Assessing the robustness of radiomics features in oncology PET

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Science and Engineering

2022

George R. Needham
Department of Physics & Astronomy

Contents

Word Count: 34,741

List of Figures

List of Tables

Table of Terms & Abbreviations

Name	Definition
ARSAC	*Administration of Radioactive Substances Advisory Committee*; the body that regulates the industry standards for nuclear medicine
CASToR	*Customisable and Advanced Software for Tomographic Reconstruction*; an open source reconstruction software for PET data
CV/CoV	*coefficient of variation*; the standard deviation divided by the mean of a dataset
DICOM	*Digital Imaging & Communications in Medicine*; a widely used file format used to store medical images
FBN/FBS	*fixed bin number/size*; referring to the discretisation method chosen for texture matrix processing
FBP	*filtered back-projection*; an analytical tomographic reconstruction technique
FDG	*fluoro-deoxyglucose*; a commonly used PET radiochemical using the positron-emitting isotope ^{18}F
FoV	*field of view*; in PET, this generally refers to the axial field of view, the depth of the PET detector cylinder
GATE	*GEANT4 Application for Tomographic Emission*; software used to simulate PET & SPECT scans using the GEANT4 framework
GEANT4	*GEometry ANd Tracking (4)*; software used to simulate the interactions of particles and matter

continues on next page

GLCM	*gray level co-occurrence matrix*; a textural analysis matrix used to show local heterogeneity
GLDM	*gray level dependence matrix*; a textural analysis matrix used to show local heterogeneity
GLRLM	*gray level run length matrix*; a textural analysis matrix used to show local heterogeneity
GLSZM	*gray level size-zone matrix*; a textural analysis matrix used to show 'regional' heterogeneity
LOR	*line of response*; the virtual line (or tube) connecting two coincidentally-triggered detector elements in a PET scanner
LSO	*lutetium orthosilicate*; a material commonly used for non-organic scintillation crystals in PET
MLEM	*maximum likelihood expectation maximisation*; an iterative reconstruction technique
MRI	*Magnetic Resonance Imaging*; a non-ionising medical imaging technique utilising the magnetic spin response of the nucleus
NECR	*noise-equivalent count rate*; a scanner performance metric derived from raw count statistics
NEMA	*National Electrical Manufacturers Association*; the group who developed an image quality standard for PET - their name is often used throughout this report to refer to their image quality phantom
NGTDM	*neighbourhood gray tone difference matrix*; a textural analysis matrix used to show local heterogeneity
OSEM	*ordered subset expectation maximisation*; an implementation of the MLEM algorithm over many subsets of the initial data
PET	*Positron Emission Tomography*; a method of medical imaging using a radioactive tracer

continues on next page

PMCC	*Pearson product-moment correlation coefficient*; a statistic used to measure linear correlation between two variables, equal to ratio of the covariance between the two and the product of each variable's standard deviation
PMT	*photomultiplier tube*; a device used to detect small numbers of visible light photons
PVE/PVC	*partial volume effect/correction*; the effect of, and corrective methods to adjust for, spatial resolution and image sampling degradation of data
qPET	*'PET quotient'*; a metric derived to describe the SUV of an ROI with respect to some background value
ROI	*region of interest*; the area, or number of voxels, selected for analysis
SPECT	*Single Photon Emission Computed Tomography*; another form of medical imaging, using tracers that emit single gamma photons as opposed to positrons
SUV	*Standardised Uptake Value*; a metric used to describe the uptake of radiotracer in a voxel or region of interest
TGV	*total glycolytic volume*; a measure used to quantify the amount of glucose metabolism in an ROI in an FDG scan
ToF	*time of flight*; referring to the enhanced timing capability of modern PET detectors, where coincidences can be localised to a smaller region in an LOR

Abstract

Radiomics is the branch of image analysis concerned with the extraction of statistics (features) concerning not just uptake and shape, but the heterogeneity of the ROI, using spatially-aware textural features to describe the distribution of voxel values. These texture features have shown potential for providing diagnostic and prognostic information. This work examines the robustness of these textural features in order to advise on their future usage in PET studies.

The signal-noise-ratio (SNR) of PET data can be approximated by the noise-equivalent count rate (NECR), a scanner- and geometry-dependent performance metric. An investigation was carried out to determine whether textural image features were correlated to the NECR, achieved by acquiring data from phantoms filled with a high activity of ^{18}F on a Siemens Biograph mCT TrueV over a 12 hour period. Four phantoms were utilised; a cylinder, the NEMA IQ (Image Quality) phantom and two variants using custom-printed tumour-like inserts for the NEMA IQ phantom. The data was recorded in successive 5 and 25 minute frames and images were reconstructed using clinically-appropriate parameters. Radiomics features were extracted using an IBSI (Image Biomarker Standardisation Initiative) compliant method. Strong correlations ($|\text{PMCC}| > 0.9$) were found with NECR for 32 out of 75 textural features for large-volume, long time frame images, enabling their characterisation. Multiplication factors were calculated enabling correction of texture features from the value obtained at clinically-expected activity levels to their expected value at peak NECR. Such correlations diminish for short time frame images and when considering smaller ROIs such as the NEMA spheres and phantom inserts. This thesis discusses these results, suggesting methods for calculating a 'tumour-specific' noise equivalent count rate to address the diminished textural feature correlations. In addition, work undertaken towards building a Monte Carlo simulation framework to improve this study is discussed.

Declaration of Authorship

I hereby confirm that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see `http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420`), in any relevant Thesis restriction declarations deposited in the University Library, the University Library's regulations (see `http://www.library.manchester.ac.uk/about/regulations/`) and in the University's policy on Presentation of Theses.

