatic placement of the synthetic lesions in the resampled healthy patient CT to create the final voxelised phantom. For the activity phantom, all voxels outside the lesion were set to zero. Figure 4.3 shows an example of the voxelised density and activity phantoms for one of the patients in the study with a 22 mm synthetic lesion in the left lung.

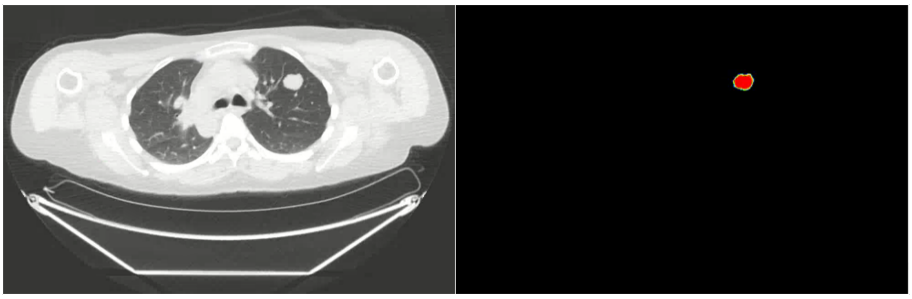

Fig. 4.3 Axial slice of the density (left) and activity (right) voxelised phantoms showing a 22 mm synthetic lesion used as the input for GATE.

Using these voxelised phantoms as input, simulated acquisitions were performed in GATE matching the location and duration of the equivalent bed positions acquired for the healthy patient scans. Processing of the output PET data from the simulation followed the steps described in Chapter 3, Section 3.2.1.

4.2.2 Quantitative Evaluation of FDG Uptake in the Reference Tissues and SPNs

Analysis of the reconstructed PET images containing the simulated lesions was performed in MIM Encore and followed that used in the original study. The consensus results from the image analysis performed by the consultant radiologist and clinical fellow in the original study were used as the reference standard for comparison of the simulated lesion measurements. Image analysis for the simulated lesions was conducted by this thesis author, a Clinical Scientist with 12 years experience in PET including the review and analysis of PET images.

In the original study by Murphy *et al* [165], the SUV_{max} in the liver and MBP were measured for each patient using large manual 2D ROIs placed in the arch of the aorta and right lobe of the liver as shown in Figure 4.4.

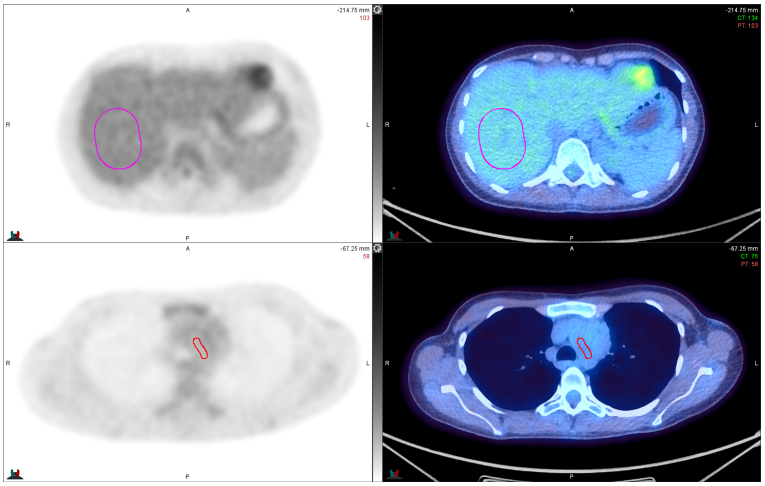

Fig. 4.4 Axial PET and fused PET-CT slices showing the placement of the 2D ROIs used to measure the SUV_{max} in the liver (top) and MBP (bottom).

The PERCIST guidelines however, recommend measurement of the blood pool in the descending aorta and a smaller volume in the right lobe of the liver [2]. Therefore, additional measurements of the SUV_{max} in the liver and MBP were made for the SPN cohort using a 3 cm diameter spherical VOI in the right lobe of the liver and a 1 cm diameter by 2 cm length cylinder VOI placed in the descending thoracic aorta as described in the practical PERCIST guidelines [87]. For consistency, the PERCIST volumes were used to measure the SUV_{max} in the liver and MBP in both the SPN and healthy cohorts for determination of the risk of malignancy. The SUV_{max} for the lung tissue was measured using a 2D ROI placed centrally in the lung in which the lesion was to be inserted, avoiding any existing structures such as cardiac tissue, lung walls or bronchi. The SUV tool in MIM was used for measuring the SUV_{max} in the nodule. This involved loosely placing a 3D sphere over the nodule to include all the uptake, the SUV tool then automatically marks the location of the maximum voxel and reports the SUV_{max}.

4.2.2.1 Derivation of the Malignancy Risk using PET

The clinical prediction model by Swensen *et al* [171] uses a probability estimate to determine the risk of malignancy (P_s) for SPNs based on clinical and radiographic factors according to the

equation:

$$P_s = \frac{1}{(1+e^{-x})}$$

where x is determined from the weighted contribution of the clinical and radiographic factors defined in Table 4.4:

$$x = -6.8272 + 0.0391(age) + 0.7917(cigarettes) + 1.3388(cancer) + 0.1274(diameter) + 1.0407(spiculation) + 0.7838(upper)$$

Clinical or Radiological Factor	Definition
Age	Patient age in years
Cigarettes	Current or former smoker
Cancer	History of extra thoracic cancer
Diameter	Diameter of SPN measured on CT (mm)
Spiculation	Spiculation
Upper	SPN located in upper lobe

Table 4.4 Definition of the clinical and radiographic factors used to determine risk of malignancy for solitary pulmonary nodules (SPNs) in the Swensen clinical prediction model [171].

The factors used for the healthy cohort in this work, as summarised in Table 4.5, were assigned from the clinical and nodule characteristics for the SPN population from Murphy *et al* [165]. The Swensen scores were therefore identical for the two patient populations.

Clinical or Radiological Characteristic	Benign (n=22)	Malignant (n=75)	All Patients (N=97)
Age (years)			
Mean ± SD:	70.9 ± 9.4	68.9 ± 10.0	69.3 ± 9.9
Range:	53-85	32-87	32-87
Cigarettes (%)	72.7	86.7	83.5
Cancer (%)	22.7	29.3	27.8
Diameter (mm)			
Mean ± SD:	11.4 ± 5.3	17.8 ± 5.5	16.4 ± 6.1
Median:	10.0	18.0	15.0
Range:	5-25	6-30	5-30
Spiculation (%)	22.7	65.3	55.7
Upper (%)	54.5	52.0	52.6

Table 4.5 Summary of clinical & radiological characteristics for malignant and benign lesions in the SPN dataset [165].

The Herder risk model [9] refines the Swensen clinical prediction model by incorporating the degree of ^{18}F-FDG uptake measured on PET to determine the probably of malignancy (P_h):

$$P_h = \frac{1}{(1+e^{-x})}$$

where x combines the Swensen score (P_s) with a weighting factor based on the degree of ^{18}F-FDG uptake in the nodule using a four-point ordinal scale from absent to intense:

$$x = -4.739 + 3.691(P_s) + 2.322(faint) + 4.617(moderate) + 4.771(intense)$$

The ^{18}F-FDG uptake for the nodules was graded using the SUV_{max} relative to the lung, liver and mediastinal blood pool (MBP) using the definitions provided by Murphy *et al* [165] and given in Table 4.6.

Score	^{18}F-FDG Uptake	Definition in terms of Nodule standardised uptake value (SUV)
1	Absent	= lung SUV_{max}
2	Faint	> lung SUV_{max} but $\leq$ MBP SUV_{max}
3	Moderate	> MBP SUV_{max} and $\leq 2 \times$ liver SUV_{max}
4	Intense	$> 2 \times$ liver SUV_{max}

Table 4.6 Ordinal scale for classification of nodule ^{18}F-FDG uptake compared to lung, liver and mediastinal blood pool (MBP) measured on PET. Table reproduced from Murphy *et al* [165].

4.2.2.2 Statistical Analysis of Results for Healthy and SPN Cohorts

Results for the SUV_{max} measurements and calculated Herder risk scores are given as mean ± standard deviation (SD). The independent two-sample *t*-test was used for comparison of continuous variables between the two patient cohorts and the dependent two-sample *t*-test for comparison between reconstructions within the same patient cohort. A *P* value <.05 was considered significant. Linear weighted Cohen's Kappa (κ) was used to assess the agreement in the ordinal classification of nodule ^{18}F-FDG uptake between the healthy patients with simulated lesions and the SPN patients with real SPNs. Levels of agreement were defined as: 0.21-0.4 = fair, 0.41-0.60 = moderate, 0.61-0.80 = good and 0.81-1.00 = very good [172]. Diagnostic performance was assessed using area under the receiver operating characteristic (ROC) curve (AUC) with 95 % pointwise confidence intervals (CIs) computed using 1000 bootstrap samples. Comparison of AUCs was performed using the non-parametric method for correlated ROC curves described by DeLong *et al* [173]. Sensitivity (Se), specificity (Sp), positive predictive values (PPV) and negative predictive values (NPV) were determined for pre-defined Herder score cut-off points of 10, 50 and 70 %. These cut-offs were chosen to match those used in the original analysis by Murphy *et al* [165]. The BTS guidelines recommend further investigation or biopsy for patients with > 10 % risk of malignancy and invasive intervention such as excision or surgical resection for patients with > 70 % risk of malignancy [7]. CIs for sensitivity and

specificity are Clopper-Pearson exact intervals [174] and standard logit intervals [175] for PPVs and NPVs. Statistical analysis was performed in MATLAB version R2022b.

4.3 Results

4.3.1 Quantitative Evaluation of FDG Uptake in the Reference Tissues

4.3.1.1 Comparison of Regions used for Reference Tissue Measurements

Table 4.7 shows the mean ± SD SUV_{max} measured in the liver and MBP for the patients in the SPN cohort using the large 2D ROIs and the PERCIST VOIs. Results are shown for both the 3D OS-EM with TOF (VPFX) and BPL with TOF and PSF (QCFX) reconstructions.

	Liver		**MBP**	
	VPFX	**QCFX**	**VPFX**	**QCFX**
2D ROIs	3.1 ± 0.6	3.0 ± 0.6	1.8 ± 0.4	1.7 ± 0.4
PERCIST VOIs	3.0 ± 0.6	2.8 ± 0.5	2.0 ± 0.4	2.0 ± 0.4
P values	*P*<.001	*P*<.001	*P*<.001	*P*<.001

Table 4.7 Mean ± SD SUV_{max} measured for the liver and mediastinal blood pool (MBP) in the SPN patient cohort using manual 2D ROIs and PERCIST VOIs.

These results show there were significant differences in the SUV_{max} for the SPN cohort when measured using the large manual 2D ROIs compared to the PERCIST VOIs in both the liver and MBP and for both reconstructions with all *P* values <.001. To determine whether this difference in SUV_{max} for the reference tissues had an impact on the results, the ^{18}F-FDG uptake in the nodules was classified for both sets of reference tissue measurements following the scheme described in Table 4.6 and using the consensus SUV_{max} for the real SPNs. Results showed very good agreement between the classification of nodule uptake derived using the SUV_{max} measured using 2D ROIs and using the PERCIST VOIs with κ=0.99, 95 % CI [0.95, 1.00] for VPFX and κ=0.93, 95 % CI [0.82, 1.00] for QCFX reconstructions. For context, this was similar to the

inter-observer agreement for the classification of nodule ^{18}F-FDG uptake derived from individual measurements of the liver and MBP by the two clinical reviewers using the manual 2D ROIs with κ=0.93, 95 % CI [0.84, 1.00] for VPFX and κ=0.96, 95 % CI [0.89, 1.0] for QCFX [165]. All further analysis of uptake for both patient cohorts used the reference tissue SUV_{max} measured using the PERCIST VOIs.

4.3.1.2 Comparison of Reference Tissue Uptake in the SPN and Healthy Patient Cohorts

Box plots of the SUV_{max} measured in the reference tissues for real patients with SPNs and healthy patients with simulated SPNs are shown in Figure 4.5. There was no significant difference between the SUV_{max} measured for the liver, lung or MBP in the two cohorts using either VPFX (P=.79, P=.88 and P=.73) or QCFX (P=.98, P=.95 and P=.78) reconstructions. These results show the cohort of patients used for insertion of the simulated lesions have comparable physiological uptake in the reference tissues to the original SPN patient cohort with real SPNs.

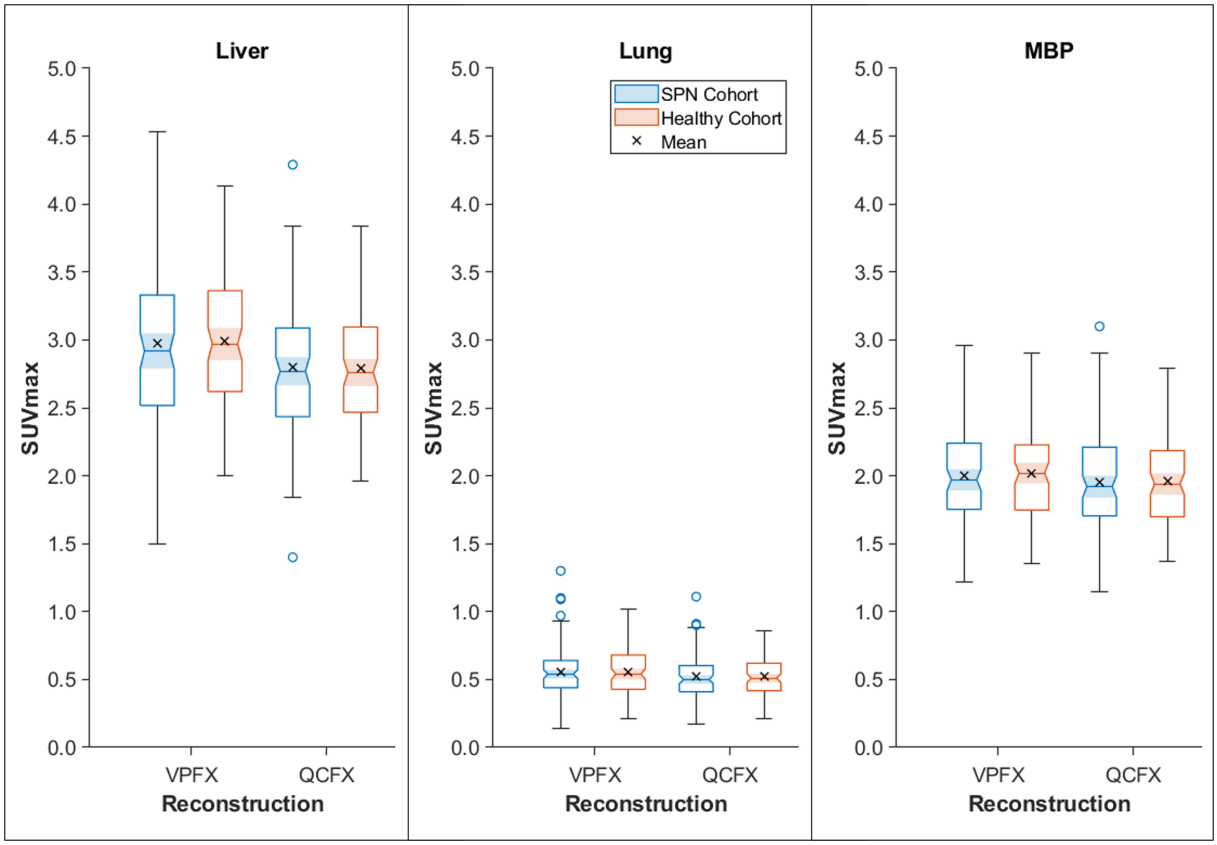

Fig. 4.5 Box plots of the SUV_{max} measured for the reference tissues in real patients with an SPN and healthy patients with a simulated SPN for VPFX and QCFX reconstructions. The × denotes the mean SUV_{max} for each group.

4.3.2 Quantitative Evaluation of FDG Uptake in the SPNs

4.3.2.1 Comparison of SUV Measured for the Real SPNs and Simulated Lesions

The Bland-Altman plot in Figure 4.6 shows the difference between the SUV_{max} measured for the simulated lesion and the assigned real SPN using the VPFX reconstruction. The differences are plotted as a function of the SUV_{max} measured for the real SPN on the VPFX reconstruction. The green dotted line represents the mean difference and the red dotted lines show ±1.96 SD limits of agreement. This plot indicates the methodology for generating the simulated lesions could reproduce the SUVs from the real SPNs with a mean difference in SUV_{max} of 0.03, 95 % CI [-1.0, 1.1]. However, as shown by the black linear regression line, as the real nodule SUV_{max} increases there is a bias in the simulated lesion SUV_{max} which goes from negative to positive.

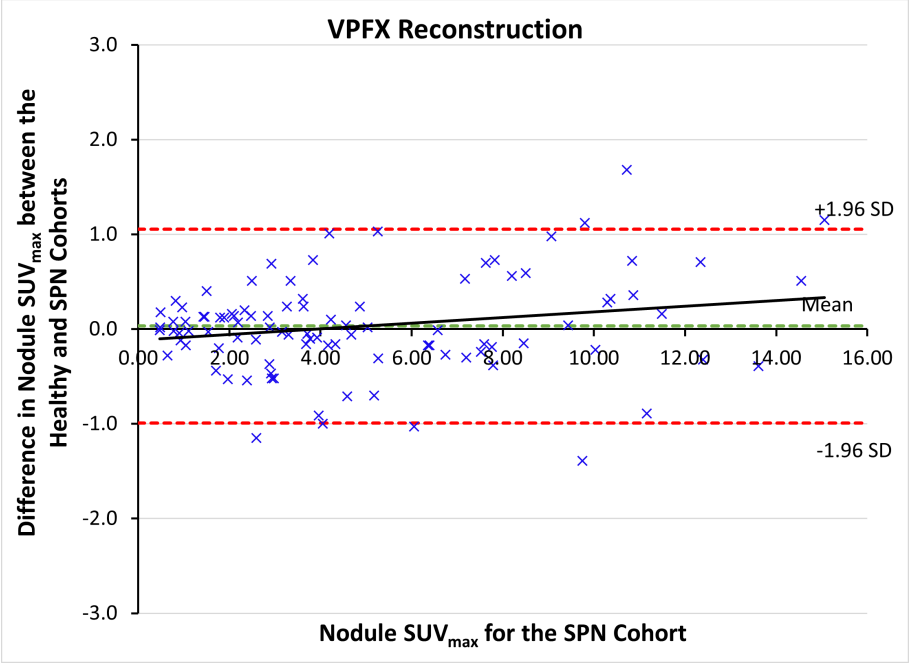

Fig. 4.6 Bland-Altman plot of the differences between the SUV_{max} measured for the real and simulated nodules using the VPFX reconstruction, plotted against the real nodule SUV_{max}. Dotted red lines show ±1.96 SD limits of agreement.

A similar bias was also seen when plotting the differences in SUV_{max} against the diameter of the real SPN as measured on CT, Figure 4.7. This suggests that the partial volume correction used for estimating the true uptake in the real lesions underestimates for smaller lesions and overestimates for larger lesions.

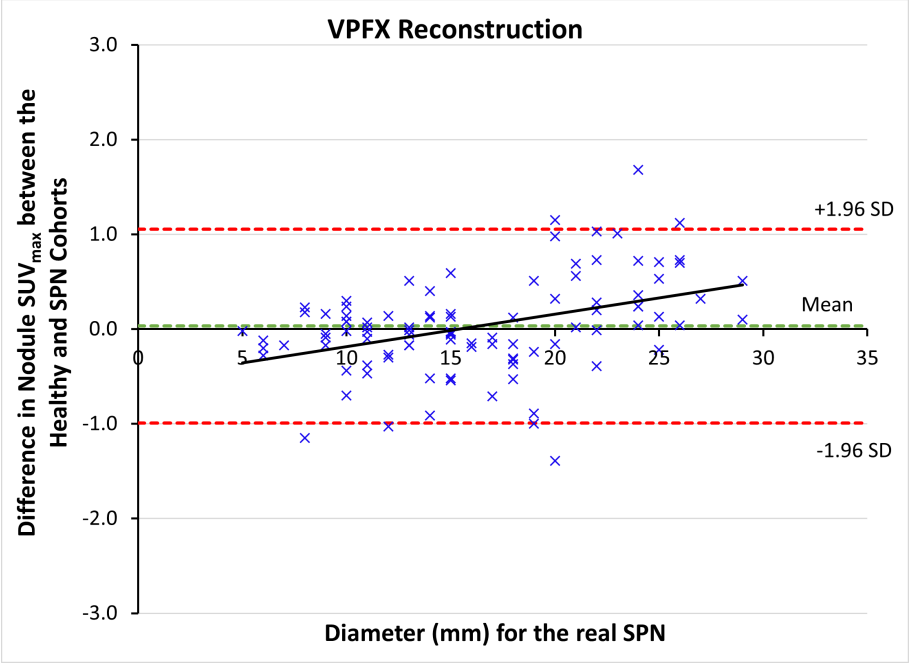

Fig. 4.7 Bland-Altman plot of the differences between the SUV_{max} measured for the real and simulated nodules using the VPFX reconstruction, plotted against the real nodule diameter. Dotted red lines show ±1.96 SD limits of agreement.

The plot in Figure 4.8 shows the SUV_{max} measured for the individual lesions on the QCFX reconstruction versus the SUV_{max} measured on the VPFX reconstruction. As shown by the linear regression lines, the change in SUV_{max} from VPFX to QCFX for the simulated lesions in the healthy cohort (orange dotted line) followed a similar trend to that of the real SPNs (green dotted line).