List of Presented Works

British Nuclear Medicine Society (BNMS) Meeting 2020 (held 2021), Young Investigator Prize Finalist (Talk)
"Validation of the Siemens e7-tools off-line PET image reconstruction software", **G. Needham**, C. Oldfield, P. Julyan, D.M. Cullen, J. Anton-Rodriguez, I. Armstrong, B. Sanghera

European Association of Nuclear Medicine (EANM) 2021 Congress, Top Rated Data Analysis Session (Talk)
"A novel methodology for assessing reproducibility of heterogeneity metrics in PET radiomics using noise-equivalent count rate, Monte Carlo simulation and 3D-printed patient-specific tumour phantoms", **G Needham**, P. Julyan, D.M. Cullen, J. Tipping, D. Hamilton, S. Pells, A. Fish

Institute of Physics (IOP) Joint APP, HEPP and NP Conference 2021 (Poster)
"Validating heterogeneity-based radiomics metrics in PET using noise-equivalent count rate, Monte Carlo simulation & 3D-printed patient-specific tumour phantoms", **G. Needham**, P. Julyan, D. Cullen, J. Tipping, D. Hamilton, S. Pells, A. Fish, E. Page, J. Anton-Rodriguez

PSMR-TBP 2022 9th Conference on PET/MR and SPECT/MR & Total Body PET 2022 (Poster)
"Assessing the robustness of radiomics feature measurements using the noise equivalent count rate, and the future role of Total Body PET", **G. Needham**, P. Julyan, D. Cullen, S. Pells, A. Fish, J. Tipping, G. Smith, D. Hamilton

Acknowledgements

Firstly I would like to thank my supervisors for helping to orchestrate this project: Peter Julyan at The Christie and Dave Cullen formerly of The University of Manchester. Pete has been a constant help in managing this project through some incredibly difficult times, and his good nature and positivity have been vital in keeping my spirits up over the past four years. I'd also like to thank Gavin Smith for all of his help and oversight towards the end of the project's run.

In acknowledgement of every word of help and advice, I would like to thank the following members of Nuclear Medicine at The Christie NHS Foundation Trust: Jill Tipping, Dave Hamilton, Jose Anton-Rodriguez, Helen Miller, Chris Oldfield, Emma Page, Nick Calvert & Heather Williams. Likewise at the University of Manchester Nuclear Physics Group: Sophia Pells, Emlyn Price, Alex Fish, Liam Barber, Ben Pietras, Mike Mallaburn, and everyone else who passed through. I'd also like to thank Rob Lyon, Anna Scaife, Patrick Parkinson, Paul Campbell and Draga Pihler-Puzović at UoM for inspiring me to take up research, as I wouldn't be here without them.

Throughout the undulating course of the last four years, I have been blessed with a network of amazing friends and family, and without their patience and guidance I would not have made it this far. A special mention to Aaron, Andy, Anna, Annie, Ben, Benoit, Callum, Calvin, Chris, Ed, Isaac, Jacob, Joe H.-E., Joe S., Josh, Laurie, Lynsey, Marie, Rhydian, Sam, Savannah, Shantam, Sohail, and Tom; to Tori, Hattie, Mum and Dad, an even more special mention; and to Alice, for putting up with me through all of the stresses and strains - I don't know how you do it.

"...there is nothing in this world which says nothing. Often - it is true - the message does not reach our soul, either because it has no meaning in and for itself, or - as is more likely - because it has not been conveyed to the right place. Every serious work rings inwardly, like the calm and dignified words: 'Here I am!'"

Wassily Kandinsky
translated from *Über das Geistige in der Kunst*, 1911

Chapter 1

Introduction

1.1 A Brief History of PET

Radioisotope imaging - the imaging modality covering positron emission tomography (PET) and planar gamma camera tomography - is a *functional* imaging process, generally providing information that cannot be gained by other techniques such as X-ray computed tomography (CT) or nuclear magnetic resonance imaging (MRI). By bonding radioisotopes to substances such as antibodies, treatment drugs or glucose, information about the activity and performance of a range of bodily functions can be determined.

The use of positrons in a medical context was first investigated in 1951 to localise brain tumours [1], and the first detector specified for medical imaging was built at Massachusetts General Hospital by Brownell & Sweet [2]. It was comprised of two opposing NaI(Tl) scintillation detectors. The resultant images were significantly different to the scans that are produced today, but the breakthrough this represented cannot be understated. It took over a decade for tomographs to be produced, requiring advances in detector technology and the groundbreaking use of algorithms to reconstruct the images to reach this stage. The first human tomograph was taken in 1974 by Phelps & Hoffman at Washington University on their PET III scanner, comprising of a ring of 48 NaI(Tl) detectors [2].

The advancement of inorganic scintillation materials, especially BGO (bismuth germanate, $Bi_4Ge_3O_{12}$) in the late 1970s and LSO (lutetium orthosilicate),

discussed later, enabled increased timing precision and material workability, which greatly helped in designing more capable PET detectors [3]. From 48 detector elements in the early 1970s to 32,448 in the scanner used in this work, the increase in precision has been significant. Likewise, faster timing in detector technology has allowed time-of-flight resolution in modern scanners down to the picosecond-level.

The development of PET-CT in the 1990s elevated PET beyond its widely-regarded status as a novel research tool into a truly useful clinical device. Prior to the introduction of the dual-modality system, the task of aligning the PET data with a CT counterpart in order to perform corrections was computationally difficult, slow and impractical [4]. The simultaneous improvement in computing power over the years has enabled a huge increase in data storage and increased speed of image reconstruction [5]. Initially, each ring of detector elements in a scanner would be separated by lead or tungsten septa[1]; the increase of computing power and storage enabled the removal of these septa, going from '2D' to '3D' PET. Faster computation has enabled better data correction methods, and more developed image reconstruction algorithms have enabled clearer and more quantitatively accurate images to be produced.

Positron imaging is not without its drawbacks. By its very nature, there is always a risk to the patient caused by the administration of radioactive material. Patient safety and dose considerations have always been at the forefront of advancing research. However, the benefits of PET imaging have always been felt to justify the means, and the use of clinical PET-CT is not slowing. NHS data showed a 15 % increase in clinical PET-CT scans given across England in 2018-19 from the previous year, and the number of scans taken increased to over 200,000 in 2021-22, despite the effect of the COVID-19 pandemic [6]. Should the rate of increase of scans return to pre-pandemic levels, research into making PET safer, more effective and more efficient will only be more crucial in years to come.

[1]These are thin blocks of material that absorb photons which approach at undesired angles, confining the accepted range to within a defined plane.

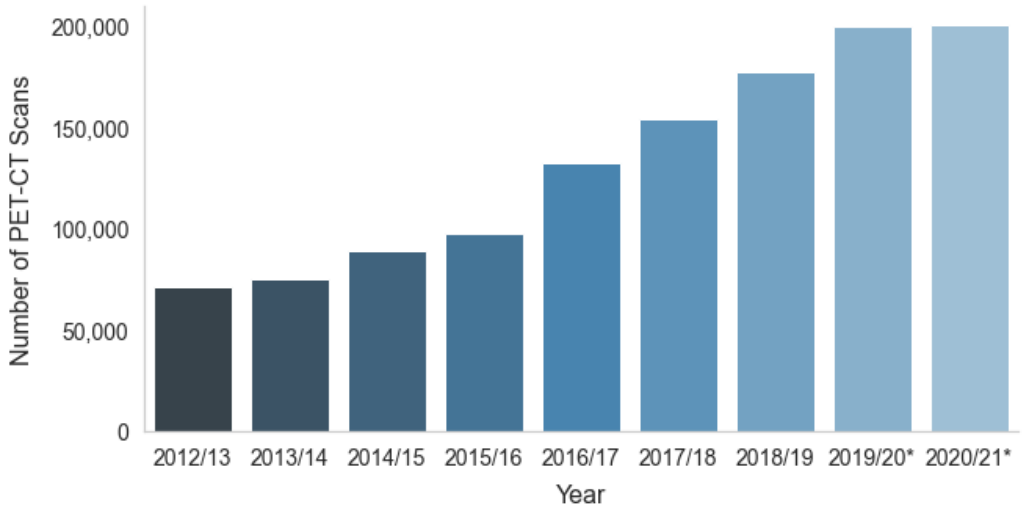


Figure 1.1: A bar chart showing the total number of PET-CT scans taken in NHS England trusts for each year of study across the previous decade [6]. Asterisk (*) indicates years affected by COVID-19 pandemic and restrictions.

1.2 The Current Landscape: Taking a PET Scan

PET is used in clinical environments to quantify the uptake and transport of substances. Advances in chemistry have enabled positron-emitting radioisotopes to be incorporated into many different molecules to test and examine a wide variety of biological processes. Table 1.1 shows some commonly used PET isotopes and associated radiochemicals. Those highlighted in this table, and by the RCR/RCP guidelines [7], are only a few of the thousands of candidates that have been identified and investigated, and new PET radiotracer development is a sizeable area of research. By using transport- and uptake-based radiochemicals, information can be interpolated regarding the location of tumours and their behaviour that is often less easily obtained from density- and material-dependent imaging modalities such as CT and MRI.

FDG is the most widely used PET radiochemical, but it is a non-specific imaging process. It becomes less useful for imaging tumours which occur near areas

[2] *Neuroendocrine tumour*

Isotope	$t_{1/2}$ mins	$E^{\beta^+}_{max}$ MeV	d_{β^+} mm	Attached to	Used to study
^{18}F	110	0.64	2.3	fluorodeoxyglucose, FDG fluorodeoxythymidine, FLT fluoride (e.g. NaF)	Glucose metabolism Cell proliferation Bone metastases
^{11}C	20	0.96	3.9	choline methionine	Prostate cancer detection Brain tumour detection
^{68}Ga	68	1.9	2.2	DOTATOC / DOTATATE	NET [2] detection

Table 1.1: Some uses of positron-emitting radioisotopes in oncology. Also listed for each isotope are the half-life ($t_{1/2}$), maximum positron energy and positron range (d_{β^+}). Further examples of common radiotracers can be found in the RCR/RCP guidelines [7].

of high natural glucose uptake such as the bladder or brain, and can be easily compromised - for instance, by increased muscular use or digestion in the period immediately before a scan. This is where other radiotracers, that may be targeting specific organs or tissues, become of great importance.

In the clinic, a patient is typically given a dosage of the order of 100s of MBq for oncological imaging. The given activity is patient- and investigation-specific. Typically activity dosages are adjusted based on patient weight, among other factors - further information is given in the Administration of Radioactive Substances Advisory Committee (ARSAC) Guidelines [8]. The patient is left to absorb this activity for an allotted period of time before the imaging process, allowing for absorption of the radiolabelled substance in the body. This period of time is different for every patient, substance and imaging protocol. For a typical ^{18}F-FDG scan, the patient is left for an hour to allow for complete absorption [9].

The scan itself is, again, dependent on what is being imaged. Almost every single modern PET scanner sold is dual-modality, with the most common joint modality being CT. A PET-CT scanner is built with a single bore construction, so that a patient laying on the bed is passed through the CT gantry before the PET acquisition. The average PET scanner has an axial field of view (FoV) of around 20 cm, requiring patients to be imaged at several bed positions, or passed through the scanner with continuous bed motion. For a typical oncological ^{18}F-FDG scan, this is done from the base of the skull to the mid-thigh, with the entire scan lasting between 15 and 30 minutes [9]. The patient is required to remain as still as possible throughout the process, to enable better co-registration between the CT and PET data, and preventing blurring and artifacts in the resultant PET image.