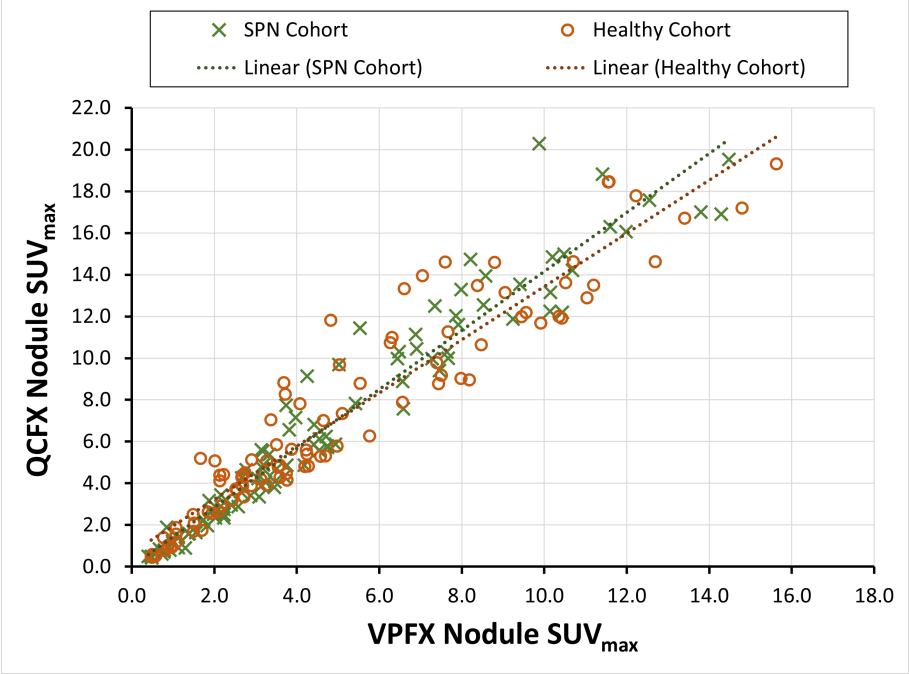

Fig. 4.8 Plot of the SUV_{max} measured for the real and simulated nodules using the QCFX reconstruction against the SUV_{max} measured using the VPFX reconstruction.

Table 4.8 shows the mean ± SD in SUV_{max} for the nodules in the SPN and healthy cohorts for both reconstructions. Results are presented for all nodules, and grouped as malignant and benign with *P* values in the last column. There was no significant difference between the nodule SUV_{max} measured for the SPN and healthy cohorts for either reconstruction across all nodules or when separating between malignant and benign lesions.

All Nodules (N=97)	SPN Cohort	Healthy Cohort	*P* value
VPFX	5.0 ± 3.6	5.0 ± 3.7	.95
QCFX	7.0 ± 5.3	7.0 ± 5.0	.999
Malignant Nodules (n=75)			
VPFX	5.9 ± 3.5	5.9 ± 3.6	.92
QCFX	8.4 ± 5.2	8.3 ± 4.8	.98
Benign Nodules (n=22)			
VPFX	1.8 ± 1.3	1.7 ± 1.3	.90
QCFX	2.5 ± 2.0	2.6 ± 2.1	.90

Table 4.8 Mean ± SD SUV_{max} for nodules in SPN and healthy cohorts measured using the VPFX and QCFX reconstructions.

As expected from the results of the original study, the SUV_{max} measured for the nodules in the SPN cohort using the QCFX reconstruction were significantly higher than that measured using the VPFX reconstruction (P<.001). This result was also seen in the SUV_{max} measurements for the healthy cohort (P<.001). This was also true when separating nodules in both SPN and healthy cohorts into malignant nodules (P<.001, P<.001) and benign nodules (P<.001, P=.002).

Figure 4.9 shows ROC curves for the diagnostic performance of SUV_{max} for predicting malignancy using the VPFX (left) and QCFX (right) reconstructions in the SPN and healthy cohorts. There was no significant difference in the area under the ROC curve (AUC) for the SPN 0.88 [0.79, 0.94] and healthy cohorts 0.90 [0.78, 0.96] when using the VPFX reconstruction (P=.86). There was also no significant difference in the AUC for the SPN 0.86 [0.76, 0.93] and healthy cohorts when using the QCFX reconstruction 0.87 [0.77, 0.94] (P=.92).

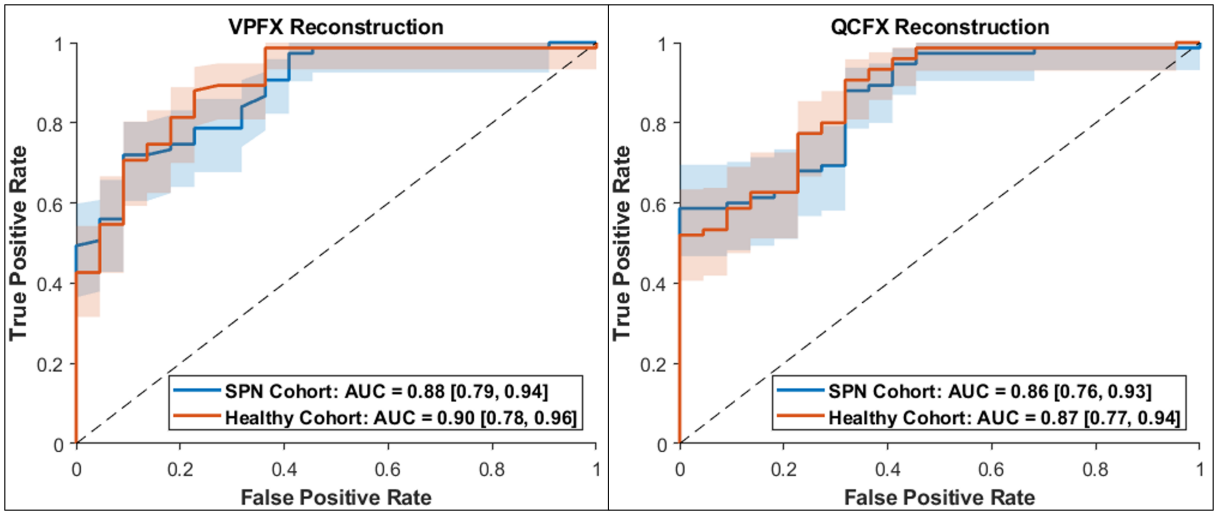

Fig. 4.9 ROC curves for the diagnostic performance of SUV_{max} for predicting malignancy using both VPFX (left) and QCFX (right) reconstructions in the SPN and healthy cohorts. Blue and orange shaded areas show 95 % CIs for the SPN and healthy ROC curves respectively.

4.3.2.2 Comparison of Classification of FDG Uptake for the Real SPNs and Simulated Lesions

Frequencies for the classification of the nodule ^{18}F-FDG uptake in the two cohorts are plotted in Figure 4.10. The number of patients in each group are identical for both the SPN and healthy patient cohorts and for both reconstructions. However, the classification was lower for one patient (moderate rather than intense) and higher for one patient (intense rather than moderate) in the healthy cohort compared to the SPN cohort when using the QCFX reconstruction.

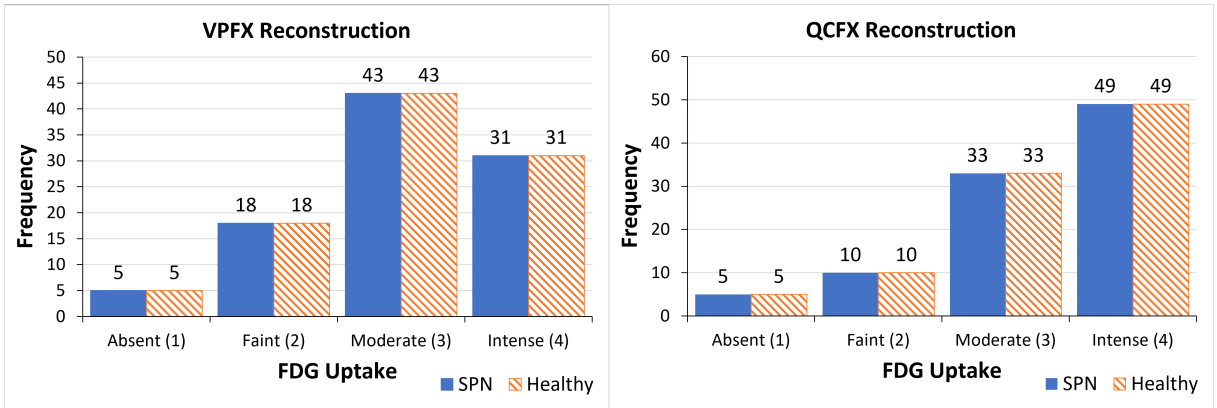

Fig. 4.10 Frequencies of ^{18}F-FDG uptake classification for the SPN and healthy cohorts using VPFX reconstruction (left) and QCFX reconstruction (right).

Agreement for the SPN and healthy cohorts was perfect for the VPFX reconstruction κ=1.00, 95 % CI [1.00, 1.00] and very good for the QCFX reconstruction κ=0.98, 95 % CI [0.92, 1.00]. This compares to the inter-observer agreement for the classification of ^{18}F-FDG uptake reported by Murphy *et al* of κ=0.85, 95 % CI [0.71, 0.99] for VPFX and κ=0.89, 95 % CI [0.74, 1.0].

Results were very similar for the SPN and healthy cohorts with the classification of ^{18}F-FDG uptake increasing in 26 (26.8 %) patients between the VPFX and QCFX reconstructions in both cohorts. This increase in classification was only seen in patients with faint or moderate uptake (scores 2 and 3) on the VPFX reconstruction and the increase was only by one group i.e. from 2 to 3 or 3 to 4. In the SPN cohort 8 (8.2 %) patients (6 malignant, 2 benign) went from faint using VPFX to moderate using QCFX and the corresponding patients with matched clinical and diagnostic characteristics in the healthy cohort also went from faint using VPFX to moderate using QCFX. 18 (18.6 %) patients went from moderate to intense uptake in both the SPN and healthy cohorts, 17 of these (14 malignant, 3 benign) were the corresponding patients with matched clinical and diagnostic characteristics.

4.3.2.3 Comparison of Herder Risk Score for the Real SPNs and Simulated Lesions

Mean ± SD Herder scores for both cohorts are shown in Table 4.9. Since the Swensen scores applied from the initial trial data and the ^{18}F-FDG uptake classification using the VPFX reconstruction were identical for the SPN and healthy cohorts, the Herder scores were also identical. There was no significant difference in Herder risk scores for the SPN and healthy cohorts using the QCFX reconstruction for all nodules (P=.997) or when grouped into malignant (P=.99) or benign nodules (P=.99).

Herder Score (%)		**VPFX**	**QCFX**	***P* value**
SPN Cohort	All nodules	69.2 ± 31.3	73.4 ± 29.2	.002
	Malignant	78.5 ± 23.1	82.5 ± 19.1	.007
	Benign	37.4 ± 35.2	42.2 ± 35.8	.07
Healthy Cohort	All nodules	69.2 ± 31.3	73.4 ± 29.2	.002
	Malignant	78.5 ± 23.1	82.6 ± 19.0	.007
	Benign	37.4 ± 35.2	42.2 ± 35.8	.07

Table 4.9 Mean ± SD Herder scores for the nodules measured on the VPFX and QCFX reconstructions for both SPN and healthy cohorts.

As found in the original SPN study, there was a significant difference in Herder scores between the VPFX and QCFX reconstructions for all nodules driven by the difference in uptake between reconstructions for malignant nodules. This was also seen in the Herder scores for the healthy cohort with *P* values for all nodules and malignant nodules <.05 as given in Table 4.9. In both cohorts, there was no significant difference in Herder scores for the VPFX and QCFX reconstructions for benign nodules (*P*=.07).

Figure 4.11 shows the ROC curves for the diagnostic performance of Herder scores for predicting malignancy in the healthy and SPN cohorts when using the VPFX (left) and QCFX (right) reconstructions. There was no significant difference between the diagnostic performance of the Herder scores for the two cohorts using VPFX (*P*>.999), with an AUC of 0.83, 95 % CI [0.72, 0.91] for both the SPN and healthy cohorts, or for QCFX (*P*=.995) with an AUC of 0.84, 95 % CI [0.73, 0.91] for both cohorts.

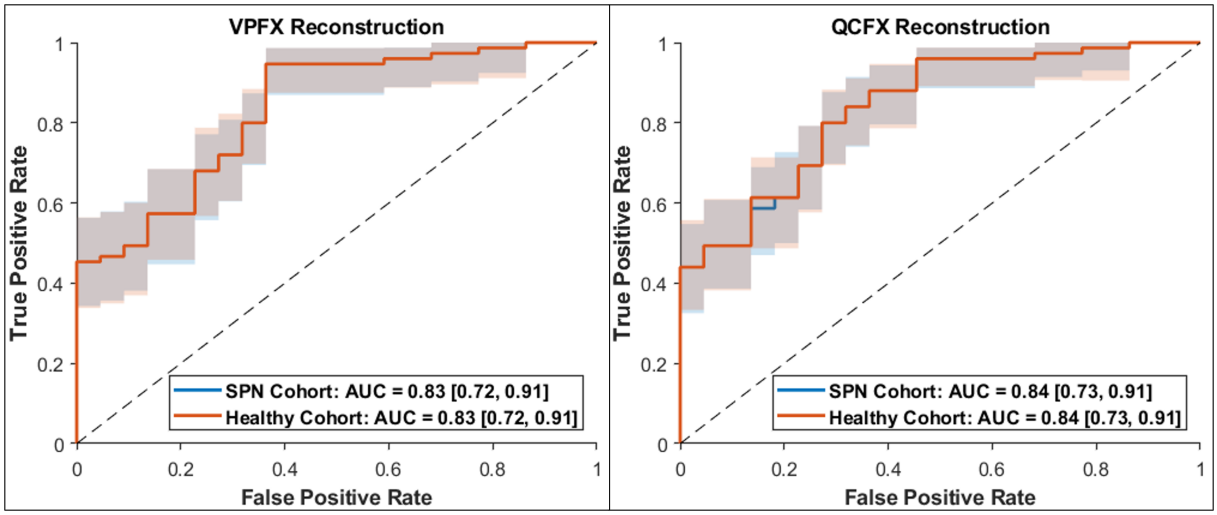

Fig. 4.11 ROC curves for the the diagnostic performance of Herder scores for predicting malignancy in the SPN and healthy cohorts using VPFX reconstruction (left) and QCFX reconstruction (right). Blue and orange shaded areas show 95 % CIs for the SPN and healthy ROC curves respectively.

Sensitivity (Se), specificity (Sp), positive predictive values (PPV) and negative predictive values (NPV) for both the SPN and healthy cohorts are presented in Table 4.10 for VPFX and Table 4.11 for QCFX. Results are shown for Herder score cut-offs of 10 %, 50 % and 70 %. No differences were observed between the diagnostic performance for the SPN and healthy patient cohorts with identical results across all Herder cut-offs for both VPFX and QCFX reconstructions.

Herder Cut-Off (%)		SPN Cohort		Healthy Cohort	
10	Se	0.99	[0.93, 1.00]	0.99	[0.93, 1.00]
	Sp	0.18	[0.05, 0.40]	0.18	[0.05, 0.40]
	PPV	0.80	[0.77, 0.83]	0.80	[0.77, 0.83]
	NPV	0.80	[0.32, 0.97]	0.80	[0.32, 0.97]
50	Se	0.89	[0.80, 0.95]	0.89	[0.80, 0.95]
	Sp	0.64	[0.41, 0.83]	0.64	[0.41, 0.83]
	PPV	0.89	[0.83, 0.94]	0.89	[0.83, 0.94]
	NPV	0.64	[0.46, 0.78]	0.64	[0.46, 0.78]
70	Se	0.76	[0.65, 0.85]	0.76	[0.65, 0.85]
	Sp	0.68	[0.45, 0.86]	0.68	[0.45, 0.86]
	PPV	0.89	[0.81, 0.94]	0.89	[0.81, 0.94]
	NPV	0.45	[0.34, 0.58]	0.45	[0.45, 0.58]

Table 4.10 Sensitivity (Se), specificity (Sp), positive predictive values (PPV) and negative predictive values (NPV) derived for Herder cut-offs of 10, 50 and 70 % for both cohorts when using the ^{18}F-FDG uptake measured on the VPFX reconstruction. Values in brackets show the 95 % CI.

Herder Cut-Off (%)		SPN Cohort		Healthy Cohort	
10	Se	0.99	[0.93, 1.00]	0.99	[0.93, 1.00]
	Sp	0.18	[0.05, 0.40]	0.18	[0.05, 0.40]
	PPV	0.80	[0.77, 0.83]	0.80	[0.77, 0.83]
	NPV	0.80	[0.32, 0.97]	0.80	[0.32, 0.97]
50	Se	0.96	[0.89, 0.99]	0.96	[0.89, 0.99]
	Sp	0.55	[0.32, 0.76]	0.55	[0.32, 0.76]
	PPV	0.88	[0.82, 0.92]	0.88	[0.82, 0.92]
	NPV	0.80	[0.55, 0.93]	0.80	[0.55, 0.93]
70	Se	0.83	[0.72, 0.90]	0.83	[0.72, 0.90]
	Sp	0.68	[0.45, 0.86]	0.68	[0.45, 0.86]
	PPV	0.90	[0.83, 0.94]	0.90	[0.83, 0.94]
	NPV	0.54	[0.39, 0.67]	0.54	[0.39, 0.67]

Table 4.11 Sensitivity (Se), specificity (Sp), positive predictive values (PPV) and negative predictive values (NPV) derived for Herder cut-offs of 10, 50 and 70 % for both cohorts when using the ^{18}F-FDG uptake measured on the QCFX reconstruction. Values in brackets show the 95 % CI.

4.4 Discussion

The work in this chapter used the insertion methodology described in Chapter 3 to generate images for a cohort of patients comprising of real clinical PET data with simulated lesions. Characteristics of the lesions were assigned based on a real population of patients with a known solitary pulmonary nodule (SPN). Validation of the cohort containing the simulated nodules was performed by comparison of quantitative measures of ^{18}F-FDG uptake and diagnostic accuracy for the assessment of malignancy risk with the cohort with real SPNs. Results showed no statistical difference in the SUV_{max}, Herder scores or diagnostic accuracy for the two cohorts. These results show the lesion insertion technique described in this thesis can produce images of a virtual patient population with results similar to those seen in a real population. Agreement in

the classification of ^{18}F-FDG uptake for the two cohorts was comparable to or better than the inter-observer agreement reported in the original study by Murphy *et al* [165].

There are a number of physical, technical and biological factors that can affect quantification of ^{18}F-FDG uptake [176]. By selecting a cohort of patients scanned on the same scanners in the same centre as the real SPN patients to use for lesion insertion, the intention was to ensure many of these factors, such as patient preparation, scanner calibration and scanning technique, were consistent. However, due to a change in the local clinical protocol from 90 to 60 minutes uptake, more patients were scanned at 60 minutes in the cohort used for lesion insertion in this thesis (71 %) compared to the cohort of patients with real SPNs (11 %). The mean uptake time for the healthy cohort was therefore shorter than that for the SPN cohort by 16.0 %. Chin *et al* compared SUV measurements for healthy tissues in patients scanned either at 1 hour or 3 hours post injection [177]. The authors found no significant difference in SUV for liver or lung tissues, but there was a moderate decrease in SUV for the aortic blood pool at 3 hours which was significant. Therefore, a large difference in uptake times between the healthy patients used for the lesion insertion in this thesis and the SPN cohort, could have resulted in a systematic bias in uptake measurements for the MBP which would in turn influence the classification of ^{18}F-FDG for patients where the SPN uptake was close to the MBP uptake. Comparison of the reference tissues for the healthy and SPN cohorts used in this thesis showed no significant difference in SUVs measured for the liver, lung or MBP indicating that the difference in uptake times was not large enough to measure a significant change in MBP uptake.

In the original study by Murphy *et al* [165], measurements of SUV_{max} for analysis of the SPN cohort were made on the clinical workstations using Hermes Hybrid Viewer (Hermes Medical Solutions) whereas the measurements for this thesis were made using MIM Encore. Therefore, to ensure SUV results agreed between the two software, quantitative assessments were performed in both software using a NEMA phantom acquired on the $D710_{real}$. All SUV measures were identical to 1 decimal place allowing results from the different software to be used interchangeably.

In this thesis, the ^{18}F-FDG uptake in the SPNs was classified using the four-point scale described by Murphy *et al* with classification confirmed using the SUV measured in the reference tissues and nodules [165]. Use of SUV measurements in combination with an ordinal scale can reduce variability between readers and avoid misinterpretation due to the influence of background activity on perceived uptake in the lesions [178]. The ordinal scale recommended by the BTS guidelines [7] relies on qualitative assessment to classify nodules using the four-point scale and this has become routine in clinical practice. In the study by Ordidge *et al*, retrospective evaluation of the ^{18}F-FDG uptake in SPNs for 100 patients was performed using the four-point BTS scale [179]. The authors found excellent interobserver agreement for three experienced reviewers with intraclass correlation coefficient (ICC)=0.97, 95 % CI [0.96, 0.98]. In a another study by Fatania *et al*, seven clinicians with varying levels of experience from four different imaging centres retrospectively scored the ^{18}F-FDG uptake on PET-CT in 50 patients with SPNs [180]. The authors found good interobserver reliability between all seven reviewers with ICC=0.78, 95 % CI [0.67, 0.85] when using the four-point BTS scale.

There are several techniques for generating synthetic lesions described in the literature, mostly geared towards making visually realistic lesions in CT or x-ray applications and for the assessment of segmentation algorithms in PET applications. These techniques involve the use of segmented lesions [181, 182] or mathematical models [183]. Methods involving segmentation of real lesions require experts to outline the lesions on real clinical images to create a library of lesions to sample from, which is time consuming. This also requires access to large datasets of images with lesions for each tumour type of interest. The segmented lesions will be dependent on the characteristics of the original imaging system on which they were acquired, in particular the finite voxel size will limit the resolution of the synthetic lesions which may result in unrealistic stepped edges. For the purposes of this project, a mathematical model was sought to allow automatic generation of the synthetic lesions avoiding the need for segmentation. Rashidnasab *et al* investigated two mathematical models for generating 3D masses for insertion into 2D mammograms which were evaluated with observer studies [184]. Of these two, the random walk (RW) technique was very similar to the tumour growth algorithm described in this thesis

except the only parameters that could be modified for the RW method were the size and density of the lesion, meaning the lesions were all very spherical in shape. The second method used diffusion limited aggregation (DLA) which was based on a fractal growth model used originally for modelling crystal growth. The observer study showed the RW method produced masses of real appearance 68 % of the time and the DLA technique 84 % of the time, compared to 94 % for real lesions. Although both these methods could produce lesions with realistic visual appearance, unlike the 3D tumour model employed for producing synthetic lesions in this thesis, neither was based on a model incorporating biological processes. It should be noted, the tumour growth model used in this thesis did not account for existing structures within the lung, however it is too complex to model an entire tumour and its neighbouring environment [185].