The images are reconstructed on a computer associated with the scanner. Each scanner manufacturer produces software that is configured to provide the highest quality image according to clinical needs. The resultant images are stored in the particular trust's PACS (Picture Archiving and Communication System). Clinicians are then able to access this data upon request. In order for this data to be used in research, full security and anonymisation protocols must be followed.

1.3 Data and Image-Based Research in Nuclear Medicine

PET can be thought of as a fully quantitative imaging process; each voxel[3] value represents the activity concentration of tracer that could be localised to that region. The resultant images therefore can be thought of as so much more than just graphical representations of metabolic processes. The PET image is a 3D radioactivity distribution matrix[4], and with a full understanding of the data, analysing patterns and emergent properties of the distribution could unlock new ways of diagnosing patients, improving the quality of treatment, or reducing radiation dose, to name a few applications.

The term *biomarkers* refers to statistics that describe physiological characteristics. Biomarkers can be image-based features, collected personal data - or even genomics, data derived from genome sequencing. Reported clinical image-based statistics, such as SUV_{max}[5], rely purely on first order statistics, meaning that they are largely based on the distribution of the raw voxel values (such as the mean and maxima). For oncology PET, this is still useful information; SUV is an easily obtainable value with a tangible relationship to physiology [12]. However,

[3] A voxel is the 3D equivalent of a pixel, used in medical imaging. The word is derived from 'volume pixel'.

[4] Depending on the scan protocol, this can even be 4D, as scans can be gated to account for patient breathing or movement. PET can also be used for kinetic modelling, measuring the rate of tracer delivery; this could give more physiologically robust information on tracer uptake which could be used to better predict treatment responses [10, 11].

[5] SUV, or standardised uptake value, is equivalent to the activity concentration normalised to the injected activity. The *max* suffix corresponds to the maximum value in a region of interest. This definition will developed further in Chapter 2.

around 2010, studies were beginning to find new ways of using more complex image-derived features to classify patient outcomes, paving the way for the more defined field of *radiomics* [13, 14]. Radiomics is the term given to a higher-throughput extraction of a large collection of image-based features, designed to give a detailed, spatially aware statistical profile of a region of interest in an image (for instance, the patient's tumour) [15]. In an age where increasingly powerful computing is more accessible, the extraction of extra data from a scan at low additional cost is extremely valuable. Artificial intelligence and machine learning, while regarded with a certain degree of doubt in their efficacy by a large part of the clinical community at present, is inevitably going to affect how many clinical processes are carried out in the future [16]. Such algorithms will require as much data as can possibly be sourced before they can be trusted in clinical practice.

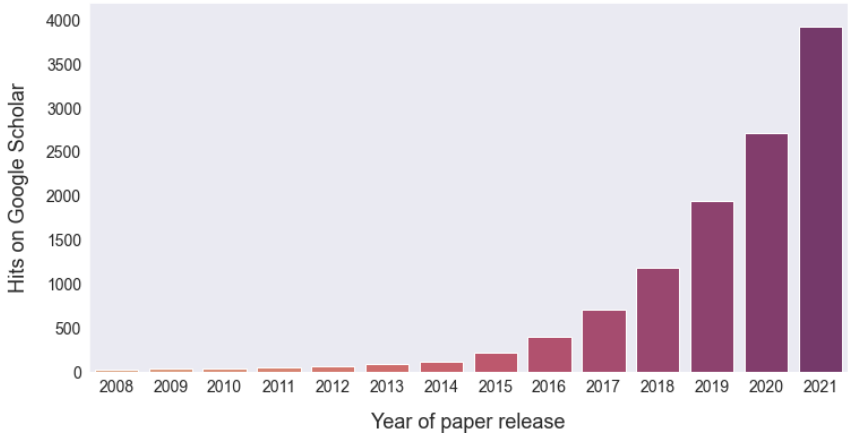


Figure 1.2: A bar chart showing the number of hits on Google Scholar with titles containing "PET" & "radiomics" against publication release year.

Interest in radiomics is growing exponentially (see Figure 1.2) because of the perceived use of this additional data [17, 18, 19]. Studies are returning with encouraging results, but the implementation of any of these research algorithms into the clinic remains distant for several reasons. Firstly, the algorithms are generally treated as a 'black-box approach', lacking clear explanation as to how the information obtained may be useful, while the datasets typically used in PET studies are very small compared to those typically used in other AI research, which makes results difficult to interpret [20]. Secondly, and especially in PET,

current harmonisation and standardisation procedures that take place are not seen as comprehensive enough to handle the sensitive and convoluted process of radiomics feature extraction [21, 22]. Concerns have been raised since 2011 that ^{18}F-FDG imaging protocols lack sufficient standardisation across the field, and introducing higher complexity to this is likely to create a real problem for interpreting study results [23].

Improving harmonisation and standardisation of all PET studies is seen as a priority across the field. The definitions of these terms is slightly different; standardisation implies that variation in all procedures is reduced to a minimum, whereas harmonisation relates to reporting standards, and the ability to reliably compare results between centres often despite different equipment. Harmonisation has already made an impact in the clinic, with SUV harmonised using the EARL (European Association of Nuclear Medicine Research Ltd.) reporting standards [24]. Initially developed in 2017, these were notably updated in 2019 to create EARL2 [25]. One beneficiary of SUV harmonisation is the Deauville criteria, a test used in lymphoma patients to determine the likelihood of treatment response based on SUV values in lesions in comparison to reference values [26]. There is no reason that clinically-relevant criteria such as Deauville could not be developed using radiomics features, should they be definitively shown to predict lesion behaviours of some kind. The process of harmonisation is dynamic, and even the most current efforts in standardisation will become obsolete once new scanner technology becomes implemented into the clinic. As such, developing a harmonisation standard for radiomics feature measurements requires constant feedback from studies which examine how these feature measurements can be more reliably stated.