With advances in computational power and machine learning methods, generative adversarial networks (GANs) provide an alternative method for generating images of realistic lesions. There are two parts to a GAN, a generator and a discriminator, both of which are neural networks. The generative model is trained using real data, such as CT images of pulmonary lesions, and generates new examples. These examples are fed into a discriminative model which then classifies the data as real or fake. The result of the classification by the discriminator is then fed back into the generator model using backpropagation to further train the generative model. As each subsequent image is created, they become more realistic and eventually the generator model is able to create images that the discriminator is unable to distinguish from the real data. Several groups have already explored the use of GANs for creating synthetic lesions including breast lesions from mammography images, [186], focal liver lesions from CT images, [187] and lung lesions mimicking ground glass nodules from CT images [188]. Results from observer studies in each case have demonstrated the lesions are difficult to distinguish from the real lesions. It should be noted however, that in all cases the authors have used 2D images for the input and output. In order to extend the GANs for use with 3D data, as required for the work in this thesis, a significant increase in the amount of training data and computational resources would be required.

The synthetic lesions were resampled to 1 mm^3 voxels for generating the voxelised phantoms for input into the Monte Carlo model. This could lose some of the finer detail from the synthetic lesions, however this is much smaller than the PET scanner resolution of 6 mm to 7 mm and is smaller than the theoretically achievable spatial resolution of 2 mm to 2 mm so should not affect the results in this work.

In the Bland-Altman plots an upwards trend in the differences between the SUV_{max} in the simulated and real nodules was observed as the SUV in the real nodule increased and also as the nodule diameter increased. This is likely due to the fact that the true uptake in the real nodules was unknown and were estimated using a partial volume correction (PVC) using RCs based on phantom data. To derive the PVC, manual measurements of the real SPN made by the clinician on CT were used which, particularly for smaller lesions, would also be subject to the partial volume effect. Measurements were to the nearest mm and the phantom used to derive the correction uses unrealistic spherical objects. The real SPNs would also be subject to patient motion which would further degrade the spatial resolution in the PET images resulting in underestimation of the SUV [189]. Despite this, results showed good agreement between the SUV_{max} measurements for the simulated lesions and the real nodules from which the characteristics were derived and there was no significant difference in the AUC derived from the two patient cohorts.

The study by Murphy *et al* [165] was made possible by the fact that raw PET data were archived and could be retrieved for retrospective reconstruction using the new Q.Clear algorithm which was included as part of a software upgrade to the existing PET scanner. In most PET sites however, raw PET data is not stored beyond a couple of weeks so this type of study would not be possible. For a single patient the reconstructed PET images are around 40 MB compared to 500 MB for the raw PET data making long term storage of raw data expensive. Additionally, it would be difficult to predict which conditions will be of interest so all data would need to be stored and carefully curated to make it useful. Use of the insertion technique described in this thesis would allow the storage of a much smaller selection of raw PET data with typical uptake for ^{18}F-FDG-PET and this same data could be used for investigation of various tumour types by insertion of lesions with different characteristics at different anatomical locations.

Further work in this thesis will use the validated methodology for generating virtual patient populations and extend it to investigate the effect of new technologies on quantitative measures in the context of clinical trials.

4.5 Conclusions

PET images of a virtual population of patients with simulated lesions have been generated and validated against a real cohort of patients scans with known SPNs making the novel method described in this thesis suitable as a surrogate for real populations of patients for virtual imaging trials.

Chapter 5

Phantom and Clinical Assessment of the Impact of New PET Technologies on Image Quality and Diagnostic Performance for Determining Risk of Malignancy in Solitary Pulmonary Nodules

5.1 Introduction

The previous chapters describe the methodology for simulation of solitary pulmonary nodules (SPNs) and their insertion into real PET data in projection space to generate images for use in virtual imaging trials (VITs). Validation comprised of generating PET images of a population of healthy patients with simulated lesions, using the methodology described in this thesis, and comparing quantitative measures of ^{18}F-FDG uptake and derivation of malignancy risk from the relative PET uptake to those measured for real SPNs in a real population of patients. Results showed the classification of FDG uptake for the real and simulated lesions was comparable

to the inter-observer agreement reported for dual readers of the real SPNs and there was no significant difference in the diagnostic performance for classification of nodules as malignant or benign using the Herder risk score. The aim of this chapter was to investigate the impact of new technologies, specifically time-of-flight (TOF), point spread function (PSF) modelling and the use of a Bayesian penalised likelihood (BPL) reconstruction, on technical image-based metrics of image quality and on diagnostic accuracy for task-based measures.

There are a number of factors that affect quantification in PET [190] which can impact on clinical diagnosis or response assessment [191]. Variation due to patient related factors can be minimised through use of matching patient preparation and acquisition protocols. Technical factors leading to the partial volume effect however are dependent on scanner geometry, the reconstruction algorithm used and the implementation of corrections within the system model. Often these will vary between manufacturers and between scanner models making it very difficult to be prescriptive in how PET scans should be acquired and reconstructed. Additionally, PET-CT scanner hardware and software have advanced significantly over the last decade with the introduction of fully 3D iterative reconstruction, time-of-flight (TOF), point spread function (PSF) modelling, continuous motion and most recently solid state photon detectors. This has resulted in a range of different technologies in clinical PET centres across the UK.

There are guidelines for ^{18}F-FDG-PET-CT for tumour imaging which give general recommendations for patient preparation, image acquisition and reconstruction to improve consistency in reporting and use of semi-quantitative measures across platforms and institutions [37]. These guidelines recommend the use of physical phantoms for harmonisation of PET image quality with the aim of allowing comparison of PET images and quantitative measures across different scanners and sites. In the UK and Europe, matching of recovery coefficients determined using the NEMA image quality phantom is the accepted method for inter-comparison between systems [37, 192]. Use of this technique has been successfully applied for multicentre lymphoma trials in the UK since 2003 [98] and is currently used by the UK PET Core Lab (http://www.ncri-pet.org.uk/) and EARL (https://earl.eanm.org/) for PET-CT scanner accreditation. Despite these recommendations to use harmonised parameters, sites still opt to use locally derived parameters which are

often subjective and based on the preference of the radiologist(s). Whilst this provides images that are visually acceptable, this makes comparison of quantitative values difficult.

Advancements in PET technology have resulted in improved sensitivity, timing resolution and spatial resolution which have been shown to result in overall CNR gain. Lower noise and good signal recovery are generally associated with optimal image quality, and potentially improved diagnostic accuracy for lesion detection. However, since the ground truth is unknown for real lesions it is difficult to assess the diagnostic accuracy in tasks that involve characterisation of lesions based on uptake or size. The aim of this chapter therefore was to investigate the impact of new technologies, in particular the use of TOF, PSF and Bayesian penalised likelihood (BPL) reconstruction algorithm, on the image quality and quantitative measurements of ^{18}F activity concentration for spheres of different sizes in the NEMA image quality phantom. Further, PET images consisting of real PET scans containing simulated lesions were generated using the techniques described in the previous chapters to investigate the influence of these new technologies on ^{18}F-FDG uptake measurements and the subsequent impact on diagnostic performance for assessment of malignancy risk in patients with SPNs.

5.2 Methods

In the first part of this work, the impact of new technologies on image quality and quantitative measurements of ^{18}F-FDG uptake were investigated using the NEMA image quality phantom. Real scans of the physical phantom with simulated spheres inserted using the validated technique in Chapter 3 were reconstructed first using OS-EM with and without incorporation of TOF information and PSF modelling, then using BPL with TOF and PSF. Images were analysed using the technical image-based metrics described in Chapter 2, Section 2.2.1. To further investigate the impact of these new technologies on diagnostic performance, PET images of a patient population with SPNs were generated using the insertion technique described in Chapters 3 and 4. These datasets were also reconstructed using OS-EM with and without incorporation of TOF information and PSF modelling and with the BPL reconstruction algorithm. The diagnostic

performance in determining malignancy risk was assessed for each of the reconstructions using receiver operating characteristic (ROC) curve analysis.

5.2.1 Assessment of the Impact of New Technologies on Image Quality using the NEMA Phantom

5.2.1.1 Generation of NEMA Phantom Images with Simulated Spheres

The background compartment of the physical NEMA image quality phantom (Data Spectrum Corporation) was filled with an ^{18}F solution, and a 5 min acquisition performed on the GE Discovery 710 PET-CT scanner (General Electric Medical Systems, GEMS, Milwaukee, WI) at the King's College London & Guy's and St Thomas' PET Centre. Four sets of wall-less spheres of water with diameters of 10, 13, 17, 22, 28 and 37 mm were simulated with different activities of ^{18}F using the GATE toolkit (version 8.2). The spheres were then inserted into the raw PET data from the real acquisition of the phantom background using the technique described in Chapter 3, Section 3.2.2. The activity concentration in the background compartment was 7.8 kBq ml^{-1} at the start of the acquisition and the activity concentrations for the inserted spheres ranged from 33 to 62 kBq ml^{-1} resulting in four phantom datasets with sphere-to-background ratios (SBRs) of 4.2, 4.9, 6.5 and 7.9 to 1.

The PET data were reconstructed using 3D OS-EM (VPHD), 3D OS-EM with TOF (VPFX), 3D OS-EM with TOF and PSF (VPFXS) and BPL with TOF and PSF (QCFX) using the GE proprietary PET Toolbox (Duetto, version 2.07). The parameters for each reconstruction are shown in Table 5.1. These reconstruction parameters were chosen as they cover a range of settings currently used in clinical practice by UK sites (data from UK PET Core Lab http://www.ncri-pet.org.uk/). Recovery coefficients (RCs) for both the VPHD and VPFX reconstructions fulfil the requirements for accreditation by the UK PET Core Lab and the VPFX reconstruction is used clinically at the King's College London & Guy's and St Thomas' PET Centre.

	Reconstruction Name			
Parameter	**VPHD**	**VPFX**	**VPFXS**	**QCFX**
Algorithm Used	OS-EM	OS-EM	OS-EM	BPL
Matrix Size	192 × 192	256 × 256	256 × 256	256 × 256
Pixel Size (mm)	3.65 × 3.65	2.73 × 2.73	2.73 × 2.73	2.73 × 2.73
Slice Thickness (mm)	3.27	3.27	3.27	3.27
Iterations	2	2	2	N/A
Subsets	24	24	16	N/A
Gaussian Post Filter (mm)	6.4	6.4	4.0	N/A
z-axis Filter (mm)	4	4	4	N/A
Strength (β) of the Penalising Factor	N/A	N/A	N/A	400
Time-of-Flight (TOF) Included	No	Yes	Yes	Yes
Point Spread Function (PSF) Modelling Included	No	No	Yes	Yes

Table 5.1 Reconstruction algorithms and parameter settings used in the assessment. OS-EM = ordered subsets-expectation maximisation, BPL = Bayesian penalised likelihood. Parameters that are not applicable for the particular reconstruction are denoted as N/A.

5.2.1.2 Assessment of Noise in the Phantom Background

The noise in the uniform background of the phantom was characterised using the background variability (BV), as defined in the NEMA standard [40], and the image roughness (IR) [76]. As discussed previously in Section 2.2, the BV measures the region-to-region variation, and the IR measures the voxel-to-voxel variation within a single image.

Twelve circular ROIs of 37 mm diameter were placed on an axial slice centred on the spheres and copied to axial slices at +19.6, +9.8, -9.8 and -19.6 mm to the central slice. Smaller ROIs with diameters matching the diameters of the spheres (10, 13, 17, 22 and 28 mm) were then drawn concentric to the 37 mm ROIs to give a total of 60 ROIs for each diameter. Figure 5.1 shows the placement of the 37 mm diameter ROIs on the axial slice centred on the spheres. The

ROIs were copied to each reconstruction for each sphere-to-background ratio (SBR) and the mean and standard deviation (SD) measured for all individual ROIs.

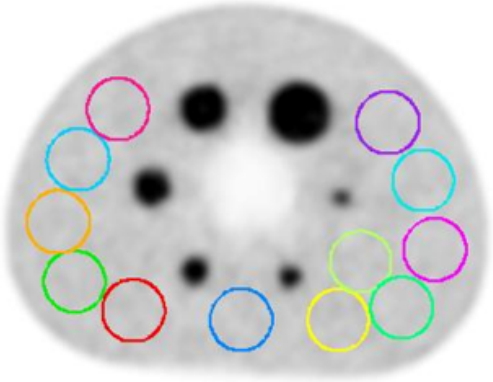

Fig. 5.1 Axial slice of the NEMA phantom centred on the spheres showing the placement of the 37 mm ROIs used to determine the image noise.

The *BV* was calculated for each ROI diameter (j) and each SBR as:

$$BV_j\,(\%) = \frac{SD_j}{C_{B,j}} \times 100$$

where $C_{B,j}$ is the mean activity concentration measured for all 60 ROIs of diameter j and SD_j is the standard deviation in the activity concentration measurements for the 60 ROIs of diameter j calculated as:

$$SD_j = \sqrt{\frac{\sum_{k=1}^{K}\left(C_{B,j,k} - C_{B,j}\right)^2}{K-1}}$$

$C_{B,j,k}$ is the mean activity concentration for each individual ROI (k) of diameter j and K is the total number of ROIs.

The *IR* was calculated as the CoV for each individual ROI (k) averaged over all 60 ROIs:

$$IR_j\,(\%) = \frac{\sum_{k=1}^{K}\frac{SD_{j,k}}{C_{B,j,k}} \times 100}{K}$$

5.2.1.3 Assessment of the Signal Recovery in the Spheres

The signal recovery for the spheres was measured using recovery coefficients (RCs). The activity concentration for the spheres was measured for each reconstruction using both the maximum voxel value within the spheres (C_{max}) and the mean for a VOI defined using a threshold of 50 % adapted for the local background (C_{A50}). The RC_{max} and RC_{A50} were then calculated for the different sphere sizes as:

$$RC = \frac{C_{measured}}{C_{true}}$$

where $C_{measured}$ is the activity concentration measured in the reconstructed PET images (either C_{max} or C_{A50}), and C_{true} the activity concentration set in the simulations. The bias in the RCs was calculated as:

$$Bias\ (\%) = \left(\frac{C_{measured}}{C_{true}} - 1\right) \times 100$$

and the variance was the standard deviation in the measurements for the different SBRs.

The signal-to-noise ratio (*SNR*) and contrast-to-noise ratio (*CNR*) were also calculated for each sphere size and reconstruction as:

$$SNR = \frac{C_{max}}{SD_j}$$

$$CNR = \frac{C_{max} - C_{B,j}}{SD_j}$$

5.2.2 Assessment of the Impact of New Technologies on Diagnostic Performance

5.2.2.1 Generation of Patient PET Images with Simulated Lesions

In addition to the 97 patients collected for the healthy cohort in Chapter 4, a further 103 anonymised raw PET data and reconstructed clinical ^{18}F-FDG-PET-CT scans were collated using the methodology described in Section 4.2 (IRAS project ID: 251611). The PET-CT scans were acquired on one of two GE Discovery 710 PET systems at the King's College London &

Guy's and St Thomas' PET Centre. Selection of the PET datasets was blinded to all clinical information. After initial inspection of the PET images, 6 patient datasets were excluded due to significant motion in the thorax or extravasation, resulting in a total of 194 datasets suitable for lesion insertion. All patient preparation followed the routine clinical protocol of fasting for 6 hours prior to injection and the blood glucose was checked to ensure blood glucose was < 11 mmol/L. The mean ± SD injected activity for all patients was 322.2 ± 21.4 MBq. 50 patients were scanned using an uptake time of 90 minutes, (86.6 ± 7.8 min), and 144 patients using 60 minutes uptake time, (62.9 ± 3.6 min) with a mean uptake time for all patients of (69.0 ± 11.6 min). The acquisition and reconstructions protocols were identical for the two PET scanners and both were accredited by the UK PET Core Lab which requires ongoing annual quality assurance to ensure consistent performance.

The PET and CT used for lesion insertion were acquired from mid-thigh or toes to base-of-brain or vertex under shallow inspiration. The CT used for attenuation correction and localisation was acquired first using 140 kV p, variable mA s (15 – 100 mA, noise index 40), 0.5 s rotation time and 1.375 pitch. CT images were reconstructed with 2.5 mm slice thickness using the standard kernel and ASiR blending of 40 %. The PET was acquired using 3 min per bed position and 11-slice (23 %) overlap. The time per bed position was increased to 4 min per bed for patients over 100 kg.

After insertion of the simulated nodules, the PET data were reconstructed using 3D OS-EM (VPHD), 3D OS-EM with TOF (VPFX), 3D OS-EM with TOF and PSF (VPFXS) and BPL with TOF and PSF (QCFX). The parameters for each reconstruction were identical to those used for the phantom assessment as shown in Table 5.1.

5.2.2.2 Clinical and Diagnostic Characteristics

Of the 194 patient datasets, 150 (77.3 %) were randomly designated as malignant and 44 (22.7 %) benign. A summary of the assigned clinical and radiological characteristics is given in Table 5.2.

Clinical or Radiological Characteristic	Benign (n=44)	Malignant (n=150)	All Patients (N=194)
Age (years)			
Mean ± SD:	70.9 ± 9.3	68.9 ± 10.0	69.3 ± 9.8
Range:	53-85	32-87	32-87
Cigarettes (%)	72.7	86.7	83.5
Cancer (%)	22.7	29.3	27.8
Diameter (mm)			
Mean ± SD:	11.4 ± 5.2	17.8 ± 5.5	16.4 ± 6.1
Median:	10.0	18.0	15.0
Range:	5-25	6-30	5-30
Spiculation (%)	22.7	65.3	55.7
Upper (%)	54.5	52.0	52.6

Table 5.2 Summary of assigned radiological characteristics for the simulated lesions and the designated clinical characteristics for the patients used for insertion.

The clinical characteristics (age, smoking status, and previous cancer) were assigned according to the proportions reported for patients with real malignant and benign SPNs in the study by Murphy *et al* [165]. The size, location, and presence of spiculation for the simulated lesions were also assigned using the radiological characteristics measured for the nodules in the real population. For the nodule size, the distribution of sizes for the real benign and malignant SPNs were modelled using lognormal and normal distributions respectively. The BTS guidelines recommend a minimum diameter of 5 mm for monitoring SPNs as the risk for nodules < 5 mm is less than 1 % [7]. Therefore, the diameters for the simulated malignant and benign lesions were randomly generated from the truncated lognormal and normal distributions using lower and upper limits of 5 mm and 30 mm respectively. The nodules were then simulated using these lesion characteristics and inserted into the real PET data using the methodology described in Chapters 3 and 4.

5.2.2.3 Quantitative Evaluation of FDG Uptake in the Reference Tissues and SPNs

All analysis was performed using MIM Encore (MIM Software Inc, version 7.1.5). SUV_{max} and SUV_{mean} were measured for the liver and MBP as described in the simplified PERCIST guide [87]. The liver volume was defined using a 3 cm diameter spherical VOI placed in the right side of the liver. MBP uptake was measured using a 1 cm diameter by 2 cm length cylindrical VOI placed in the descending thoracic aorta. Additionally, the SUV_{max} and SUV_{mean} were measured for the lung using a 3 cm diameter spherical VOI placed in the right lung avoiding the bronchials. The standard clinical OS-EM reconstruction with TOF (VPFX with 2 iterations, 24 subsets and 6.4 mm Gaussian filter) and the CT were used to guide placement of the reference regions which were then copied to the other reconstructions.

SUV_{max} and SUV_{mean} were measured for the simulated SPNs. For delineation of the lesions to derive SUV_{mean}, a commercially available gradient-based algorithm, PETedge+ (MIM Encore v7.1.5, MIM Software Inc, Cleveland, OH), was chosen. To define the SPN volume a large spherical VOI was manually placed over the nodule on the VPFX PET scan, the VOI was automatically copied to the other reconstructions and then resized using the PETedge+ algorithm. The resultant volumes were not manually adjusted.

The nodule ^{18}F-FDG uptake measured using both SUV_{max} and SUV_{mean} was classified for each reconstruction according to the ordinal scale shown in Table 5.3 using both SUV_{max} and SUV_{mean} for the reference tissues.

Score	**^{18}F-FDG Uptake**	**Definition in terms of Nodule standardised uptake value (SUV)**	**Weighting Factor**
1	Absent	= lung SUV	0
2	Faint	> lung SUV but $\leq$ MBP SUV	2.322
3	Moderate	> MBP SUV and $\leq 2 \times$ liver SUV	4.617
4	Intense	$> 2 \times$ liver SUV	4.771

Table 5.3 Ordinal scale for classification of ^{18}F-FDG uptake measured on PET. Nodule uptake was classified using both the SUV_{max} and SUV_{mean} for the lung, liver and mediastinal blood pool (MBP).

5.2.2.4 Derivation of Malignancy Risk

First the risk of malignancy (P_s) was calculated for the SPNs using the clinical prediction model by Swensen *et al* [171]:

$$P_s = \frac{1}{(1+e^{-x})}$$

where x was determined from the assigned clinical and radiographic factors as:

$$x = -6.8272 + 0.0391(age) + 0.7917(cigarettes) + 1.3388(cancer) + 0.1274(diameter) + 1.0407(spiculation) + 0.7838(upper)$$

The probably of malignancy (P_h) was then calculated using the Herder risk model [9]:

$$P_h = \frac{1}{(1+e^{-(-4.739+3.691(P_s)+y)})}$$

where P_s is the Swensen risk derived from the clinical and radiological characteristics and y is the weighting factor based on the degree of ^{18}F-FDG uptake using the ordinal scale as given in Table 5.3.

5.2.2.5 Statistical Analysis

Results are presented as mean ± standard deviation (SD) for continuous variables and frequencies and percentages (%) for ordinal data. Pairwise comparison of the SUV and Herder risk was performed for each of the reconstructions with a P value <.05 considered statistically significant. The diagnostic performance for each of the reconstructions was assessed using sensitivity (Se), specificity (Sp), positive predictive values (PPV) and negative predictive values (NPV) for pre-defined cut-off points. Results are reported using 95 % confidence intervals (CIs), with Clopper-Pearson exact intervals for sensitivity and specificity [174] and standard logit intervals for PPVs and NPVs [175]. The overall diagnostic performance for discriminating malignancy are given as area under the receiver operating characteristic (ROC) curve (AUC) with 95 % pointwise CIs computed using 1000 bootstrap samples. Comparison of AUCs was performed

using the non-parametric method for correlated ROC curves described by DeLong *et al* [173]. Optimal cut-off values were determined for ROC curves using Youden's J statistic [193]. All statistical analysis was performed in MATLAB version R2022b.

5.3 Results

5.3.1 Assessment of the Impact of New Technologies on Image Quality using the NEMA Phantom

5.3.1.1 Assessment of Noise in the Phantom Background

Figure 5.2 shows axial PET slices of the NEMA image quality phantom reconstructed using the different parameters for SBRs of 4.2:1 (top) and 7.9:1 (bottom). On visual assessment, the spheres were visible across all reconstructions and for all SBRs. The spheres appear sharper going from VPHD→VPFX→VPFXS→QCFX, and there was no perceptible difference in noise between the reconstructions.

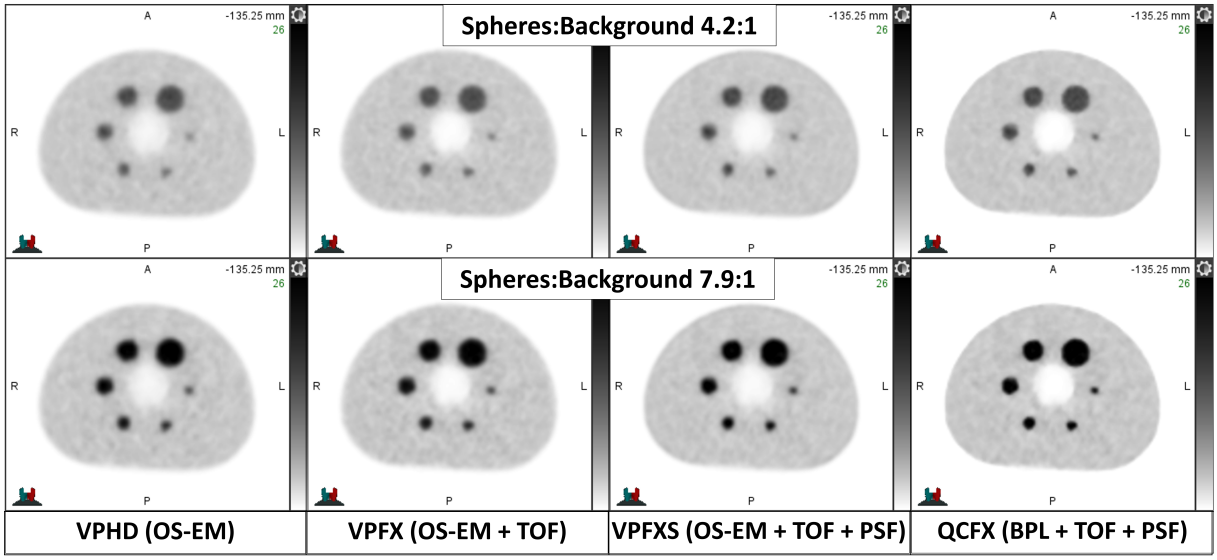

Fig. 5.2 Axial slices showing the different reconstructions of the NEMA image quality phantom. OS-EM = ordered subsets-expectation maximisation, BPL = Bayesian penalised likelihood, TOF = time-of-flight and PSF = point spread function. The sphere-to-background ratio (SBR) is 4.2:1 for the top row and 7.9:1 for the bottom row.

Table 5.4 shows the mean background variability (BV) and image roughness (IR) measured for the different reconstructions using the 37 mm ROIs. Results are quoted as the mean ± SD for the different SBRs.

Reconstruction	BV (%)	IR (%)
VPHD	2.98 ± 0.10	7.13 ± 0.44
VPFX	2.18 ± 0.13	7.61 ± 0.49
VPFXS	2.13 ± 0.10	6.65 ± 0.44
QCFX	1.96 ± 0.08	7.60 ± 0.55

Table 5.4 Background variability (BV) and image roughness (IR) measured using the 37 mm ROIs for the four different reconstructions. Values are the mean ± SD for the different sphere-to-background ratios (SBRs).

The results in Table 5.4 show a decrease in BV for the OS-EM reconstruction when incorporating TOF information (VPFX) and PSF modelling (VPFXS) and this was seen for all ROI sizes. Overall, the OS-EM reconstructions had higher BV than the BPL reconstruction (QCFX), which contains both TOF and PSF. These results indicate the accuracy of the corrections (attenuation and scatter) are improved with the addition of TOF and PSF within the reconstruction resulting in less region-to-region variation in the images. Conversely the VPHD reconstruction, which had the highest BV, had lower IR than both the VPFX and QCFX reconstructions. This is because the pixel size for the VPHD reconstruction was significantly larger (3.65 mm × 3.65 mm) than that used for the VPFX and QCFX reconstructions (2.73 mm × 2.73 mm) thus reducing the pixel-to-pixel noise. The VPFXS reconstruction had the lowest IR of all the reconstructions which is a consequence of increasing noise correlation for neighbouring pixels when using PSF modelling. A similar decrease in IR with the addition of PSF was seen in the phantom experiments by Tong *et al* [76]. This effect was not observed for the QCFX reconstruction due to the regularising prior which controls the image noise.

5.3.1.2 Assessment of the Signal Recovery in the Spheres

RC_{max} and RC_{A50} are plotted against the sphere diameters for each of the reconstructions and sphere-to-background ratios (SBRs) in Figure 5.3.

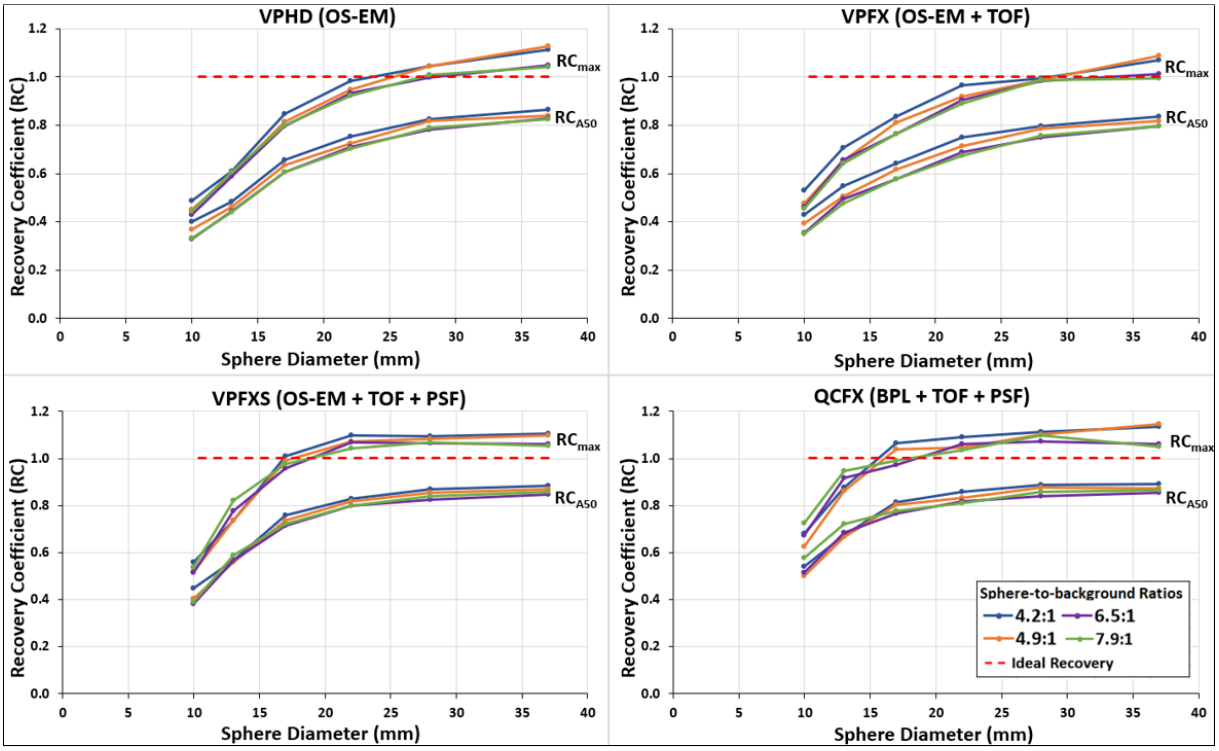

Fig. 5.3 Plots showing the recovery coefficients (RCs) for the maximum voxel value (RC_{max}) and mean (RC_{A50}) as a function of sphere diameter. Results are plotted for the four reconstructions using the different sphere-to-background ratios (SBRs). OS-EM = ordered subsets-expectation maximisation, BPL = Bayesian penalised likelihood, TOF = time-of-flight and PSF = point spread function.

All reconstructions showed decreasing signal recovery as sphere size decreased for both RC_{max} and RC_{A50}. The QCFX reconstruction demonstrated the greatest signal recovery across all sphere sizes followed by the VPFXS reconstruction as a result of the inclusion of PSF modelling. The biases in measured activity concentration for VPHD and VPFX were similar across all sphere sizes with maximum differences of 6.4 % in RC_{max} and 4.8 % in RC_{A50}. For the RC_{max} the biases in measured activity concentration were $\leq$ 8.3 % for spheres $\geq$ 22 mm using VPHD and $\leq$ 9.8 % for spheres $\geq$ 17 mm using VPFX. The biases in measured activity concentration were also similar for the VPFXS and QCFX reconstructions but only for spheres $\geq$ 17 mm with maximum differences of 3.3 % in RC_{max} and 5.6 % in RC_{A50}. For spheres $\leq$ 17 mm, the rate

of decrease in RC_{max} and RC_{A50} was greater for VPFXS than for QCFX. There were no trends observed in the variance in RC_{max} for the different SBRs when using any of the reconstructions reconstructions. For RC_{A50} the variance was similar for all sphere sizes, but increased for the 10 mm sphere using all reconstructions. The VPFXS had the lowest variance whilst VPFX had the highest.

Figure 5.4 shows the SNR and CNR for the 4.2:1 and 7.9:1 SBRs. Overall, the VPFXS and QCFX reconstructions both outperformed the VPHD and VPFX reconstructions in terms of SNR and CNR. Both the VPHD and VPFX reconstructions had similar signal recovery, whereas the VPHD had lower IR which resulted in higher SNR and CNR. It should be noted however, the VPHD reconstruction had greater BV than the VPFX reconstruction which is not accounted for in the SNR or CNR calculations. Again, due to the low IR and good signal recovery, the VPFXS reconstruction had the highest SNR and CNR for spheres ≥ 17 mm. The QCFX and VPFXS demonstrated similar SNR and CNR for the smaller spheres (10 mm and 13 mm) due to the greater decrease in recovery seen for VPFXS.

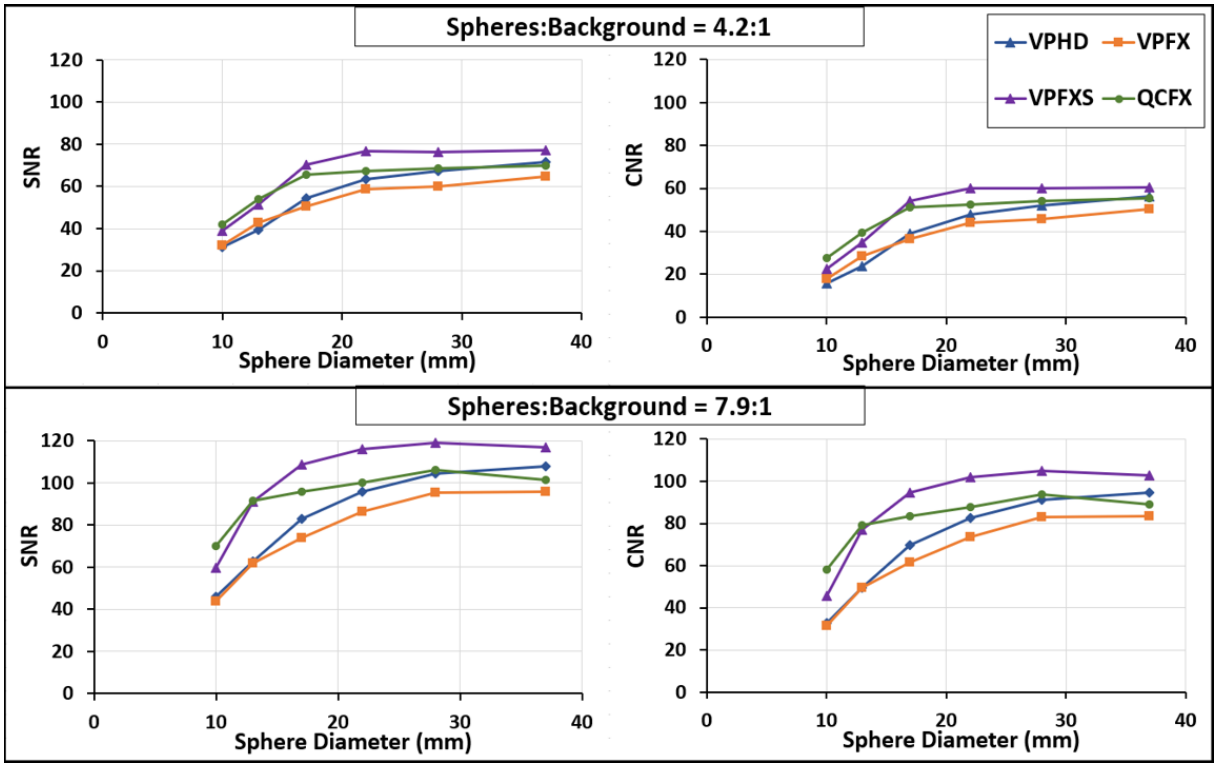

Fig. 5.4 Plots of the SNR and CNR measured using the maximum voxel value as a function of sphere diameter. Results are plotted for the four reconstructions with SBR of 4.2:1 for the top row and 7.9:1 for the bottom row.

5.3.2 Assessment of the Impact of New Technologies on Diagnostic Performance

5.3.2.1 Comparison of Diagnostic Performance for Derivation of Malignancy Risk using the Swensen Score

The Swensen risk scores for all patients were derived using the assigned clinical and diagnostic characteristics. Figure 5.5 shows the ROC curve for the diagnostic performance for the Swensen risk scores in predicting malignancy. The AUC using the Swensen score alone was 0.78, 95 % CI [0.70, 0.84].

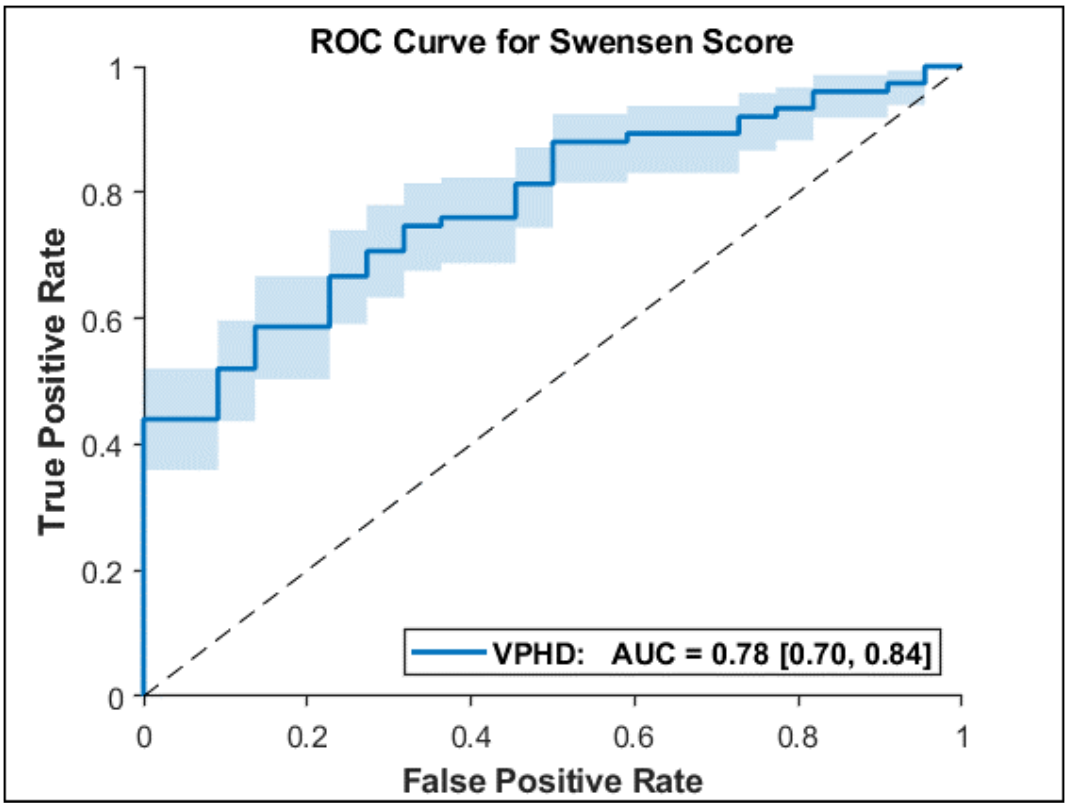

Fig. 5.5 ROC curve for the diagnostic performance of Swensen score in predicting malignancy. The AUC is shown with 95 % CIs.

5.3.2.2 Comparison of FDG Uptake in the Reference Tissues

The SUV_{max} and SUV_{mean} measured for the reference tissues in all patients using the four different reconstructions are presented as box plots in Figure 5.6.

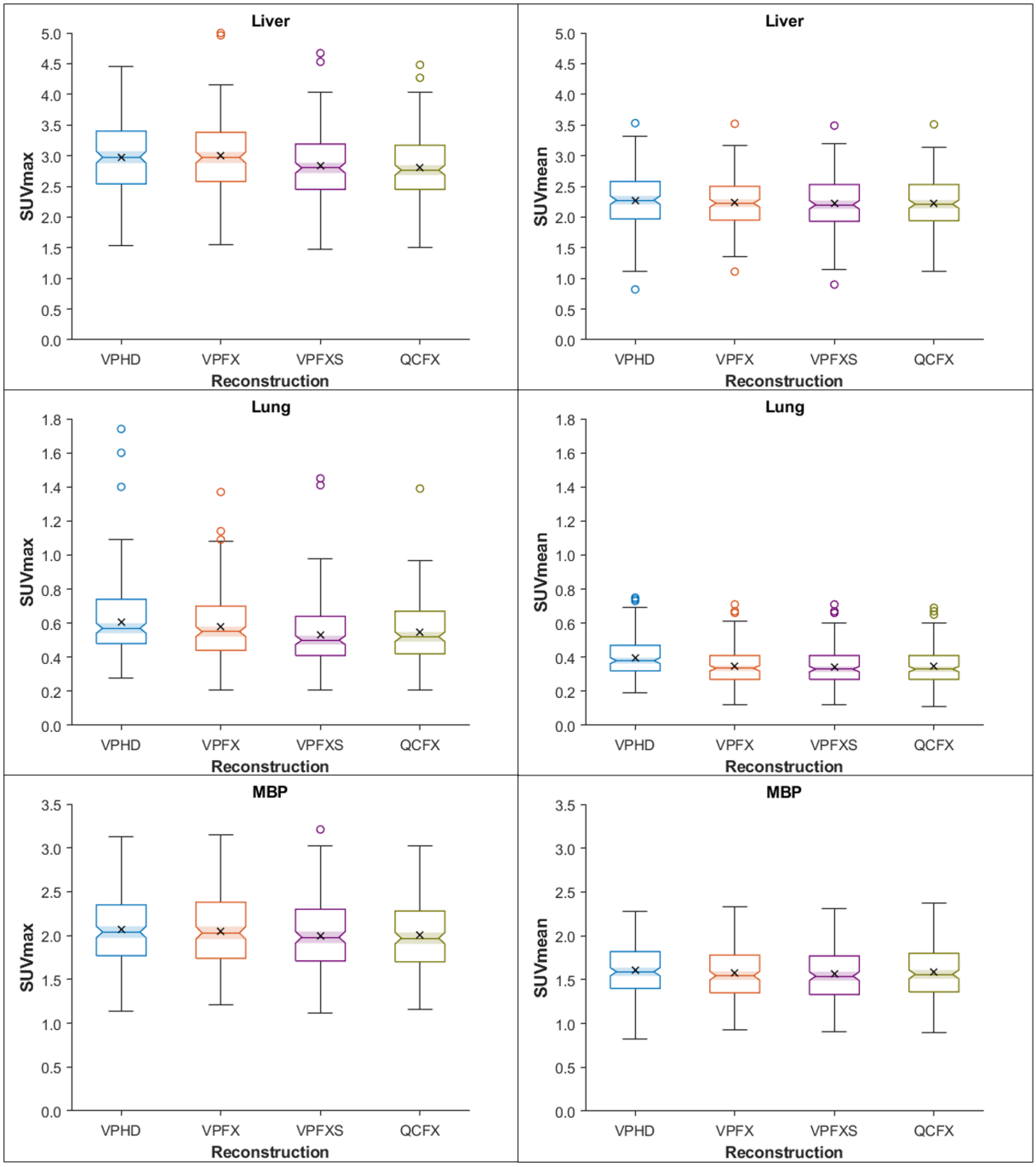

Fig. 5.6 Box plots showing the SUV_{max} and SUV_{mean} measured for the reference tissues using the four reconstructions. The $\times$ denotes the mean value for each reconstruction.

Table 5.5 shows the *P* values for the pairwise comparisons of the uptake measured for the reference tissues using the four different reconstructions. When using SUV_{max}, there was no significant difference in the MBP uptake measured for any of the reconstructions. There were also no significant differences in the SUV_{max} measured for the liver or lung when using the

VPHD and VPFX reconstructions or when using the VPFXS and QCFX reconstructions. There were however significant differences in the uptake for both liver and lung using SUV_{max} with the VPHD and QCFX reconstructions and the VPFX and VPFXS reconstructions. For VPHD and VPFXS reconstructions, there was no significant difference in the liver SUV_{max}, but there was for the lung and the opposite was seen for the VPFX and QCFX reconstructions.

	SUV_{max}			SUV_{mean}		
Reconstructions	**Liver**	**Lung**	**MBP**	**Liver**	**Lung**	**MBP**
VPHD vs. VPFX	.95	.55	.92	.88	<.001	.77
VPHD vs. VPFXS	.07	<.001	.21	.61	<.001	.63
VPHD vs. QCFX	.02	.01	.28	.67	<.001	.94
VPFX vs. VPFXS	.02	.03	.56	.97	.99	>.99
VPFX vs. QCFX	.004	.28	.65	.98	.99	.98
VPFXS vs. QCFX	.94	.78	>.99	>.99	>.99	.93

Table 5.5 *P* values for pairwise comparison of the uptake measured in the reference tissues when using the four different reconstructions. Results are shown for measurements of SUV_{max} and SUV_{mean}.

There were no significant differences between the SUV_{mean} measured in the liver or MBP for any of the reconstructions. There were also no significant differences between the SUV_{mean} in the lung for any of the reconstructions that included TOF (i.e. VPFX, VPFXS and QCFX), but there were significant differences in the SUV_{mean} measured in the lung for the VPHD reconstruction when compared to the other three reconstructions with all *P* values <.001.