1.4 Thesis Overview

Most of the interest in radiomics surrounds the ability to powerfully quantify and describe the heterogeneity of a tumour using detailed and nuanced features derived from *texture matrices*. Considered and detailed approaches to characterising tumour heterogeneity has long been proposed as a way by which to better diagnose cancer behaviour, with radiomics answering this need [27]. As shall be

explored further in Chapter 3, many radiomics features describing textural heterogeneity require convoluted and complex mathematical derivations from the original images. PET imaging by nature is difficult due to the noisy data that is collected, and therefore radiomics texture features are notoriously difficult to quote with reliable and accurate uncertainties.

The work in this thesis sought to investigate the robustness of these radiomics texture features. To determine a feature's robustness means to determine the feature's stability given changes to the experimental conditions. It is known that PET scanners exhibit different noise characteristics depending on the amount of activity within the scanner. This motivated an initial investigation determining texture features' instability when changing the total activity level in a scan, while keeping the distribution of the activity the same. These ground truth radioactivity distributions can be created from either phantom scans (plastic models with varying degrees of anthropomorphism) or Monte Carlo computational simulations. The work evolved to consider how novel measures of noise could be used to improve the use of these texture features, and how the phantom study could be augmented with further simulation techniques.

Chapter 2 will detail all of the physics theory and background information required to interpret the results of this thesis. Following this, Chapters 3 and 4 establish the motivation behind the work and the experimental techniques. The methods of creating a set of custom tumour phantoms is described in the latter, detailing how a 3D printed model of patient geometries can be created. Chapter 5 lays out and discusses the initial findings of the texture feature robustness analysis, while the work done over the course of this PhD into developing new noise metrics and creating useful Monte Carlo simulations are elaborated upon in Chapters 6 and 7 respectively. The final chapter will describe the findings of this work, along with recommendations to those currently working in the field that result; finally, conclusions are made, summarising how current cutting edge technology and research in PET could augment this work.

Chapter 2

PET Background & Theory

In this part of the thesis, the physics theory behind the work will be laid out. It covers the basic principles of β^+ emission, PET scanning, PET data corrections, how PET images are created, and an overview of how image-based quantitative PET research is performed. A key reference for this chapter is the `pyradiomics` documentation, in which the mathematical descriptions of all features (also listed in Table 2.2) can be found [28].

2.1 β^+ Decay

Positrons are emitted from proton-rich isotopes by random decay facilitated by the weak nuclear interaction. Also released in this emission, displayed in Equation 2.1, is an electron neutrino.

$$ {}^{A}_{Z}X \longrightarrow {}^{A}_{Z-1}Y + {}^{0}_{+1}e^{+} + {}^{0}_{0}\nu \tag{2.1} $$

The positron is emitted with a kinetic energy dependent on the dynamics of the other 2 bodies, and as such positron energies are described by a spectrum, an example of which can be seen in Figure 2.1. An emitted positron travels a short but significant distance from the parent nucleus, known as the *positron range* [29]. The range is dependent on the initial kinetic energy of the positron, the density and atomic number of surrounding matter, presence of magnetic field among other factors [30]. The positron loses energy along this distance via scattering

interactions until it is thermalised and subsequently annihilated by a nearby electron. To conserve momentum in this interaction, the product photons have equal energy of 511 keV and are released back-to-back, or *collinearly*[6].

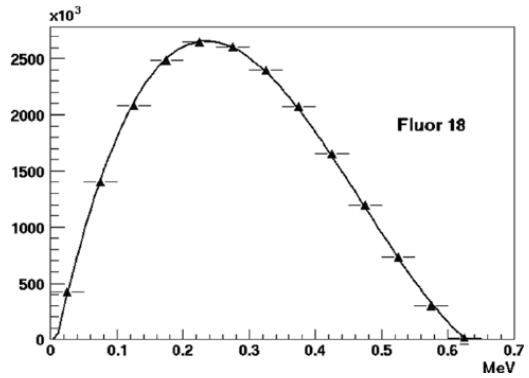
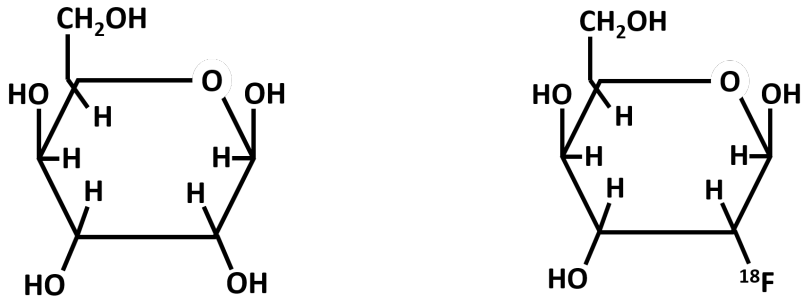


Figure 2.1: Example spectrum of kinetic energies for emitted positrons from an ^{18}F nucleus [31].

2.2 Positron Emission Tomography (PET) Principles

PET scans are performed using *radiopharmaceuticals*. These are chemicals, which may be drugs or proteins, which are *radio-labelled* with a positron-emitting isotope. Some examples of these are listed in Table 1.1. One example, and the most commonly used in PET, is FDG, shown in Figure 2.2. FDG is an analogue of glucose, the simple sugar used by all cells in the human body for metabolism. Hanahan & Weinberg's 2000 study [32] laid out six 'hallmarks of cancer' (later increased to ten hallmarks in a subsequent reevaluation [33]) outlining the processes that distinguish cancerous cells from regular tissues. One of these latter hallmarks is titled *Reprogramming Energy Metabolism* [33]. It is observed that tumours have an increased metabolism compared to the surrounding tissues - a process originally likened to fermentation in a seminal paper by Warburg in 1956 [34]. As such, glucose will have a higher uptake in cancerous tissue than the comparative 'normal' tissues surrounding the tumour.