5.3.2.3 Comparison of FDG Uptake in the Nodules

The nodule SUV_{max} and SUV_{mean} measured using the four reconstructions are shown in the box plots in Figure 5.7. Results are shown for all nodules (top row) and separated into malignant (middle row) and benign nodules (bottom row).

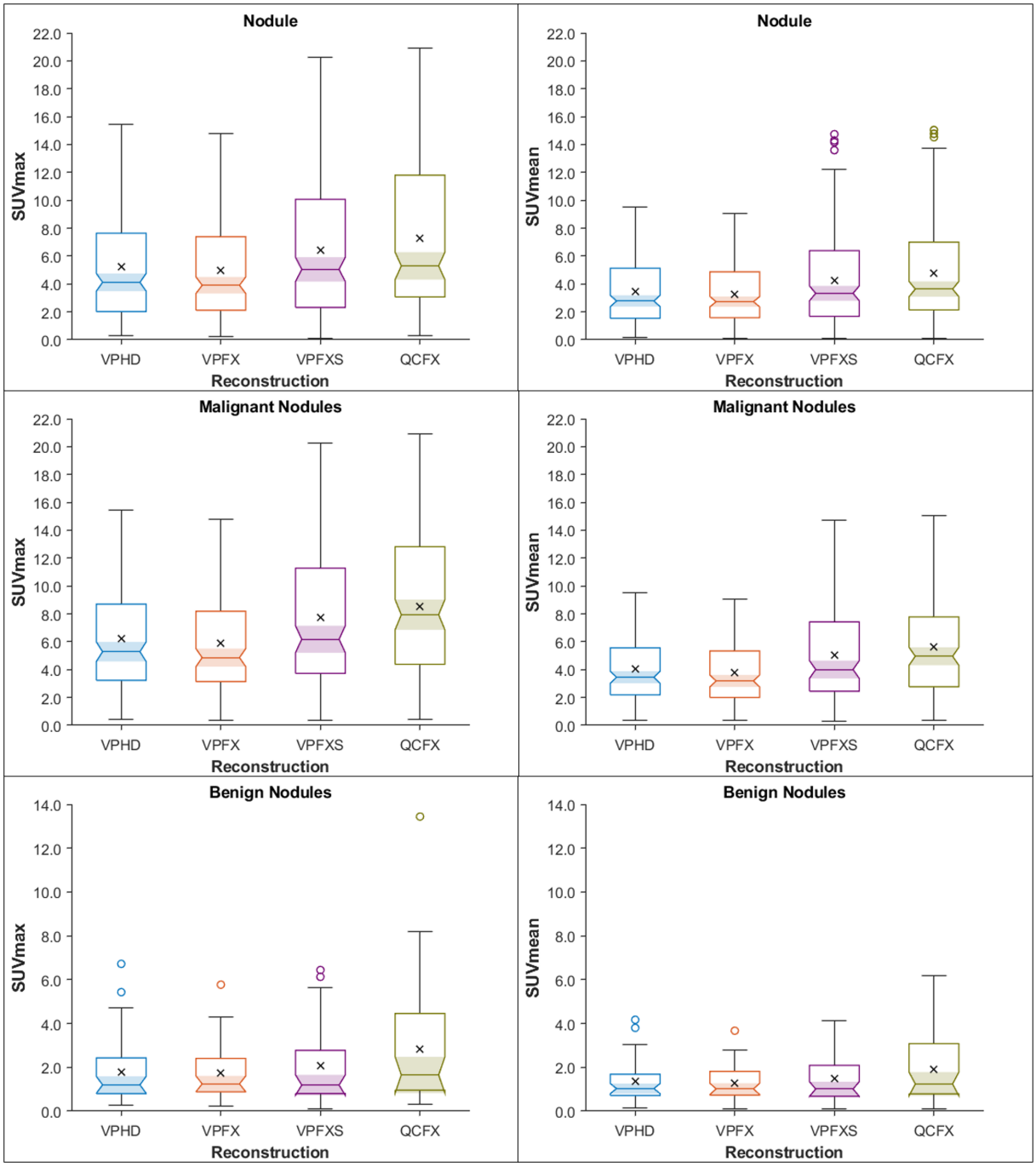

Fig. 5.7 Box plots showing the SUV_{max} and SUV_{mean} measured for all nodules (top row), malignant nodules (middle row) and benign nodules (bottom row) for the four reconstructions. The $\times$ denotes the mean value for each reconstruction.

As can be seen by these box plots, the measured nodule uptake was higher using the VPFXS and QCFX reconstructions for both SUV_{max} and SUV_{mean}. This result was more pronounced in the malignant lesions which tended to have higher uptake than the benign lesions. *P* values for

the pairwise comparisons of the uptake measured when using the different reconstructions are shown in Table 5.6. There were no significant differences between the SUV_{max} measured for all, malignant or benign nodules when using VPHD and VPFX or VPFXS and QCFX. There were significant differences between the SUV_{max} measured in the nodules for all other reconstructions, except in benign nodules for VPHD and VPFXS or VPFX and VPFXS.

	SUV_{max}			**SUV_{mean}**		
Reconstructions	**Benign**	**Malignant**	**All Nodules**	**Benign**	**Malignant**	**All Nodules**
VPHD vs. VPFX	>.99	.91	.94	.99	.88	.90
VPHD vs. VPFXS	.90	.01	.03	.95	.01	.03
VPHD vs. QCFX	.047	<.001	<.001	.09	<.001	<.001
VPFX vs. VPFXS	.87	.001	.006	.86	.001	.003
VPFX vs. QCFX	.04	<.001	<.001	.045	<.001	<.001
VPFXS vs. QCFX	.22	.36	.28	.27	.28	.23

Table 5.6 *P* values for pairwise comparison of the uptake measured in the nodules when using the four different reconstructions. Results are presented for SUV_{max} and SUV_{mean} across all nodules and separated into malignant and benign nodules.

Results showed there was no significant difference between the SUV_{mean} measured using VPHD and VPFX or the VPFXS and QCFX reconstructions either for all nodules nor when separated into malignant and benign nodules. For benign nodules there was only a significant difference observed in the SUV_{mean} measured using VPFX and QCFX. When considering all nodules together, there were significant differences in the SUV_{mean} measured using either the VPHD or VPFX reconstructions and both the VPFXS and QCFX reconstructions. This was driven by the difference in SUV_{mean} for the malignant nodules.

Figure 5.8 shows the ROC curves for the diagnostic performance of SUV_{max} (left) and SUV_{mean} (right) in predicting malignancy using the four different reconstructions. There was no significant difference between the AUCs derived using SUV_{max} and SUV_{mean} when using VPHD (*P*=.17), VPFX (*P*=.74), VPFXS (*P*=.94) or QCFX (*P*=.75) reconstructions.

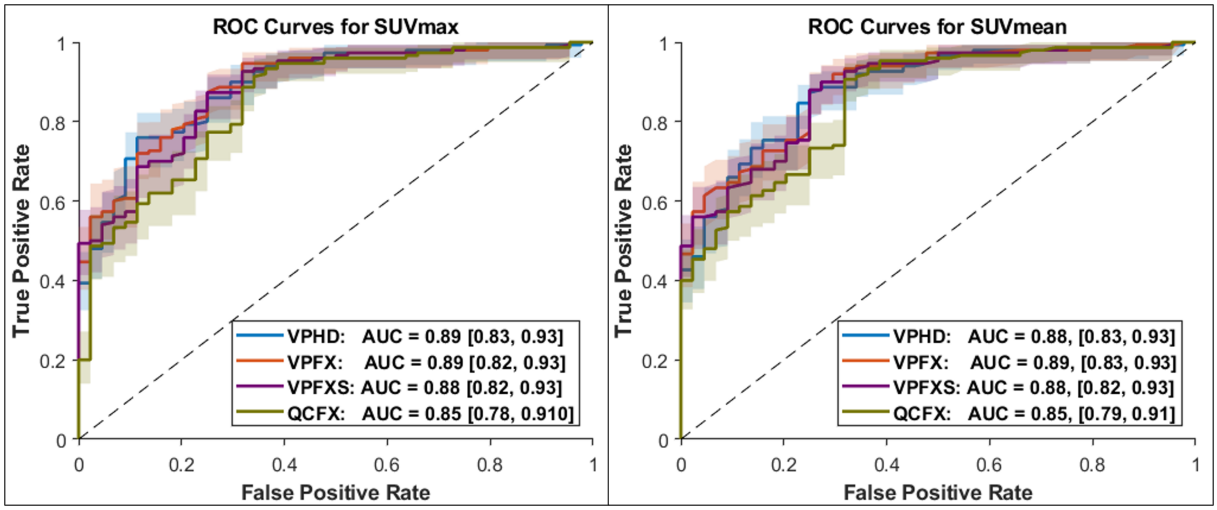

Fig. 5.8 ROC curves for the diagnostic performance of SUV_{max} (left) and SUV_{mean} (right) in predicting malignancy using the four different reconstructions. AUCs are shown with 95 % CIs.

Table 5.7 shows the *P* values for the pairwise comparisons of the diagnostic performance of both SUV_{max} and SUV_{mean} when using the four different reconstructions. As can be seen, there was no significant difference in the diagnostic performance of the VPHD and VPFX or VPHD and VPFXS reconstructions when using either SUV_{max} or SUV_{mean}. There were however significant differences in the diagnostic performance of both SUV_{max} and SUV_{mean} when using VPFX and QCFX and VPFXS and QCFX reconstructions. There were also significant differences in diagnostic performance between VPHD and QCFX and VPFX and VPFXS but only when using SUV_{max}.

Reconstructions	**SUV_{max}**	**SUV_{mean}**
VPHD vs. VPFX	.75	.32
VPHD vs. VPFXS	.36	.90
VPHD vs. QCFX	.03	.07
VPFX vs. VPFXS	.046	.13
VPFX vs. QCFX	.008	.006
VPFXS vs. QCFX	.01	.01

Table 5.7 *P* values for pairwise comparison of the diagnostic performance of both SUV_{max} and SUV_{mean} when using the four different reconstructions.

Optimal cut-offs for SUV_{max} and SUV_{mean} were derived using Youden's J statistic and are given in Table 5.8 for each reconstruction. Despite the higher sensitivity seen for QCFX, the Youden's index was slightly worse than the other reconstructions due to the lower specificity and this was seen for both SUV_{max} and SUV_{mean}. The optimal cut-offs and the sensitivity / specificity for the VPHD, VPFX and VPFXS were very similar when using SUV_{mean}, however this was not the case for SUV_{max} where there were noticeable differences in the sensitivity and specificity at the optimal cut-off points and the cut-offs for SUV_{max} ranged from 1.67 for VPFX to 3.18 for VPHD.

	Reconstruction	**SUV Cut-off**	**Specificity (Sp)**	**Sensitivity (Se)**	**Youden's Index**
SUV_{max}	VPHD	3.18	0.89	0.76	0.65
	VPFX	1.67	0.68	0.95	0.63
	VPFXS	2.59	0.75	0.87	0.62
	QCFX	2.52	0.66	0.91	0.57
SUV_{mean}	VPHD	1.62	0.75	0.87	0.62
	VPFX	1.66	0.75	0.88	0.63
	VPFXS	1.79	0.75	0.88	0.63
	QCFX	1.89	0.68	0.91	0.59

Table 5.8 Optimal cut-offs for SUV_{max} and SUV_{mean} using the four different reconstructions. Cut-offs were derived using Youden's J statistic assuming equal weights for false negative and false positive results.

5.3.2.4 Comparison of Diagnostic Performance for Derivation of Malignancy Risk using the Herder Score

Figure 5.9 shows the ROC curves for the diagnostic performance of Herder risk scores derived using a) SUV_{max} for both nodules and the reference tissues, b) nodule SUV_{max} and SUV_{mean} for the reference tissues, and c) SUV_{mean} for both nodules and the reference tissues. The AUCs for the different reconstructions are shown along with 95 % CIs.

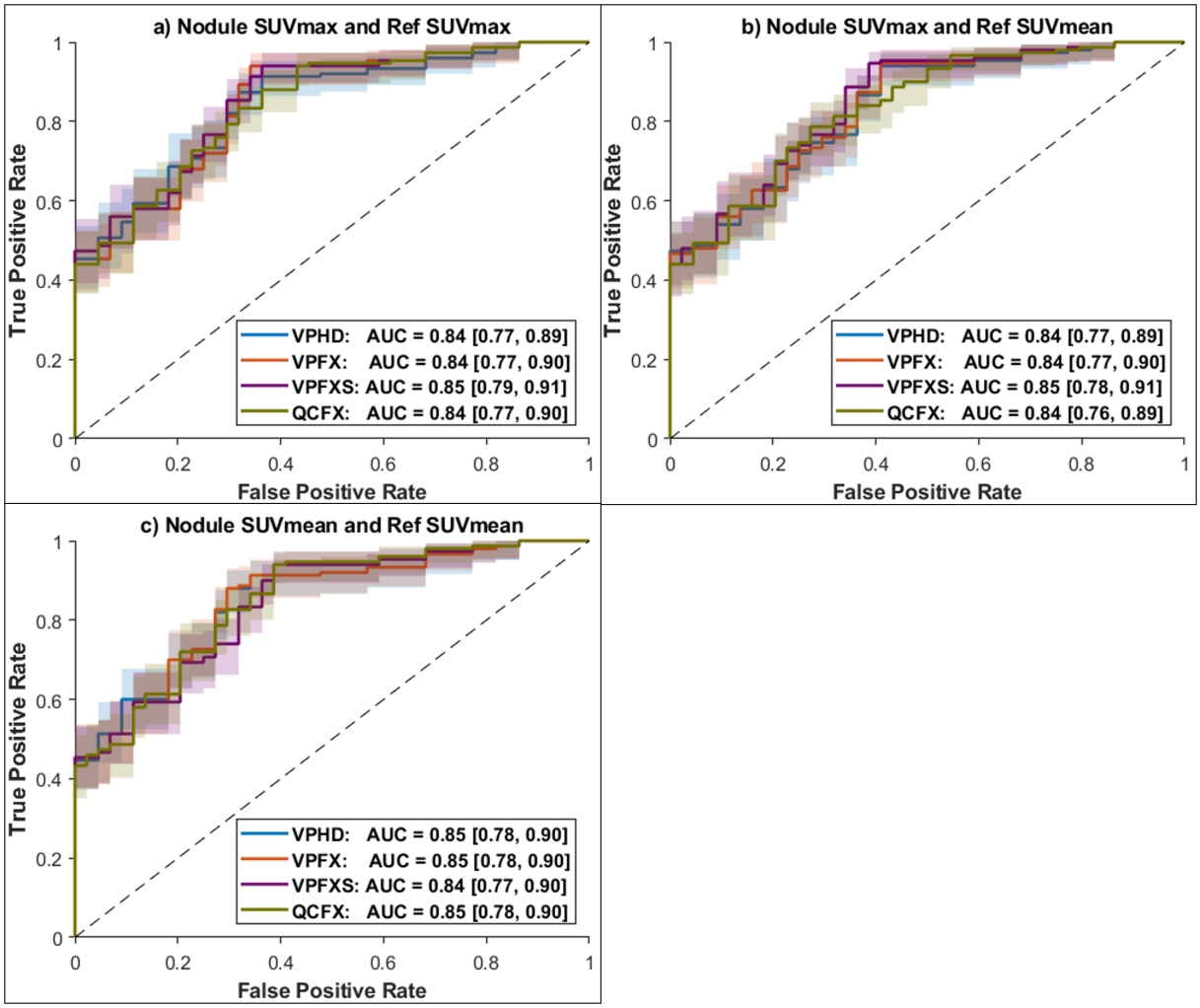

Fig. 5.9 ROC curves for the diagnostic performance of Herder score derived using a) nodule SUV_{max} and SUV_{max} for the reference tissues, b) nodule SUV_{max} and SUV_{mean} for the reference tissues, and c) nodule SUV_{mean} and SUV_{mean} for the reference tissues in predicting malignancy using the four different reconstructions. AUCs are shown with 95 % CIs.

Pairwise comparison showed no significant difference in AUCs for any of the reconstructions when using any of the three methods for classifying the ^{18}F-FDG uptake in the nodules as demonstrated by the *P* values in Tables 5.9 and 5.10.

Reconstructions	Nodule SUV_{max} + Ref SUV_{max}	Nodule SUV_{max} + Ref SUV_{mean}	Nodule SUV_{mean} + Ref SUV_{mean}
VPHD vs. VPFX	.83	.19	.81
VPHD vs. VPFXS	.40	.18	.52
VPHD vs. QCFX	.96	.91	.90
VPFX vs. VPFXS	.18	.35	.63
VPFX vs. QCFX	.81	.80	.89
VPFXS vs. QCFX	.17	.053	.44

Table 5.9 *P* values for pairwise comparison of the diagnostic performance of Herder score for the different reconstructions. Results are shown for Herder risk score derived using SUV_{max} for both nodules and reference tissues, SUV_{max} for the nodules and SUV_{mean} for the reference tissues and SUV_{mean} for both nodules and reference tissues.

Classification Method			Reconstruction			
			VPHD	VPFX	VPFXS	QCFX
Nodule SUV_{max} + Ref SUV_{max}	vs.	Nodule SUV_{max} + Ref SUV_{mean}	.67	.74	.78	.41
Nodule SUV_{max} + Ref SUV_{max}	vs.	Nodule SUV_{mean} + Ref SUV_{mean}	.26	.81	.11	.36
Nodule SUV_{max} + Ref SUV_{mean}	vs.	Nodule SUV_{mean} + Ref SUV_{mean}	.35	.64	.25	.17

Table 5.10 *P* values for pairwise comparison of the diagnostic performance of Herder score derived using SUV_{max} for both nodules and reference tissues, SUV_{max} for the nodules and SUV_{mean} for the reference tissues and SUV_{mean} for both nodules and reference tissues. Results are shown for the four different reconstructions.

Optimal cut-offs for the Herder risk score derived using: SUV_{max} for both the nodules and reference tissues; nodule SUV_{max} and SUV_{mean} for the reference tissues; and SUV_{mean} for both nodules and reference tissues are given in Table 5.11 for each reconstruction. The diagnostic performance of the Herder score was slightly worse than when using SUV alone for all methods

of classification. Use of nodule SUV_{max} and SUV_{mean} for the reference tissues demonstrated the greatest difference in cut-offs between reconstructions and had the lowest performance across all reconstructions. When using SUV_{max} for both the nodules and reference tissues, VPFX and VPFXS showed similar performance and the optimal cut-offs were identical. The optimal cut-offs were identical for VPHD, VPFX and QCFX when using SUV_{mean} for both nodules and reference tissues. As was seen previously for the SUV results, QCFX showed a higher sensitivity, but lower specificity compared to the other reconstructions.

	Reconstruction	**Herder Cut-off (%)**	**Specificity (Sp)**	**Sensitivity (Se)**	**Youden's Index**
Nodule SUV_{max}	VPHD	49.5	0.68	0.87	0.56
+ Ref SUV_{max}	VPFX	23.9	0.66	0.94	0.60
	VPFXS	23.9	0.64	0.94	0.58
	QCFX	63.8	0.64	0.88	0.52
Nodule SUV_{max}	VPHD	23.5	0.59	0.94	0.53
+ Ref SUV_{mean}	VPFX	28.2	0.59	0.95	0.54
	VPFXS	50.7	0.61	0.95	0.56
	QCFX	75.7	0.73	0.79	0.51
Nodule SUV_{mean}	VPHD	49.5	0.71	0.88	0.59
+ Ref SUV_{mean}	VPFX	49.5	0.71	0.88	0.59
	VPFXS	23.9	0.61	0.94	0.55
	QCFX	49.5	0.61	0.94	0.55

Table 5.11 Optimal cut-offs for Herder score using the SUV_{max} for both the nodules and reference tissues, nodule SUV_{max} and SUV_{mean} for the reference tissues and SUV_{mean} for both nodules and reference tissues for the four different reconstructions. Cut-offs were derived using Youden's J statistic assuming equal weights for false negative and false positive results.

5.4 Discussion

5.4.1 Assessment of the Impact of New Technologies on Image Quality using the NEMA Phantom

The sphere sizes used for the phantom experiment ranged from 10 mm to 37 mm and were based on those in the real physical phantom which are derived from the specifications in the NEMA NU-2 standard [40] However, the range of lesion sizes used in this thesis for the assessment of diagnostic performance were from 5 mm to 30 mm, therefore in future studies it would be of interest to extend the range to include smaller sphere sizes especially as this is the range where the greatest difference is seen between the new and old technologies. Overall the incorporation of TOF in the OS-EM reconstruction didn't show a noticeable improvement in bias, SNR or CNR compared to the OS-EM without TOF. Both the reconstructions with PSF performed better than the non-PSF reconstructions, with QCFX performing better in terms of minimising bias and VPFXS performing better for SNR and CNR.

5.4.2 Assessment of the Impact of New Technologies on Diagnostic Performance

The AUC when using the Swensen score alone was 0.78, 95 % CI [0.70, 0.85] for the population used in this thesis which had a prevalence of malignancy of 77.3 %. In the original study by Swensen *et al* [171], in which the clinical risk score was derived, the authors reported an AUC of 0.83 ± 0.02, however the prevalence for malignancy in the population studied was much lower at 26.4 %. The later study by Herder *et al* [9] which developed the Herder risk score combining the clinical risk with ^{18}F-FDG-PET uptake had a prevalence of malignancy of 57.5 % and reported an AUC of 0.79, 95 % CI [0.70, 0.85], which was similar to the AUC for the population used in this thesis. In current clinical practice, only a subset of patients with SPNs would be referred for ^{18}F-FDG-PET-CT therefore it is to be expected that populations in studies investigating the usefulness of PET will be biased towards malignant lesions as was the case for the population in

the study by Murphy *et al* [165] whose characteristics were used for the population used in this thesis.

The number of patients used for the investigation of diagnostic performance in this thesis was relatively small (N=194) and there were significantly more malignant (77.3 %) than benign lesions (22.7 %). As such, there were not enough patients available to split the study into different nodule size groups for further analysis. Weir *et al* investigated the impact of nodule size on the diagnosis of SPNs in a prospective multicentre trial [194]. They found as lesion size decreased there was a reduction in sensitivity coupled with an increase in specificity. The authors suggested the use of different diagnostic thresholds according to the nodule size, with optimal SUV_{max} cut-offs of 1.75 for nodules < 12 mm, 2.55 for nodules 12-16 mm and 3.6 for nodules > 16 mm. In this prospective multicentre study, the scanners used reconstructions that were harmonised using the NEMA image quality phantom and prevented the use of PSF and BPL. Based on the results from this thesis, the optimal cut-offs are likely to be different when incorporating these new technologies. It is therefore important if using specific SUV cut-offs to re-validate and adjust as required when using the latest PET technologies. With a larger cohort of patients, the technique used in this thesis could be used in place of real patients for investigating the optimal cut-offs using these new technologies.

An automated method for delineation of the lesion was chosen to define the SUV_{mean} to avoid the requirement of manual delineation by an expert which is time-consuming and can be prone to large inter-observer variability [195, 196]. Use of a threshold based on the maximum voxel value are popular for segmentation of lesions in PET as they are widely available across vendor software and easy to implement, however they do not perform well with small and/or low contrast lesions [95, 197] and the optimal threshold is dependent on the characteristics of the scanner and the lesion-to-background ratio [198, 199]. Instead, a gradient-based algorithm, PETedge+ (MIM Encore v7.1.5, MIM Software Inc, Cleveland, OH), was chosen. To define the volume a large spherical VOI was manually placed over the lesion and automatically resized using the PETedge+ algorithm. Gradient-based methods can also be biased due to spatial resolution blur which distorts the real lesion boundary [96], however they have been shown to be more

accurate and robust compared to thresholding techniques particularly in small and/or low contrast lesions [200].

The PERCIST guidelines recommend using SUV_{peak} which consists of a 1 cm^3 sphere in the hottest region of the lesion which doesn't necessarily include the SUV_{max} [87]. However, due to the small size of many of the lesions in this study, the SUV_{peak} volume would extend beyond the lesion uptake and include lung background within the volume losing the benefit of increasing the resolution with the new PET technologies. An alternative method has been proposed by Hasenclever *et al* and has been applied for assessment of residual uptake for patients with Hodgkin's Lymphoma [91]. In this case, the authors defined the SUV_{peak} as the average over the maximum SUV voxel and the three hottest adjacent ones.

In a recent study of lymphoma patients, Zwezerijnen *et al* [201] concluded that the liver SUV_{mean} was the most reproducible metric compared to SUV_{max} and SUV_{peak} when considering VOI size, location, reconstruction protocol and image noise level. The authors recommended the use of a 3 cm diameter spherical VOI to determine the SUV_{mean}, as was used in this thesis, and it is recommended for use in clinical trials. Whilst tumour SUV_{max} has been shown to have high inter-observer reproducibility for the same image [202], it has been shown to have poor inter-study reproducibility compared to SUV_{mean} [203, 204]. In a meta-analysis by Lodge *et al*, repeatability for SUV_{max} and SUV_{mean} were found to be similar [205]. This was due to the fact that whilst SUV_{mean} uses volume averaging across voxels, and therefore is less influenced by noise than SUV_{max}, delineation of tumours was less consistent than selection of the single hottest voxel.

In the original study by Herder *et al*, the authors used visual grading for grouping the SPN uptake according to a 4-point ordinal scale of absent, faint, moderate and intense [9]. No further detail was provided by the authors to indicate how the visual grading was performed making it difficult to reproduce as the scale is subjective and may be different depending on the readers experience and thresholding of the PET images. The BTS guidelines recommend qualitative assessment using the following four-point scale [7]:

- Absent: uptake indiscernible from background lung tissue
- Faint: uptake less than or equal to mediastinal blood pool
- Moderate: uptake greater than mediastinal blood pool
- Intense: uptake markedly greater than mediastinal blood pool

Again, this scale could be interpreted differently by different readers, particularly when differentiating between moderate and intense uptake. For the purposes of this thesis, the SPN uptake was classified according to the ordinal scale provided by Murphy *et al* [165]. This scale closely matches the BTS guideline, except it specifies the use of SUV_{max} as cut-offs for the reference tissues and intense uptake is defined as > 2 x the liver SUV_{max}. The BTS scale is widely used in UK clinical practice and has been shown to have good interobserver agreement [180, 179]. In the multi-observer study by Fatania *et al* [180], the authors found no significant difference in the diagnostic accuracy for classifying SPN ^{18}F-FDG uptake according to the BTS four-point scale (AUC=0.768) compared to using SUV_{max} (AUC=0.794), *P*=.43. Therefore it is reasonable to expect the results from this thesis using the SUV_{max} to confirm the classification of SPN uptake reflect those of visual assessment using the BTS scale which is more representative of clincial practice.

The Youden's J statistic was used to derive the optimal cut-offs for the SUV and Herder risk scores. The index applies equal weight to the sensitivity and specificity, however, use of a diagnostic threshold that favours sensitivity over specificity is generally used in the clinical management of SPNs to minimise the number of false-negative test results. This is because the consequences of missing a malignant SPN and missing an opportunity for treatment is more undesirable than the consequences of a false-positive test result which may involve an unnecessary biopsy or surgery [11].

5.5 Conclusions

The uptake measured for the reference tissues using SUV_{max} was found to be variable across the different reconstructions and did not reflect the results for background variability (BV) and image roughness (IR) from the phantom measurements. The SUV_{mean} however appeared to be more stable across reconstructions for both the MBP and liver and may prove a more reliable measurement.

Compared to the Swensen score alone, incorporation of the degree of ^{18}F-FDG resulted in an improved diagnostic performance regardless of the reconstruction used or the method of measurement. Interestingly, use of SUV without incorporation of the clinical and radiological characteristics demonstrated better diagnostic performance, suggesting that the weighting of the PET component in the Herder score equation could be increased to improve the diagnostic accuracy further. There was no significant difference seen in the diagnostic performance of the Herder score when using TOF, PSF modelling or the BPL reconstruction although optimal cut-offs were different depending on the method of uptake classification used and the reconstruction.

There were significant differences between the diagnostic performance for the different technologies when using SUV particularly for the BPL reconstruction which also demonstrated the lowest AUC for both SUV_{max} and SUV_{mean}. This is due to the fact that whilst the sensitivity increased, the specificity also decreased compared to the other reconstructions. Ultimately, this is due to the fact that ^{18}F-FDG is a non-specific radiotracer and by improving the recovery of the signal in smaller objects, more benign lesions are being classified as malignant. Use of appropriate cut-offs for newer technologies would therefore be required to ensure that there isn't an unacceptable decrease in specificity resulting in significant false positives. Results showed no significant difference in the diagnostic performance of nodule SUV_{max} and SUV_{mean} when using the same reconstruction suggesting either metric could be used for measuring nodule uptake.

Chapter 6

Overall Conclusions & Discussion

6.1 Overall Conclusions

Continuous improvements in PET scanner instrumentation and more accurate corrections within the system model have led to a range of technologies in clinical use. Evaluation of the impact of these technologies on clinically relevant measures is challenging, particularly as repeat scanning of individuals is unethical due to the risks of ionising radiation. Current methods of assessment often use small cohorts of patients where the ground truth is unknown or phantom studies which don't take account of the wide variation in patient sizes and noise levels observed in patients. Novel methods for assessing the impact of new technologies on clinically relevant measures are needed to ensure they are implemented appropriately. The aim of this thesis was to develop the methodology for the simulation of realistic lesions and their insertion into patient datasets to allow the generation of a large number of PET images that could then be used in place of real PET scans for evaluating new technologies in virtual imaging trials. To ensure the images generated using the proposed methodology could be reliably used in place of patient PET images acquired on the real PET scanner, the methodology was validated by comparing the diagnostic performance for a cohort of patients with simulated lesions against a cohort of patients with known solitary pulmonary nodules (SPNs). A virtual imaging trial was then performed to investigate the use of time-of-flight (TOF), point spread function (PSF) modelling

and a Bayesian penalised likelihood (BPL) reconstruction algorithm on quantitative measures of ^{18}F-FDG uptake and the diagnostic performance in separating malignant and benign lesions for a cohort of patients with simulated SPNs.

A model of the clinical PET system was designed using the GATE toolkit and this was used for performing Monte Carlo simulations of the lesions in this thesis. The toolkit incorporates well-validated physics models which can accurately simulate an object containing radioactivity and the response of the imaging system. To generate images as close as possible to the real scanner, data from the simulations were binned into sinograms and inserted into existing real PET data accounting for attenuation, normalisation, randoms and intrinsic radiation prior to reconstruction using the GE Duetto toolbox. The performance of the model was tested using simulated phantoms following the NEMA NU-2 standard. The results presented in Chapter 3 showed the spatial resolution for the model was within 1.00 mm of the real PET scanner when using the clinical reconstruction (OS-EM with TOF, 2 iterations, 24 subsets and 6.4mm Gaussian post filter). The model also had identical system sensitivity and the axial slice sensitivity was within 1.6 ± 2.3 % of the real PET scanner. Real pulmonary nodules from 10 patients were segmented and used as input for simulating lesions that were then inserted into the contralateral lung of the same patient. This method was shown to be able to generate clinically realistic lesions with SUV_{max} and SUV_{mean} measurements within -0.20 ± 0.29 and 0.00 ± 0.19 of the real lesions they were modelled on.

The next stage of the thesis was to show that the proposed methodology could generate PET images combining the physiological and anatomical variation observed in patients with realistic simulated lesions where the characteristics (^{18}F-FDG uptake, size, appearance and location) were defined using the those from real tumours. Validation of the methodology consisted of comparing the diagnostic performance for a virtual population of patients with simulated lesions against a real cohort of patients scans with known SPNs. SUV_{max} measurements for the simulated lesions showed good agreement with the SPNs and there was no significant difference in the diagnostic performance for predicting malignancy. As such, the novel method described in this thesis is

suitable for generating virtual PET images to use as a surrogate for populations of patients in virtual imaging trials.

One of the key aims of this thesis was to evaluate the impact of TOF, PSF modelling and a BPL reconstruction algorithm on image quality and diagnostic accuracy. First a phantom assessment was performed to investigate the impact of these technologies on noise, signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR) as well as bias in quantitative measures. Next, PET images for a population of patients with simulated SPNs were generated using the novel method described in Chapters 3 and 4. The diagnostic accuracy in determining malignancy was assessed using SUV alone and in combination with the clinical and radiological characteristics using the Herder risk score. The phantom results demonstrated incorporation of PSF and use of the BPL reconstruction outperformed OS-EM with and without TOF. The OS-EM with PSF showed the maximum SNR and CNR across all sphere sizes, whilst the BPL reconstruction demonstrated the smallest bias. This suggests that use of PSF would improve diagnostic accuracy for lesion detection tasks, whereas the BPL reconstruction would be optimal for accurate quantification. In the assessment of SPNs the location of the lesions is already known from previous radiological imaging and the degree of PET uptake is used to help in classifying the lesions as malignant or benign. The diagnostic performance of PET for classifying the SPNs was shown to be better than using clinical and radiological characteristics alone and this was true for all technologies and SUV metrics. Overall, use of SUV performed better when used alone rather than incorporating with the clinical and radiological characteristics. This is likely due to the fact that the original Herder score was derived using old PET technology consisting of a PET-only bismuth germanate (BGO) system acquiring in 2D mode and the diagnostic sensitivity has significantly improved with newer 3D PET-CT systems. The reduction in diagnostic performance when using the Herder score may however be balanced out by the fact that there was no significant difference in the results seen for the different reconstructions or the different SUV metrics.

The majority of publications investigating the diagnostic accuracy of ^{18}F-FDG-PET in determining the risk of malignancy in SPNs omitted details about the methodology used for analysis of the PET uptake or were missing important technical information about the PET

acquisition and reconstruction. However, determination of uptake using visual methods alone has been shown to have good interobserver reproducibility when using the four-point scale recommended by the BTS guidelines [180] and represents current UK clinical practice [7]. There are several technical, physical and biological factors that can affect quantification using SUVs [176] however with careful standardisation of protocols and scanner harmonisation, the within-subject repeatability of tumour SUVs can be reduced to give a coefficient of variation of approximately 10 % [205]. Published studies should include details about the PET acquisition and reconstruction parameters used to provide the reader with an idea of the repeatability and technology included. Recommended information to include would be whether the study was single- or multi-centre and, if the latter, whether harmonisation of scanners and protocols was performed. There are a number of professional bodies that promote harmonisation schemes for PET, such as the European Association of Nuclear Medicine Research Ltd (EARL) accreditation program, the Society of Nuclear Medicine and Molecular Imaging (SNMMI) Clinical Trials Network and the American College of Radiology (ACR) Imaging Network, all of which have slightly different standards and use various phantoms so details of the scheme should be included or if a local harmonisation methodology was used this should be provided as supplemental information. A full table of scanner models and reconstruction parameters used in the study would be ideal, but in cases where this isn't feasible, the data should at least describe the range of technology included in the study: for example PET-CT and PET-MR, TOF, PSF, BPL or solid state photon detectors.

There are a number of clinical applications for PET in oncology [1]. The focus of the work in this thesis was to investigate the impact of new PET technologies on the characterisation of SPNs using a four-point scale. Similarly, response assessment of lymphoma uses a five-point scale, known as the Deauville score [206], which compares the uptake in the lymph nodes to reference tissues. Recent studies have shown management changes in 4.3 % of patients when using BPL [207] and 5 % for OS-EM with PSF [208] when compared to OS-EM alone. The current guidelines for management of lymphoma patients were developed before the introduction of these new PET technologies. Virtual imaging trial (VIT) could therefore have a role in

predicting the impact of these technologies on quantitative assessment of treatment response in lymphoma and subsequent changes in therapeutic decisions without the need to perform large-scale clinical trials. The novel methodology described in this thesis could be applied to perform these VITs. A statistical model such as that described by Tomei *et al* [209] or a generative adversarial network (GAN) (see Section 4.4) could be used to generate lymphoma lesions with realistic sizes, locations and uptakes to be used as input for the Monte Carlo model.

PET-CT is also used in staging and detection of recurrence or metastatic disease for a number of oncological applications. As discussed in Chapter 2, improvements in PET hardware and software have led to increased spatial resolution and contrast recovery providing better small lesion detection [210]. To predict the performance of human observers in task-based assessments, such as lesion detection, the PET data generated using the methodology described in this thesis could be used to develop mathematically derived model observers [211] or convolutional neural network-based model observers, such as that described by Han *et al* [212]. These could then be applied to datasets reconstructed using different algorithms or using advanced corrections removing the reliance on experienced reviewers.

3D PET scans are typically acquired over 2-3 minutes with patients breathing freely, whilst the CT is acquired within a fraction of the respiratory cycle. As a result of respiratory motion, the uptake in lesions is blurred on PET and can be mis-registered with CT leading to inaccurate attenuation correction [156]. Respiratory gating using external devices, consisting of a camera or pressure belt, that track movement of the chest wall have been available for some years. However, clinical adoption has been limited due to the complexities of the device set up [7]. More recently data-driven respiratory gating (DDG) has become available on clinical PET systems, where software is used to extract the respiratory waveform from the PET data, removing the need for an external device. In recent studies, the performance of two DDG solutions was compared to external gating devices. Büther *et al* retrospectively assessed the Siemens DDG solution on the Biograph mCT for 56 patients with suspected malignancies in the thorax and abdomen [213]. The authors found the DDG method was able to generate accurate respiratory signals and produce results comparable to the ANZAI belt system (Anzai Medical Co. Ltd, Tokyo, Japan). In another

retrospective study by Walker *et al*, the DDG solution on the GE Discovery 690 and 710 was compared to the Varian camera-based, real-time position management (RPM) system (Varian Medical Systems) in 144 patients who underwent wholebody ^{18}F-FDG-PET-CT [214]. Results showed the DDG system had a lower failure rate compared to the external device, a larger increase in SUV_{max} and improved image quality as rated by an experienced radiologist. Routine use of DDG in four-dimensional (4D) PET-CT has the potential to improve lesion detection and quantification, particularly in small pulmonary nodules located close to the diaphragm where respiratory motion is greatest [215]. VITs could be used in this context to compare 3D and 4D-PET-CT along with advanced reconstruction algorithms and incorporation of PSF to assess the impact of on the diagnostic accuracy for characterisation of SPNs. The lesion simulation methodology described in this thesis could be adapted to incorporate respiratory motion by using real patient respiratory patterns to split the simulations into shorter time frames based on the respiratory cycle and shifting the voxelised phantom containing the lesion between frames.

With careful validation the use of artificial intelligence (AI) is likely to become widely adopted for the formation of images and reconstruction in medical imaging, including PET [216]. By training with existing PET data, AI can take advantage of the vast amount of prior information and incorporate this in the image reconstruction to improve spatial resolution and reduce noise. Further, the AI can learn to optimise the image for a given clinical task. AI also has a role in the interpretation of medical images and, with validation and regulatory approval, adoption of commercial AI applications for medical image assessment will lead to more accurate and efficient use of data for diagnosis and characterisation of diseases. This has the potential to improve patient management and reduce unnecessary additional imaging or interventions. In particular, there have been a number of AI applications developed to detect and characterise pulmonary nodules based on CT from lung cancer screening [217]. In the UK an NIHR-funded multi-centre prospective observational cohort study, Determining the Impact of Optellum's Lung Cancer Prediction Solution (DOLCE), has just opened. The intention of the study is to compare physician decision making using the current standard of care with an AI-based lung cancer prediction (LCP) tool in 5-30mm solid and part-solid pulmonary nodules. If the study shows

there are clinical and health-economic benefits of the LCP tool this will provide the evidence to support clinical translation of this emerging AI technology.

The latest generation of clinical PET-CT scanners now incorporate silicon photomultiplier (SiPM) detectors in place of photomultiplier tubes (PMTs) [218]. As a result, the TOF resolution, spatial resolution and sensitivity have significantly improved compared to the equivalent PMT-based system evaluated in this thesis. Further work looking at these newer systems would be of great interest as data is currently very limited in the literature and a number of UK sites are installing these systems. With continued manufacturer co-operation, modification of the model in GATE would be feasible to set up the latest GE MI PET-CT scanner (General Electric Medical Systems, GEMS, Milwaukee, WI). Access to the raw and reconstructed PET data for patients scanned on the GE MI would need to be sought, but if that could be achieved, virtual imaging trials using the methodology proposed in this thesis could be performed to investigate how the significant improvements in TOF resolution, spatial resolution and sensitivity would impact on clinically relevant measures. This would be particularly valuable in the context of clinical trials either for helping inform the appropriate technical inclusion/exclusion criteria when designing PET studies, or for deciding whether to include certain PET data retrospectively.

References

1. Dizdarevic, S., Scarsbrook, A. & Barrington, S. *Evidence-based indications for the use of PET-CT in the United Kingdom 2022* tech. rep. (The Royal College of Radiologists, Royal College of Physicians, British Nuclear Medicine Society, Administration of Radioactive Substances Advisory Committee, 2022).
2. Wahl, R. L., Jacene, H., Kasamon, Y. & Lodge, M. A. From RECIST to PERCIST: Evolving Considerations for PET Response Criteria in Solid Tumors. *Journal of Nuclear Medicine* **50,** 122S–150S (2009).
3. Troost, E. G. C., Thorwarth, D. & Oyen, W. J. G. Imaging-Based Treatment Adaptation in Radiation Oncology. *Journal of Nuclear Medicine* **56,** 1922–1929 (2015).
4. Gallamini, A., Zwarthoed, C. & Borra, A. Positron emission tomography (PET) in oncology. *Cancers* **6,** 1821–1889 (2014).
5. Johnson, P. *et al.* Adapted Treatment Guided by Interim PET-CT Scan in Advanced Hodgkin's Lymphoma. *New England Journal of Medicine* **374,** 2419–2429 (2016).
6. Radford, J. *et al.* Results of a Trial of PET-Directed Therapy for Early-Stage Hodgkin's Lymphoma. *New England Journal of Medicine* **372,** 1598–1607 (2015).
7. Callister M E J. Baldwin, D. R. A. A. R. BTS Guidelines for the Investigation and Management of Pulmonary Nodules. *Thorax* **70** (2015).
8. McWilliams, A. *et al.* Probability of Cancer in Pulmonary Nodules Detected on First Screening CT. *New England Journal of Medicine* **369,** 910–919 (2013).
9. Herder, G. J. *et al.* Clinical Prediction Model To Characterize Pulmonary Nodules. *Chest* **128,** 2490–2496 (2005).
10. Al-Ameri, A. *et al.* Risk of malignancy in pulmonary nodules: A validation study of four prediction models. *Lung Cancer* **89,** 27–30 (2015).
11. Gould, M. K., Maclean, C. C., Kuschner, W. G., Rydzak, C. E. & Owens, D. K. Accuracy of Positron Emission Tomography for Diagnosis of Pulmonary Nodules and Mass Lesions. *JAMA* **285,** 914 (7 Feb. 2001).
12. Maisey, M. & Dakin, M. The First 5 Years of a Dedicated Clinical PET Centre. *Clinical Positron Imaging* **1,** 59–69 (1998).
13. Bettinardi, V. *et al.* Physical Performance of the new hybrid PET-CT Discovery 690. *Medical Physics* **38,** 5394–5411 (2011).
14. Jakoby, B. W. *et al.* Physical and clinical performance of the mCT time-of-flight PET/CT scanner. *Physics in Medicine and Biology* **56,** 2375–2389 (2011).
15. Rausch, I. *et al.* Performance evaluation of the Biograph mCT Flow PET/CT system according to the NEMA NU2-2012 standard. *EJNMMI Physics* **2,** 26 (2015).

16. Moses, W. W. Fundamental limits of spatial resolution in PET. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **648,** S236–S240 (2011).

17. Muehllehner, G. & Karp, J. S. Positron emission tomography. *Physics in Medicine and Biology* **51,** R117–R137 (2006).

18. Van der Vos, C. S. *et al.* Quantification, improvement, and harmonization of small lesion detection with state-of-the-art PET. *European Journal of Nuclear Medicine and Molecular Imaging* **44,** 4–16 (2017).

19. Weber, W. A. Is There Evidence for Evidence-Based Medical Imaging? *Journal of Nuclear Medicine* **52,** 74S–76S (2011).

20. Hillner, B. E. *et al.* The National Oncologic PET Registry (NOPR): Design and Analysis Plan. *Journal of Nuclear Medicine* **48,** 1901–1908. ISSN: 0161-5505 (Nov. 2007).

21. Hicks, R. J. The Injustice of Being Judged by the Errors of Others: The Tragic Tale of the Battle for PET Reimbursement. *Journal of Nuclear Medicine* **59,** 418–420. ISSN: 0161-5505 (Mar. 2018).

22. Barrett, H. H., Myers, K. J., Hoeschen, C., Kupinski, M. A. & Little, M. P. Task-based measures of image quality and their relation to radiation dose and patient risk. *Physics in Medicine and Biology* **60,** R1–R75 (2015).