[6]There is in fact a slight deviation from collinearity, due to the initial positron momentum.



(a) Glucose

(b) 2-deoxy-2[^{18}F]fluoroglucose / FDG

Figure 2.2: Molecular diagrams showing the difference between glucose (left) and FDG (right).

The metabolism process involves breaking up the glucose molecule by a series of chemical reactions - a process known as glycolysis. One of these reactions involves the 'OH' group that has been removed to form FDG (hence 'deoxy'). As this stage of the glycolytic pathway, the FDG molecule cannot continue to be broken apart. Consequently, FDG builds up within the cell until the ^{18}F atom decays into an oxygen atom, whereby glycolysis may continue.

2.3 PET Scanners

The aim of a PET detector is to locate where collinear 511 keV photons are simultaneously detected - a *coincidence* event. A *line of response*, or LOR, is identified by the straight line between the coincidentally-fired detectors, as seen in Figure 2.3. The location of the emission source is realised by the overlapping of multiple LORs. The early PET scanners used two opposing planar detector arrays, but this approach was quickly eschewed in favour of a cylindrical array of detectors around the subject. These cylindrical rings were originally separated by lead septa, preventing coincidence detection across 'slices' - thus the technique is referred to as 2D PET. Aided by technological advancement, modern scanners have removed the need for septa, and as such are now dubbed 3D PET. This mode of acquisition increases the sensitivity (more photons reach the detectors) but increases the likelihood of random & scattered events, which will be discussed further in later sections.

Modern scanners, such as the Siemens Biograph mCT used in this work,

use scintillation crystal detectors made from lutetium orthosilicate, or LSO. The mechanism behind scintillation detection is thus; a gamma photon hits the crystal, exciting electrons which relax and emit photons of a lower energy - around 30 photons for each keV of incident radiation [3]. The scintillator crystal should be transparent to these product photons, which are then measured by a photomultiplier tube (PMT) and an electronic signal produced. LSO as a material meets these criteria; it is able to transmit the visible photons easily and has a good sensitivity to gamma photons (due in part to its comparatively high density). One drawback is the material's poor energy resolution. For PET imaging this is not highly problematic as the only photon energy of interest is 511 keV; a reasonably wide energy window (typically $350-650$ keV) can be used to record an event

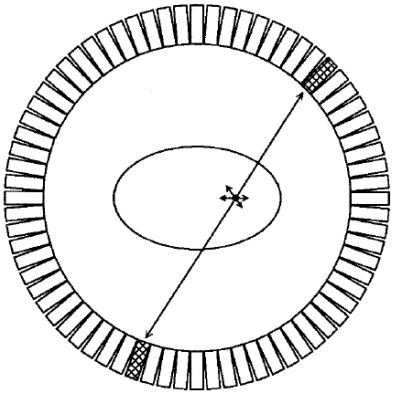
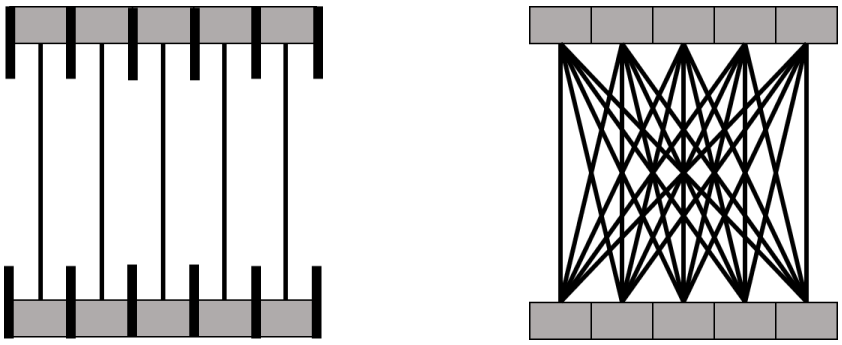


Figure 2.3: An emission event occurs within a ring of detectors. Two detector elements fire, forming an LOR [5].



(a) 2D PET, with lead septa

(b) 3D PET with lead septa removed

Figure 2.4: Basic sensitivity diagrams for 2D & 3D PET, demonstrating the inter-ring acquisition capability for each modality.

with the assumption that the only photons present in this range originate from a positron annihilation. This is certainly the case for ^{18}F, which decays via β^+ with a 97% branching ratio (the other 3% through electron capture). However, other commonly used PET isotopes, including ^{124}I or ^{89}Zr, have some degree of gamma emission in their decay schemes which may add inaccuracy and noise to the recorded data.

The Biograph mCT uses a modular system; four rings each containing 52 'block' detector modules. One module, illustrated in Figure 2.5, consists of a block of LSO with a series of deep cuts into the surface filled with a light-reflecting material, emulating a 13×13 array of crystals. This block is attached to a 2×2 array of PMTs. The ratios

$$R_x = \frac{A+B}{A+B+C+D} \tag{2.2}$$

$$R_y = \frac{A+C}{A+B+C+D} \tag{2.3}$$

of the signals from the 4 PMTs (A-D) are used to determine which of the 169 individual 2D crystal elements have been triggered by the incident photon [35].

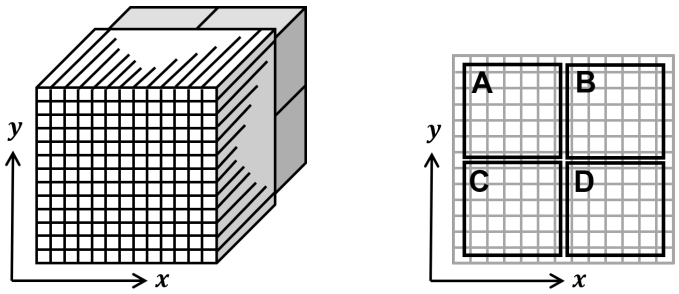


Figure 2.5: Diagram of a detector module consisting of a cut LSO block attached to 4 PMTs.

2.4 Computed Tomography (CT) and PET

X-ray computed tomography, commonly known as CT, is the method of internal 3D imaging based on the attenuation of X-rays through the subject. X-rays interact with matter mainly via Compton scattering and the photoelectric effect,

causing incident photon beams to attenuate. The energy deficit between the outgoing beam and the detected beam is related to a material-dependent attenuation coefficient μ, a measure of the fractional loss in intensity per unit length. These μ values are conventionally converted into Hounsfield units (HU), by normalising to the attenuation coefficients of air and water,

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}}. \tag{2.4}$$

These values can be interpreted as a measure of density of material within the body; for instance, bone has an HU value of around $1800 - 2000$ while lung tissue, for example, has a typical HU value of around 400 [36].

A CT scanner consists of a rotating gantry of an X-ray source collimated to a fan-shaped beam opposite an arc-shaped array of detectors. A 3D projection set of μ values are then collected and reconstructed using similar methods to those used in PET.

The first PET-CT scanners were built in 1998, and the joint modality scanners are now accepted as the industry standard in clinical practice [4]. While having the CT scanner attached to the PET gantry there is an advantage relating to the ease of alignment when creating attenuation maps. The main advantage to a clinician is the contextualisation of the PET data, provided by seeing the uptake hotspots in parallel with the material-based information provided by CT. In general, CT images also have a higher resolution than PET images; the voxel size in a CT image obtained from the Siemens Biograph mCT is $1 \times 1 \times 3$ mm^3 compared to $4 \times 4 \times 3$ mm^3 in the PET image.

2.5 PET Data Corrections

It is not the case that every coincidence detected by the PET scanner corresponds to a *true* line of response. It is necessary to apply corrections to the recorded data to account for the *scatter* and *random* coincidences that are recorded in addition to these. Further corrections are applied to account for *attenuation* in the subject medium, *dead time* in the detectors and *normalisation* of the detector geometry. This section details the origins of these erroneous coincidences, and how they are

corrected for in the data.

2.5.1 Scatter

It is common for a photon travelling through a medium to be scattered; it is estimated that 15% of recorded events in 2D PET, and up to 50% of recorded events in 3D PET, are erroneous scatter measurements [37]. This is usually through the Compton effect, reducing the energy of the scattered photon by an amount relative to the angle of scatter,

$$E_{\gamma'} = \frac{E_\gamma}{1 + \left(\frac{E_\gamma}{m_e c^2}\right)(1 - \cos(\theta))}. \tag{2.5}$$

If the angle of deviation is large enough, the energy may be reduced to an extent such that it falls short of the energy window set by the detectors to record single events. However should this not be the case, a false LOR will be recorded, as shown in Figure 2.6.

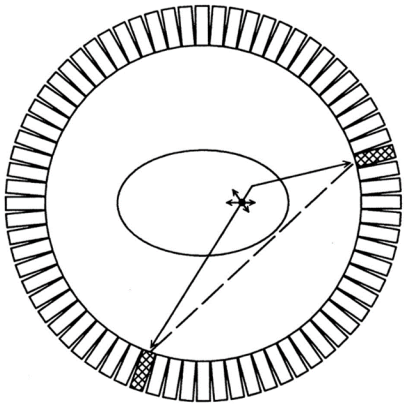


Figure 2.6: Diagram of a scattered event and the erroneous LOR recorded [5].

Scatter may be accounted for in several ways. The traditional approach, and one that is used primarily in SPECT imaging, is that of energy window manipulation. If data is recorded for an energy window set below that of the expected photopeak, the entirety of the data recorded in this lower energy window can be attributed to scatter. This *scatter fraction* is then subtracted from the main energy

window dataset. This is known as 'dual energy window' scatter correction, and can also be used with two lower-energy windows (thus 'triple energy window' scatter correction). Energy window techniques do not tend to be used with modern LSO-based PET detectors, due in part to the poor energy resolution of the scintillation material.

Another method of scatter correction involves estimating the total scatter by using the recorded coincidences for LORs outside the subject, such as that shown by Figure 2.6. These out-of-body LORs, the tails of the intensity distribution at any projection, are trivially attributed to scatter, therefore the scatter fraction within the subject can be interpolated from this data and removed as such (seen in Figure 2.7).

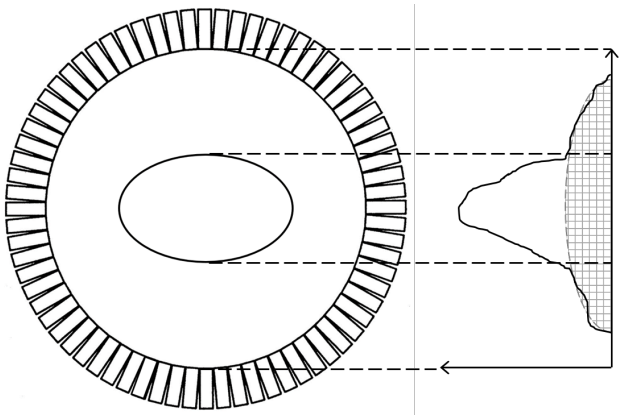


Figure 2.7: The grey region is interpolated from the tails of the given projection and is attributed to scatter.