23. Elangovan, P. *et al.* A Modelling Framework for Evaluation of 2D-Mammography and Breast Tomosynthesis Systems. *Maidment A.D.A., Bakic P.R., Gavenonis S. (eds) Breast Imaging. IWDM 2012. Lecture Notes in Computer Science* **7361,** 338–345 (2012).

24. Elangovan, P. *et al.* Development and validation of a modelling framework for simulating 2D-mammography and breast tomosynthesis images. *Physics in Medicine and Biology* **59,** 4275–4293 (2014).

25. Harrison, R. L. *et al.* A Virtual Clinical Trial of FDG-PET Imaging of Breast Cancer: Effect of Variability on Response Assessment. *Translational Oncology* **7,** 138–146 (2014).

26. Wangerin, K. A. *et al.* Effect of 18F-FDG Uptake Time on Lesion Detectability in PET Imaging of Early-Stage Breast Cancer. *Tomography* **1,** 53–60 (2015).

27. Wangerin, K. A. *et al.* A virtual clinical trial comparing static versus dynamic PET imaging in measuring response to breast cancer therapy. *Physics in Medicine and Biology* **62,** 3639–3655 (2017).

28. Kurland, B. F. *et al.* Multicenter Clinical Trials Using 18F-FDG PET to Measure Early Response to Oncologic Therapy: Effects of Injection-to-Acquisition Time Variability on Required Sample Size. *Journal of Nuclear Medicine* **57,** 226–230 (2016).

29. Venkatesh, K. P., Raza, M. M. & Kvedar, J. C. Health digital twins as tools for precision medicine: Considerations for computation, implementation, and regulation. *npj Digital Medicine* **5.** ISSN: 23986352 (2022).

30. Rahmim, A. *et al.* Theranostic digital twins for personalized radiopharmaceutical therapies: Reimagining theranostics via computational nuclear oncology. *Frontiers in Oncology* **12,** 1–7. ISSN: 2234-943X (Dec. 2022).

31. Warburg, O. Über den Stoffwechsel der Carcinomzelle. *Die Naturwissenschaften* **12,** 1131–1137. ISSN: 0028-1042 (Dec. 1924).

32. Conti, M. & Eriksson, L. Physics of pure and non-pure positron emitters for PET: A review and a discussion. *EJNMMI Physics* **3** (2016).

33. Levin, C. S. & Hoffman, E. J. Calculation of positron range and its effect on the fundamental limit of positron emission tomography system spatial resolution spatial resolution. *Phys. Med. Biol.* **44,** 781 (1999).

34. Van Sluis, J. *et al.* Performance Characteristics of the Digital Biograph Vision PET/CT System. *Journal of Nuclear Medicine* **60,** 1031–1036 (2019).

35. Watson, C. C. *et al.* NEMA NU 2 Performance Tests for Scanners with Intrinsic Radioactivity. *Journal of nuclear medicine* **45,** 822–826 (2004).

36. Anger, H. O. Scintillation camera. *Review of Scientific Instruments* **29,** 27–33 (1958).

37. Boellaard, R. *et al.* FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *European Journal of Nuclear Medicine and Molecular Imaging* **42,** 328–354 (2015).

38. Ross, S., Stearns, C., Manjeshwar, R. & Iatrou, M. *A method of overlap correction for fully 3D OSEM reconstruction of PET data* in *IEEE Symposium Conference Record Nuclear Science 2004.* **6** (IEEE, 2005), 3497–3500.

39. Soret, M., Bacharach, S. L. & Buvat, I. Partial-Volume Effect in PET Tumor Imaging. *Journal of Nuclear Medicine* **48,** 932–945 (2007).

40. NEMA. *NEMA NU 2-2018 Performance Measurements of Positron Emission Tomographs (PET)* tech. rep. (National Electrical Manufacturers Association, 2018).

41. Erlandsson, K., Buvat, I., Pretorius, P. H., Thomas, B. A. & Hutton, B. F. A review of partial volume correction techniques for emission tomography and their applications in neurology, cardiology and oncology. *Physics in Medicine and Biology* **57,** R119–R159 (2012).

42. Cysouw, M. C. *et al.* Accuracy and precision of partial-volume correction in oncological PET/CT studies. *Journal of Nuclear Medicine* **57,** 1642–1649 (2016).

43. Conti, M. & Bendriem, B. The new opportunities for high time resolution clinical TOF PET. *Clinical and Translational Imaging* **7,** 139–147 (2019).

44. Qi, J. & Leahy, R. M. Iterative reconstruction techniques in emission computed tomography. *Phys. Med. Biol. Phys. Med. Biol* **51,** 541–578 (2006).

45. Colsher, J. G. Fully-three-dimensional positron emission tomography. *Phys. Med. Biol* **25,** 103–104 (1980).

46. Cherry, S. R. & Dahlbom, M. in *PET* (ed Phelps, M.) 1–117 (Springer New York, New York, NY, 2006). ISBN: 978-0387-32302-2. http://link.springer.com/10.1007/0-387-34946-4_1.

47. Shepp, L. A. & Vardi, Y. Maximum Likelihood Reconstruction for Emission Tomography. *IEEE Transactions on Medical Imaging* (1982).

48. Lange, K. & Carson, R. EM reconstruction algorithms for emission and transmission tomography. *Journal of computer assisted tomography* (1984).

49. Hudson, H. M. & Larkin, R. S. Ordered Subsets of Projection Data. *IEEE transactions on medical imaging* **13,** 601–609 (1994).

50. Noise properties of the EM algorithm. I. Theory. *Physics in Medicine and Biology* **39,** 833–846 (1994).

51. Chen, C.-T. *et al.* Bayesian Image Reconstruction in Positron Emission Tomography. *IEEE Transactions on Nuclear Science* **37,** 636–641 (1990).

52. Mumcuoğlu, E., Leahy, R., Zhou, Z. & Cherry, S. R. Fast Gradient-Based Methods for Bayesian Reconstruction of Transmission and Emission PET Images. *IEEE Transactions on Medical Imaging* **13,** 687–701 (1994).

53. Ross, S. *Q.Clear [Whitepaper]* GE Healthcare, Waukesha, WI, 2014. www.gehealthcare.com.

54. Nuyts, J., Bequé, D., Dupont, P. & Mortelmans, L. A concave prior penalizing relative differences for maximum-a-posteriori reconstruction in emission tomography. *IEEE Transactions on Nuclear Science* **49,** 56–60 (2002).

55. Teoh, E. J., McGowan, D. R., Macpherson, R. E., Bradley, K. M. & Gleeson, F. V. Phantom and Clinical Evaluation of the Bayesian Penalized Likelihood Reconstruction Algorithm Q.Clear on an LYSO PET/CT System. *Journal of Nuclear Medicine* **56,** 1447–1452 (2015).

56. Burger, C. *et al.* PET attenuation coefficients from CT images: Experimental evaluation of the transformation of CT into PET 511-keV attenuation coefficients. *European Journal of Nuclear Medicine* **29,** 922–927 (2002).

57. Theodorakis, L. *et al.* A review of PET normalization: Striving for count rate uniformity. *Nuclear Medicine Communications* **34,** 1033–1045 (2013).

58. Kinahan, P. E., Townsend, D. W. & Bailey, D. L. Efficiency normalization techniques for 3D PET data. *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.* **2,** 1021–1025 (1995).

59. Badawi, R. D. & Marsden, P. K. Developments in component-based normalization for 3D PET. *Physics in medicine and biology* **44,** 571–94 (1999).

60. Badawi, R. D., Lodge, M. A. & Marsden, P. K. Algorithms for calculating detector efficiency normalization coefficients for true coincidences in 3D PET. *Physics in Medicine and Biology* **43,** 189–205 (1998).

61. Meikle, S. R. & Badawi, R. D. in *Positron Emission Tomography* 93–126 (Springer-Verlag, London, 2006). http://link.springer.com/10.1007/1-84628-007-9_5.

62. Stearns, C. *et al. Random coincidence estimation from single event rates on the Discovery ST PET/CT scanner* in *2003 IEEE Nuclear Science Symposium Conference Record* **5** (2003), 3067–3069.

63. Stearns, C. W. & Lonn, A. H. R. *Randoms from singles estimation for long PET scans* in *2011 IEEE Nuclear Science Symposium Conference Record* (2011), 3739–3741.

64. Rahmim, A., Qi, J. & Sossi, V. Resolution modeling in PET imaging: Theory, practice, benefits, and pitfalls. *Medical Physics* **40,** 064301 (2013).

65. Thielemans, K. *et al. Impact of PSF modelling on the convergence rate and edge behaviour of EM images in PET* in *IEEE Nuclear Science Symposuim & Medical Imaging Conference* (IEEE, 2010), 3267–3272.

66. Andersen, F. L., Klausen, T. L., Loft, A., Beyer, T. & Holm, S. Clinical evaluation of PET image reconstruction using a spatial resolution model. *European Journal of Radiology* **82,** 862–869 (2013).

67. Schaefferkoetter, J., Casey, M., Townsend, D. & El Fakhri, G. Clinical impact of time-of-flight and point response modeling in PET reconstructions: a lesion detection study. *Physics in Medicine and Biology* **58,** 1465–1478 (2013).

68. Munk, O. L., Tolbod, L. P., Hansen, S. B. & Bogsrud, T. V. Point-spread function reconstructed PET images of sub-centimeter lesions are not quantitative. *EJNMMI Physics* **4,** 1–12 (2017).

69. Barrington, S. F., Sulkin, T., Forbes, A. & Johnson, P. W. All that glitters is not gold - new reconstruction methods using Deauville criteria for patient reporting. *European Journal of Nuclear Medicine and Molecular Imaging* **45,** 316–317 (2018).

70. Leveque, L. *et al.* On the Subjective Assessment of the Perceived Quality of Medical Images and Videos. *2018 10th International Conference on Quality of Multimedia Experience, QoMEX 2018* (2018).

71. Sah, B.-R. R. *et al.* Clinical evaluation of a block sequential regularized expectation maximization reconstruction algorithm in 18F-FDG PET/CT studies. *Nuclear medicine communications* **38,** 57–66 (2017).

72. Chilcott, A. K., Bradley, K. M. & McGowan, D. R. Effect of a Bayesian Penalized Likelihood PET Reconstruction Compared With Ordered Subset Expectation Maximization on Clinical Image Quality Over a Wide Range of Patient Weights. *American Journal of Roentgenology* **210,** 153–157 (2018).

73. Chang, T. *et al.* Effects of injected dose, BMI and scanner type on NECR and image noise in PET imaging. *Physics in Medicine and Biology* **56,** 5275–5285 (2011).

74. Rogasch, J. M. *et al.* Moving the goalposts while scoring—the dilemma posed by new PET technologies. *European Journal of Nuclear Medicine and Molecular Imaging* (2021).

75. Rahmim, A. & Tang, J. Noise propagation in resolution modeled PET imaging and its impact on detectability. *Physics in medicine and biology* **58,** 6945–68. eprint: NIHMS150003 (2013).

76. Tong, S., Alessio, A. M. & Kinahan, P. E. Noise and signal properties in PSF-based fully 3D PET image reconstruction: an experimental evaluation. *Physics in medicine and biology* **55,** 1453–1473 (2010).

77. Strother, S., Casey, M. & Hoffman, E. Measuring PET scanner sensitivity: relating countrates to image signal-to-noise ratios using noise equivalents counts. *IEEE Transactions on Nuclear Science* **37,** 783–788 (1990).

78. Queiroz, M. A., Wollenweber, S. D., von Schulthess, G., Delso, G. & Veit-Haibach, P. Clinical image quality perception and its relation to NECR measurements in PET. *EJNMMI physics* **1,** 103 (2014).

79. Hoffman, E. J., Huang, S.-C. & Phelps, M. E. Quantitation in Positron Emission Computed Tomography: 1. Effect of Object Size. *Journal of computer assisted tomography* **3,** 299–308 (1979).

80. Frings, V. *et al.* Repeatability of metabolically active volume measurements with 18F-FDG and 18F-FLT PET in non-small cell lung cancer. *Journal of Nuclear Medicine* **51,** 1870–1877 (2010).

81. Boellaard, R., Krak, N. C., Hoekstra, O. S. & Lammertsma, A. A. Effects of noise, image resolution, and ROI definition on the accuracy of standard uptake values: a simulation study. *Journal of nuclear medicine* **45,** 1519–27 (2004).

82. Trevethan, R. Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice. *Frontiers in Public Health* **5,** 1–7 (2017).

83. Obuchowski, N. A. & Bullen, J. A. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Physics in Medicine & Biology* **63** (2018).

84. Halligan, S., Altman, D. G. & Mallett, S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *European Radiology* **25,** 932–939. ISSN: 0938-7994 (Apr. 2015).

85. Vickers, A. J., Van Calster, B. & Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ (Online)* **352,** 3–7. ISSN: 17561833 (2016).

86. Aide, N. *et al.* EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. *European Journal of Nuclear Medicine and Molecular Imaging* **44,** 17–31 (2017).

87. Joo Hyun, O., Lodge, M. A. & Wahl, R. L. Practical percist: A simplified guide to PET response criteria in solid tumors 1.0. *Radiology* **280,** 576–584 (2016).

88. Kinahan, P. E. & Fletcher, J. W. Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy. *Seminars in Ultrasound, CT and MRI* **31,** 496–505 (2010).

89. Lodge, M. A., Chaudhry, M. A. & Wahl, R. L. Noise Considerations for PET Quantification Using Maximum and Peak Standardized Uptake Value. *Journal of Nuclear Medicine* **53,** 1041–1047 (2012).

90. Westerterp, M. *et al.* Quantification of FDG PET studies using standardised uptake values in multi-centre trials: Effects of image reconstruction, resolution and ROI definition parameters. *European Journal of Nuclear Medicine and Molecular Imaging* **34,** 392–404 (2007).

91. Hasenclever, D. *et al.* qPET - A quantitative extension of the Deauville scale to assess response in interim FDG-PET scans in lymphoma. *European Journal of Nuclear Medicine and Molecular Imaging* **41,** 1301–1308 (2014).

92. Henriksson, E. *et al.* 2-Deoxy-2-[18F]fluoro-D-glucose uptake and correlation to intratumoral heterogeneity. *Anticancer Research* **27,** 2155–2159 (2007).

93. Mikhaeel, N. G. *et al.* Combination of baseline metabolic tumour volume and early response on PET/CT improves progression-free survival prediction in DLBCL. *European Journal of Nuclear Medicine and Molecular Imaging* **43,** 1209–1219 (2016).

94. Barrington, S. F. & Meignan, M. Time to Prepare for Risk Adaptation in Lymphoma by Standardizing Measurement of Metabolic Tumor Burden. *Journal of Nuclear Medicine* **60,** 1096–1102 (2019).

95. Foster, B., Bagci, U., Mansoor, A., Xu, Z. & Mollura, D. J. A Review on Segmentation of Positron Emission Tomography Images. *Compt Biol Med* **50,** 76–96 (2014).

96. Hatt, M. *et al.* Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. *Medical Physics* **44,** e1–e42 (2017).

97. Valladares, A., Beyer, T. & Rausch, I. Physical imaging phantoms for simulation of tumor heterogeneity in PET, CT, and MRI: An overview of existing designs. *Medical Physics* **47,** 2023–2037 (2020).

98. Barrington, S. F. *et al.* Establishment of a UK-wide network to facilitate the acquisition of quality assured FDG-PET data for clinical trials in lymphoma. *Annals of Oncology* **22,** 739–745 (2011).

99. Lee, C. & Lee, J.-K. Computational anthropomorphic phantoms for radiation protection dosimetry: evolution and prospects. *Nuclear engineering and technology* **38,** 239 (2006).

100. Xu, X. G. An exponential growth of computational phantom research in radiation protection, imaging, and radiotherapy: a review of the fifty-year history. *Physics in Medicine and Biology* **59,** R233–R302 (2014).

101. ICRP. *ICRP Publication 110: Adult Reference Computational Phantoms* tech. rep. (International Commission on Radiological Protection, 2010).

102. Segars, W. P., Sturgeon, G., Mendonca, S., Grimes, J. & Tsui, B. M. 4D XCAT phantom for multimodality imaging research. *Medical Physics* **37,** 4902–4915 (2010).

103. Segars, W. P. *et al.* Application of the 4-D XCAT Phantoms in Biomedical Imaging and Beyond. *IEEE Transactions on Medical Imaging* **37,** 680–692 (2018).

104. Buvat, I. & Lazaro, D. Monte Carlo simulations in emission tomography and GATE: An overview. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **569,** 323–329 (2006).

105. Gillam, J. E. & Rafecas, M. Monte-Carlo simulations and image reconstruction for novel imaging scenarios in emission tomography. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **809,** 76–88 (2016).

106. Buvat, I. & Castiglioni, I. Monte Carlo simulations in SPET and PET. *Quarterly Journal of Nuclear Medicine* **46,** 48–61 (2002).

107. Agostinelli, S. *et al.* Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506,** 250–303 (3 2003).

108. Allison, J. *et al.* Geant4 developments and applications. *IEEE Transactions on Nuclear Science* **53,** 270–278 (2006).

109. Allison, J. *et al.* Recent developments in Geant4. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **835,** 186–225 (2016).

110. Jan, S. *et al.* GATE: a simulation toolkit for PET and SPECT. *Physics in Medicine and Biology* **49,** 4543–4561 (2004).

111. Jan, S. *et al.* GATE V6: A major enhancement of the GATE simulation platform enabling modelling of CT and radiotherapy. *Physics in Medicine and Biology* **56,** 881–901 (2011).

112. Bataille, F., Comtat, C., Jan, S. & Trebossen, R. *Monte Carlo Simulation for the ECAT HRRT using GATE* in *IEEE Symposium Conference Record Nuclear Science 2004.* **4** (IEEE, 2004), 2570–2574.

113. Lamare, F., Turzo, A., Bizais, Y., Rest, C. C. L. & Visvikis, D. Validation of a Monte Carlo simulation of the Philips Allegro/GEMINI PET systems using GATE. *Physics in Medicine and Biology* **51,** 943–962 (2006).

114. Schmidtlein, C. R. *et al.* Validation of GATE Monte Carlo simulations of the GE Advance Discovery/LS PET scanners. *Medical Physics* **33,** 198–208 (2006).

115. Rehfeld, N. S., Stute, S., Apostolakis, J., Soret, M. & Buvat, I. Introducing improved voxel navigation and fictitious interaction tracking in GATE for enhanced efficiency. *Physics in Medicine and Biology* **54,** 2163–2178 (2009).

116. Camarasu-Pop, S., Glatard, T., Mościcki, J. T., Benoit-Cattin, H. & Sarrut, D. Dynamic Partitioning of GATE Monte-Carlo Simulations on EGEE. *Journal of Grid Computing* **8,** 241–259 (2010).

117. Camarasu-Pop, S. *et al.* Monte Carlo simulation on heterogeneous distributed systems: A computing framework with parallel merging and checkpointing strategies. *Future Generation Computer Systems* **29,** 728–738 (2013).

118. Glatard, T. *et al.* A virtual imaging platform for multi-modality medical image simulation. *IEEE Transactions on Medical Imaging* **32,** 110–118 (2013).

119. Holford, N., Ma, S. C. & Ploeger, B. A. Clinical Trial Simulation: A Review. *Clinical Pharmacology & Therapeutics* **88,** 166–182 (2010).

120. Samei, E. *et al. Virtual imaging trials: an emerging experimental paradigm in imaging research and practice* in *Medical Imaging 2020: Physics of Medical Imaging* (eds Bosmans, H. & Chen, G.-H.) (SPIE, 2020), 65. ISBN: 9781510633919.

121. Badano, A. In silico imaging clinical trials: cheaper, faster, better, safer, and more scalable. *Trials* **22,** 1–7 (2021).

122. Badano, A. *et al.* Evaluation of Digital Breast Tomosynthesis as Replacement of Full-Field Digital Mammography Using an In Silico Imaging Trial. *JAMA Network Open* **1,** e185474 (2018).

123. Sharma, D. *et al.* Technical Note: In silico imaging tools from the VICTRE clinical trial. *Medical Physics* **46,** 3924–3928 (2019).

124. Elangovan, P. *et al.* Design and validation of realistic breast models for use in multiple alternative forced choice virtual clinical trials. *Physics in Medicine and Biology* **62,** 2778–2794 (2017).

125. Mackenzie, A., Kaur, S., Elangovan, P., Dance, D. R. & Young, K. C. *Comparison of synthetic 2D images with planar and tomosynthesis imaging of the breast using a virtual clinical trial* in *Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment* (eds Nishikawa, R. M. & Samuelson, F. W.) (SPIE, 2018), 16.

126. Abadi, E. *et al.* Virtual clinical trial for quantifying the effects of beam collimation and pitch on image quality in computed tomography. *Journal of Medical Imaging* **7,** 1 (04 2020).

127. Abadi, E. *et al. Optimization of energy thresholds in photon-counting CT via a virtual clinical trial* in (eds Bosmans, H. & Chen, G.-H.) (SPIE, 2020), 70. ISBN: 9781510633919.

128. Elston, B., Comtat, C., Harrison, R. & Kinahan, P. in *Monte Carlo Calculations in Nuclear Medicine: Applications in Diagnostic Imaging* (eds Ljungberg, M., Strand, S.-E. & King, M.) 201–220 (CRC Press, Boca Raton, FL, FL, 2012).

129. Comtat, C. *et al. Simulating whole-body PET scanning with rapid analytical methods* in *1999 IEEE Nuclear Science Symposium. Conference Record. 1999 Nuclear Science Symposium and Medical Imaging Conference (Cat. No.99CH37019)* **3** (IEEE, 1999), 1260–1264.

130. Thielemans, K. *et al.* STIR: Software for tomographic image reconstruction release 2. *Physics in Medicine and Biology* **57,** 867–883 (2012).

131. Abadi, E. *et al.* Virtual clinical trials in medical imaging: a review. *Journal of Medical Imaging* **7,** 1 (04 2020).

132. Frangi, A. F., Tsaftaris, S. A. & Prince, J. L. Simulation and Synthesis in Medical Imaging. *IEEE Transactions on Medical Imaging* **37,** 673–679 (2018).

133. Badano, A. In silico imaging: Definition, possibilities and challenges. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **648,** S276–S280 (2011).

134. Zubal, I. G. *et al.* Computerized three-dimensional segmented human anatomy. *Medical Physics* **21,** 299–302 (1994).

135. Silva-Rodríguez, J., Domínguez-Prado, I., Pardo-Montero, J., Ruibal, Á. & Aguiar, P. Impact of muscular uptake and statistical noise on tumor quantification based on simulated FDG-PET studies. *Radiation Physics and Chemistry* **131,** 28–34 (2017).

136. Tomei, S. *et al.* OncoPET_DB: A Freely Distributed Database of Realistic Simulated Whole Body 18F-FDG PET Images for Oncology. *IEEE Transactions on Nuclear Science* **57,** 246–255 (2010).

137. Le Maitre, A. *et al.* Incorporating patient-specific variability in the simulation of realistic whole-body 18F-FDG distributions for oncology applications. *Proceedings of the IEEE* **97,** 2026–2038 (2009).

138. Pfaehler, E., De Jong, J. R., Dierckx, R. A., van Velden, F. H. & Boellaard, R. SMART (SiMulAtion and ReconsTruction) PET: an efficient PET simulation-reconstruction tool. *EJNMMI Physics* **5** (2018).

139. Berthon, B. *et al.* PETSTEP: Generation of synthetic PET lesions for fast evaluation of segmentation methods. *Physica Medica* **31,** 969–980 (2015).

140. Tzanoukos, G. *et al.* Modeling of solitary pulmonary nodules in PET/CT images using Monte Carlo methods. *13th IEEE International Conference on BioInformatics and BioEngineering, IEEE BIBE 2013* (2013).

141. Sarrut, D. *et al.* Advanced Monte Carlo simulations of emission tomography imaging systems with GATE. *Physics in Medicine and Biology* **66** (2021).

142. Kemp, B. J., Kim, C., Williams, J. J., Ganin, A. & Lowe, V. J. NEMA NU 2-2001 performance measurements of an LYSO-based PET/CT system in 2D and 3D acquisition modes. *Journal of nuclear medicine* **47,** 1960–7 (2006).

143. Pépin, A., Stute, S., Jan, S. & Comtat, C. Normalization of Monte Carlo PET data using GATE. *IEEE Nuclear Science Symposium Conference Record,* 4196–4200 (2011).

144. Schneider, W., Bortfeld, T. & Schlegel, W. Correlation between CT numbers and tissue parameters needed for Monte Carlo simulations of clinical dose distributions. *Physics in Medicine and Biology* **45,** 459–478 (2000).

145. Stute, S., Tylski, P., Grotus, N. & Buvat, I. *LuCaS: Efficient Monte Carlo simulations of highly realistic PET tumor images* in *2008 IEEE Nuclear Science Symposium Conference Record* (IEEE, 2008), 4010–4012.

146. NEMA. *NEMA NU 2-2012 Performance Measurements of Positron Emission Tomographs (PET)* tech. rep. (National Electrical Manufacturers Association, 2012).

147. Gong, K., Cherry, S. R. & Qi, J. On the assessment of spatial resolution of PET systems with iterative image reconstruction. *Physics in Medicine and Biology* **61,** N193–N202 (2016).

148. Bailey, D. L., Jones, T. & Spinks, T. J. A method for measuring the absolute sensitivity of positron emission tomographic scanners. *European Journal of Nuclear Medicine* **18,** 374–379 (1991).

149. Daube-Witherspoon, M. E. *et al.* PET performance measurements using the NEMA NU 2-2001 standard. *Journal of Nuclear Medicine* **43,** 1398–1409 (2002).

150. Healthcare, G. *Discovery PET/CT 600, 710 Series Optima PET/CT 560 Series NEMA Test Procedures* tech. rep. (GE Healthcare, Waukesha, 2012), 1–58.

151. Scheuermann, J. S., Saffer, J. R., Karp, J. S., Levering, A. M. & Siegel, B. a. Qualification of PET scanners for use in multicenter cancer clinical trials: the American College of Radiology Imaging Network experience. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* **50,** 1187–1193 (2009).

152. Committee, F.-P. T. *FDG-PET/CT as an Imaging Biomarker Measuring Response to Cancer Therapy, Quantitative Imaging Biomarkers Alliance, Version 1.13, Technically Confirmed Version.* tech. rep. (QIBA, 2016). RSNA.ORG/QIBA.

153. Swensen, S. J. *et al.* Lung Nodule Enhancement at CT: Multicenter Study. *Radiology* **214,** 73–80 (2000).

154. Srinivas, S. M. *et al.* A recovery coefficient method for partial volume correction of PET images. *Annals of Nuclear Medicine* **23,** 341–348 (2009).

155. Hofheinz, F., Dittrich, S., Pötzsch, C. & Hoff, J. V. D. Effects of cold sphere walls in PET phantom measurements on the volume reproducing threshold. *Physics in Medicine and Biology* **55,** 1099–1113 (2010).

156. Nehmeh, S. A. & Erdi, Y. E. Respiratory Motion in Positron Emission Tomography/Computed Tomography: A Review. *Seminars in Nuclear Medicine* **38,** 167–176 (2008).

157. Vauclin, S. *et al.* Monte-Carlo simulations of clinically realistic respiratory gated 18F-FDG PET: Application to lesion detectability and volume measurements. *Computer Methods and Programs in Biomedicine* **118,** 84–93 (2015).

158. Merlin, T. *et al.* CASToR: a generic data organization and processing code framework for multi-modal and multi-dimensional tomographic reconstruction. *Physics in Medicine & Biology* **63,** 185005 (2018).

159. Khan, T. *et al.* Diagnosis and management of peripheral lung nodule. *Annals of Translational Medicine* **7,** 348–348 (2019).

160. Cecil, R. L., Goldman, L. & Schafer, A. I. *Goldman's Cecil Medicine* 24th ed. (eds Goldman, L. & Schafer, A. I.) (Elsevier, 2012).

161. Prieto, E. *et al.* Impact of time-of-flight and point-spread-function in SUV quantification for oncological PET. *Clinical Nuclear Medicine* **38,** 103–109 (2013).

162. Rogasch, J. M. *et al.* The association of tumor-to-background ratios and SUVmax deviations related to point spread function and time-of-flight F18-FDG-PET/CT reconstruction in colorectal liver metastases. *EJNMMI Research* **5,** 31 (2015).

163. Kuhnert, G. *et al.* Impact of PET/CT image reconstruction methods and liver uptake normalization strategies on quantitative image analysis. *European Journal of Nuclear Medicine and Molecular Imaging* **43,** 249–258 (2016).

164. Van den Belt, T. *et al.* Effect of reconstruction settings and therapy on 18F-FDG PET/CT blood pool and liver SUV in patients with diffuse large B-cell lymphoma. *Journal of Nuclear Medicine* **58,** 608–608 (2017).

165. Murphy, D. *et al.* The effect of a novel Bayesian penalised likelihood PET reconstruction algorithm on the assessment of malignancy risk in solitary pulmonary nodules according to the British Thoracic Society guidelines. *European Journal of Radiology* **117,** 149–155 (2019).

166. Poleszczuk, J., Macklin, P. & Enderling, H. Agent-Based Modeling of Cancer Stem Cell Driven Solid Tumor Growth. *Methods Mol Biol.,* 335–346 (2016).

167. Enderling, H., Park, D., Hlatky, L. & Hahnfeldt, P. The importance of spatial distribution of stemness and proliferation state in determining tumor radioresponse. *Mathematical Modelling of Natural Phenomena* **4,** 117–133 (2009).

168. Del Monte, U. Does the cell number 109 still really fit one gram of tumor tissue? *Cell Cycle* **8,** 505–506 (2009).

169. Sim, Y. T. & Poon, F. W. Imaging of solitary pulmonary nodule-a clinical review. *Quantitative imaging in medicine and surgery* **3,** 316–26 (2013).

170. Choromańska, A. & Macura, K. J. Evaluation of solitary pulmonary nodule detected during computed tomography examination. *Polish Journal of Radiology* **77,** 22–34 (2012).

171. Swensen, S. J., Silverstein, M. D., Ilstrup, D. M., Schleck, C. D. & Edell, E. S. The Probability of Malignancy in Solitary Pulmonary Nodules. Application to small radiologically indeterminate nodules. *Archives of Internal Medicine* **157,** 849 (1997).

172. Landis, J. R. & Koch, G. G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **33,** 159 (1977).

173. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44,** 837–45 (Sept. 1988).

174. Newcombe, R. G. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* **17,** 2635–2650 (1998).

175. Mercaldo, N. D., Lau, K. F. & Zhou, X. H. Confidence intervals for predictive values with an emphasis to case–control studies. *Statistics in Medicine* **26,** 2170–2183 (2007).

176. Boellaard, R. Standards for PET Image Acquisition and Quantitative Data Analysis. *Journal of Nuclear Medicine* **50,** 11S–20S (2009).

177. Chin, B. B., Green, E. D., Turkington, T. G., Hawk, T. C. & Coleman, R. E. Increasing uptake time in FDG-PET: Standardized uptake values in normal tissues at 1 versus 3 h. *Molecular Imaging and Biology* **11,** 118–122 (2009).

178. Barrington, S. F. & Kluge, R. FDG PET for therapy monitoring in Hodgkin and non-Hodgkin lymphomas. *European Journal of Nuclear Medicine and Molecular Imaging* **44,** 97–110. ISSN: 1619-7070 (Aug. 2017).

179. Ordidge, K. L. *et al.* Interobserver agreement of the visual Herder scale for the assessment of solitary pulmonary nodules on 18F Fluorodeoxyglucose PET/computed tomography. *Nuclear Medicine Communications* **41,** 235–240. ISSN: 0143-3636 (Mar. 2020).

180. Fatania, K. *et al.* Multi-observer concordance and accuracy of the British Thoracic Society scale and other visual assessment qualitative criteria for solid pulmonary nodule assessment using FDG PET-CT. *Clinical Radiology* **75,** 878.e21–878.e28. ISSN: 00099260 (Nov. 2020).

181. Chen, B. *et al.* Validation of a Projection-domain Insertion of Liver Lesions into CT Images. *Academic Radiology* **23,** 1221–1229 (2016).

182. Papadimitroulas, P. *et al.* Investigation of realistic PET simulations incorporating tumor patient's specificity using anthropomorphic models: Creation of an oncology database. *Medical Physics* **40,** 1–13 (2013).

183. Solomon, J. & Samei, E. A generic framework to simulate realistic lung, liver and renal pathologies in CT imaging. *Physics in Medicine and Biology* **59,** 6637–6657 (2014).

184. Rashidnasab, A. *et al.* Simulation and assessment of realistic breast lesions using fractal growth models. *Physics in Medicine and Biology* **58,** 5613–5627 (2013).

185. Cornelis, F. *et al.* In vivo mathematical modeling of tumor growth from imaging data: soon to come in the future? *Diagnostic and interventional imaging* **94,** 593–600 (2013).

186. Alyafi, B. *et al. Quality analysis of DCGAN-generated mammography lesions* in *15th International Workshop on Breast Imaging (IWBI2020)* (eds Van Ongeval, C., Marshall, N. & Bosmans, H.) (SPIE, May 2020), 19. ISBN: 9781510638310. eprint: 1911.12850.

187. Frid-Adar, M. *et al.* GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*. ISSN: 18728286. eprint: 1803.01229 (2018).

188. Wang, Z. *et al.* Generation of synthetic ground glass nodules using generative adversarial networks (GANs). *European Radiology Experimental* **6,** 59. ISSN: 2509-9280 (Nov. 2022).

189. Bettinardi, V., Castiglioni, I., De Bernardi, E. & Gilardi, M. C. PET quantification: Strategies for partial volume correction. *Clinical and Translational Imaging* **2,** 199–218 (2014).

190. Rogasch, J. M. *et al.* Influences on PET Quantification and Interpretation. *Diagnostics* **12,** 1–25 (2022).

191. Boellaard, R. Need for standardization of 18F-FDG PET/CT for treatment response assessments. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* **52 Suppl 2,** 93S–100S (2011).

192. Makris, N. E., Huisman, M. C., Kinahan, P. E., Lammertsma, A. A. & Boellaard, R. Evaluation of strategies towards harmonization of FDG PET/CT studies in multicentre trials: Comparison of scanner validation phantoms and data analysis procedures. *European Journal of Nuclear Medicine and Molecular Imaging* **40,** 1507–1515 (2013).

193. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3,** 32–35. ISSN: 0008-543X (1950).

194. Weir-McCall, J. R. *et al.* Impact of solitary pulmonary nodule size on qualitative and quantitative assessment using 18F-fluorodeoxyglucose PET/CT: the SPUTNIK trial. *European Journal of Nuclear Medicine and Molecular Imaging* (Nov. 2020).

195. Caldwell, C. B. *et al.* Observer variation in contouring gross tumor volume in patients with poorly defined non-small-cell lung tumors on CT: the impact of 18 FDG-hybrid PET fusion. *International Journal of Radiation Oncology*Biology*Physics* **51,** 923–931 (2001).

196. Riegel, A. C. *et al.* Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion. *International Journal of Radiation Oncology Biology Physics* **65,** 726–732 (2006).

197. Hatt, M. *et al.* Impact of Tumor Size and Tracer Uptake Heterogeneity in 18 F-FDG PET and CT Non–Small Cell Lung Cancer Tumor Delineation. *Journal of Nuclear Medicine* **52,** 1690–1697 (2011).

198. Öllers, M. *et al.* The integration of PET-CT scans from different hospitals into radiotherapy treatment planning. *Radiotherapy and Oncology* **87,** 142–146 (2008).

199. Schaefer, A. *et al.* A contrast-oriented algorithm for FDG-PET-based delineation of tumour volumes for the radiotherapy of lung cancer: Derivation from phantom measurements and validation in patient data. *European Journal of Nuclear Medicine and Molecular Imaging* **35,** 1989–1999 (2008).

200. Wanet, M. *et al.* Gradient-based delineation of the primary GTV on FDG-PET in non-small cell lung cancer: A comparison with threshold-based approaches, CT and surgical specimens. *Radiotherapy and Oncology* **98,** 117–125 (2011).

201. Zwezerijnen, G. J. *et al.* Reproducibility of [18F]FDG PET/CT liver SUV as reference or normalisation factor. *European Journal of Nuclear Medicine and Molecular Imaging* **50,** 486–493 (2023).

202. Benz, M. R. *et al.* Treatment Monitoring by 18 F-FDG PET/CT in Patients with Sarcomas: Interobserver Variability of Quantitative Parameters in Treatment-Induced Changes in Histopathologically Responding and Nonresponding Tumors. *Journal of Nuclear Medicine* **49,** 1038–1046 (2008).

203. Nahmias, C. & Wahl, L. M. Reproducibility of standardized uptake value measurements determined by 18F-FDG PET in malignant tumors. *Journal of Nuclear Medicine* **49,** 1804–1808 (2008).

204. Krak, N. C. *et al.* Effects of ROI definition and reconstruction method on quantitative outcome and applicability in a response monitoring trial. *European Journal of Nuclear Medicine and Molecular Imaging* **32,** 294–301 (2005).

205. Lodge, M. A. Repeatability of SUV in oncologic 18F-FDG PET. *Journal of Nuclear Medicine* **58,** 523–532 (2017).

206. Meignan, M., Gallamini, A., Meignan, M., Gallamini, A. & Haioun, C. Report on the first international workshop on interim-PET scan in lymphoma. *Leukemia & lymphoma* **50,** 1257–1260 (2009).

207. Wyrzykowski, M., Siminiak, N., Kaźmierczak, M., Ruchała, M. & Czepczyński, R. Impact of the Q.Clear reconstruction algorithm on the interpretation of PET/CT images in patients with lymphoma. *EJNMMI Research* **10,** 99. ISSN: 2191-219X (Dec. 2020).

208. Enilorac, B. *et al.* Does PET reconstruction method affect Deauville score in lymphoma patients? *Journal of Nuclear Medicine* **59,** 1049–1055 (2018).

209. Tomei, S. *et al. Development of a database of realistic simulated whole body 18F-FDG PET images for lymphoma* in *2008 IEEE Nuclear Science Symposium Conference Record* (IEEE, 2008), 4958–4963.

210. Van der Vos, C. S. *et al.* Quantification, improvement, and harmonization of small lesion detection with state-of-the-art PET. *European Journal of Nuclear Medicine and Molecular Imaging* **44,** 4–16. ISSN: 16197089 (2017).

211. He, X. & Park, S. Model Observers in Medical Imaging Research. *Theranostics* **3,** 774–786. ISSN: 1838-7640 (2013).

212. Han, M. & Baek, J. A convolutional neural network-based anthropomorphic model observer for signal-known-statistically and background-known-statistically detection tasks. *Physics in Medicine & Biology* **65,** 225025. ISSN: 0031-9155 (Nov. 2020).

213. Büther, F. *et al.* Clinical Evaluation of a Data-Driven Respiratory Gating Algorithm for Whole-Body PET with Continuous Bed Motion. *Journal of Nuclear Medicine* **61,** 1520–1527. ISSN: 0161-5505 (Oct. 2020).

214. Walker, M. D., Morgan, A. J., Bradley, K. M. & McGowan, D. R. Data-driven respiratory gating outperforms device-based gating for clinical 18F-FDG PET/CT. *Journal of Nuclear Medicine* **61,** 1678–1683. ISSN: 2159662X (2020).

215. Frood, R., McDermott, G. & Scarsbrook, A. Respiratory-gated PET/CT for pulmonary lesion characterisation—promises and problems. *The British Journal of Radiology,* 20170640. ISSN: 0007-1285 (Feb. 2018).

216. Reader, A. J. & Schramm, G. Artificial Intelligence for PET Image Reconstruction. *Journal of Nuclear Medicine* **62,** 1330–1333. ISSN: 0161-5505 (Oct. 2021).

217. Schreuder, A., Scholten, E. T., van Ginneken, B. & Jacobs, C. Artificial intelligence for detection and characterization of pulmonary nodules in lung cancer CT screening: ready for practice? *Translational Lung Cancer Research* **10,** 2378–2388. ISSN: 22186751 (May 2021).

218. Surti, S. & Karp, J. S. Update on latest advances in time-of-flight PET. *Physica Medica* **80,** 251–258. ISSN: 1120-1797 (2020).

Appendix A

Macro used to Define the PET Scanner Geometry in GATE

```
## Define the WORLD geometry:
# Needs to be slightly larger than inserted scanner and phantom objects

   /gate/world/geometry/setXLength                          90. cm
   /gate/world/geometry/setYLength                          90. cm
   /gate/world/geometry/setZLength                          52. cm
   /gate/world/setMaterial                                  Air

## Define the scanner DETECTOR GEOMETRY:

# PET system => cylindricalPET [LEVEL 0]
   /gate/world/daughters/name                               cylindricalPET
   /gate/world/daughters/insert                             cylinder
   /gate/cylindricalPET/placement/setTranslation            0.0 0.0 0.0 cm
   /gate/cylindricalPET/geometry/setRmax                    445. mm
   /gate/cylindricalPET/geometry/setRmin                    400. mm
```

```
   /gate/cylindricalPET/geometry/setHeight              157.1 mm
   /gate/cylindricalPET/setMaterial                     Air

# GE PET module => rsector [LEVEL 1]
   /gate/cylindricalPET/daughters/name                  rsector
   /gate/cylindricalPET/daughters/insert                box
   /gate/rsector/placement/setTranslation               417.6 0.0 0.0 mm
   /gate/rsector/geometry/setXLength                    25.0 mm
   /gate/rsector/geometry/setYLength                    78.1 mm
   /gate/rsector/geometry/setZLength                    156.7 mm
   /gate/rsector/setMaterial                            Air

# GE PET unit => module [LEVEL 2]
   /gate/rsector/daughters/name                         module
   /gate/rsector/daughters/insert                       box
   /gate/module/placement/setTranslation                0.0 0.0 0.0 mm
   /gate/module/geometry/setXLength                     25.0 mm
   /gate/module/geometry/setYLength                     78.1 mm
   /gate/module/geometry/setZLength                     156.7 mm
   /gate/module/setMaterial                             Air

# GE PET block => submodule [LEVEL 3]
   /gate/module/daughters/name                          submodule
   /gate/module/daughters/insert                        box
   /gate/submodule/placement/setTranslation             0.0 0.0 0.0 mm
   /gate/submodule/geometry/setXLength                  25.0 mm
   /gate/submodule/geometry/setYLength                  38.35 mm
   /gate/submodule/geometry/setZLength                  38.35 mm
   /gate/submodule/setMaterial                          Air
```

```
# PET crystal => crystal [LEVEL 4]
   /gate/submodule/daughters/name                          crystal
   /gate/submodule/daughters/insert                        box
   /gate/crystal/placement/setTranslation                  0.0 0.0 0.0 mm
   /gate/crystal/geometry/setXLength                       25. mm
   /gate/crystal/geometry/setYLength                       4.26 mm
   /gate/crystal/geometry/setZLength                       6.39 mm
   /gate/crystal/setMaterial                               LYSO

# Repeat the crystal array (9 x 6) => crystal [LEVEL 4] # no gaps
   /gate/crystal/repeaters/insert                          cubicArray
   /gate/crystal/cubicArray/setRepeatNumberX               1
   /gate/crystal/cubicArray/setRepeatNumberY               9
   /gate/crystal/cubicArray/setRepeatNumberZ               6
   /gate/crystal/cubicArray/setRepeatVector                0.0 4.26 6.39 mm

# Repeat the block array (2 x 4) => submodule [LEVEL 3]
   /gate/submodule/repeaters/insert                        cubicArray
   /gate/submodule/cubicArray/setRepeatNumberX             1
   /gate/submodule/cubicArray/setRepeatNumberY             2
   /gate/submodule/cubicArray/setRepeatNumberZ             4
   /gate/submodule/cubicArray/setRepeatVector              0.0 39.75 39.45 mm

# Repeat the unit array (1 x 1) => module [LEVEL 2]
   /gate/module/repeaters/insert                           cubicArray
   /gate/module/cubicArray/setRepeatNumberX                1
   /gate/module/cubicArray/setRepeatNumberY                1
   /gate/module/cubicArray/setRepeatNumberZ                1
```

```
   /gate/module/cubicArray/setRepeatVector              0.0 0.0 0.0 mm

# Repeat the module array => rsector [LEVEL 1]
   /gate/rsector/repeaters/insert                       ring
   /gate/rsector/ring/setRepeatNumber                   32

# Attach the modules (rsector), units (module), blocks (submodule) and crystals
    to the PET system
   /gate/systems/cylindricalPET/rsector/attach          rsector
   /gate/systems/cylindricalPET/module/attach           module
   /gate/systems/cylindricalPET/submodule/attach        submodule
   /gate/systems/cylindricalPET/crystal/attach          crystal

# Attach the crystal as a sensitive detector
   /gate/crystal/attachCrystalSD
```