A preferential method of scatter correction uses Monte Carlo-based methods, statistically simulating the scatter fraction of data in any given system. This relies on an extensive knowledge of the system that is being scanned and thus can become very computationally expensive. This becomes complicated further by considering phenomena such as multiple scatter, and scatter from outside of the field of view (FOV)[7].

[7]This is usually accounted for by shielding the outermost sides of the detectors with lead. This does however invariably leave a small range of angles of acceptance out of the axial extent of the scanner.

2.5.2 Randoms

A random event also results in an erroneous coincidence being recorded, seen in Figure 2.8. This happens when two detectors are triggered within the coincidence timing window, but the photons do not originate from the same source. The two 'lost' partner photons may have either scattered outside the detector, been attenuated and lost within the subject, or may have travelled straight through the PET detector without being recorded.

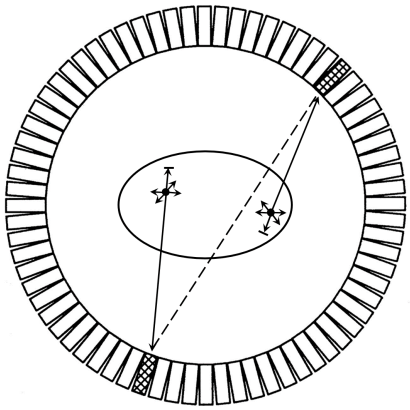


Figure 2.8: Diagram of a random event and the erroneous LOR recorded (adapted from [5]).

Two methods widely used for randoms correction are *elementwise* and *delayed window* estimation. For an elementwise estimation, the rate of random coincidence events, R_r, between two detectors in a coincidence window τ can be expressed as

$$R_r = 2\tau \cdot R_1 \cdot R_2 \tag{2.6}$$

for each detector's singles rate $R_{1,2}$. The total timing window is twice the coincidence window to cover the entire time ($\pm\tau$) that detector 2 can be triggered in order to be in coincidence with detector 1. The random rate for each LOR is computed from these individual singles rates, and can thus be accounted for [38].

Delayed window estimation is the contemporarily preferred method for randoms estimation. This method involves duplicating one of the data channels and delaying it by a time period longer than the coincidence window. Any 'coincidences' detected by the processing electronics in this delayed channel must therefore be random, and can be removed from the data. The disadvantage of

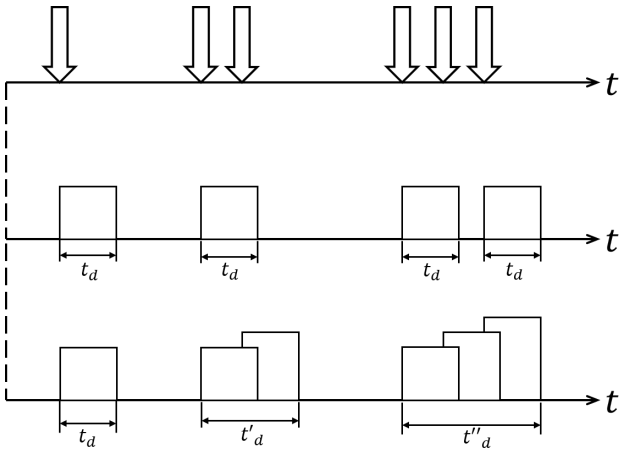


Figure 2.9: A demonstration of dead time models for non-paralyzable (middle) and paralyzable (lower) detectors when given events (upper) within their characteristic dead time per event, t_d.

this method is the doubling of the statistical Poisson noise for each LOR from duplicating the data, but it is commonly viewed to be a more accurate estimation method than the alternatives [38].

2.5.3 Dead Time

Each scintillation detector element has an inherent dead time, the time after the detection of a photon where the detector is unable to detect another incident photon. Detectors can be modelled as either *paralyzable* or *non-paralyzable* during this time. These models are shown in Figure 2.9. A non-paralyzable detector will not 'see' a subsequent event during its dead time, whereas a paralyzable detector will have its effective dead time extended by the next event; from the six events in Figure 2.9, the non-paralyzable detector will detect four, whereas the paralyzable detector only detects three.

The obvious problem with dead time centres on the loss of singles in the system. However, by predicting the behaviour of each model, the theoretical actual count rate, n, can be estimated from the measured values, m at each detector. In the non-paralyzable model, the detector will be rendered 'dead' for a period of time $m \cdot t_d$; the rate that singles events are therefore lost to the dead time of the

detector is therefore $n \cdot m \cdot t_d$ [38]. This is hence equivalent to

$$n - m = nmt_d \tag{2.7}$$

and can be rearranged to give

$$n = \frac{m}{1 - mt_d}. \tag{2.8}$$

A paralyzable detector, on the other hand, can be modelled with the Poisson statistics that govern the interval between successive event detection. It can be established that the probability of an event occurring in the interval dt is

$$F(t)dt = \mathrm{P}(0) \times ndt \tag{2.9}$$

where $F(t)$ is the distribution function, $\mathrm{P}(0)$ is the probability of no events occurring in the interval, and n the Poisson rate of event occurrence. The measured count rate in a paralyzable detector is thus

$$m = ne^{-nt_d}. \tag{2.10}$$

This must be solved for n numerically [3]. Further analysis of how these equations are implemented in correction algorithms may be found in [39].

In order to use this information in correcting for dead time, the characteristic dead time t_d must be found. There are many ways that this can be achieved, but prominently used is the 'two-source method'. Using two sources of different activities, the dead time can be determined by the difference between the measured rates both independently and measured together. The nonlinear response of dead time models ensures that the measured count rate will be less than the sum of the individual count rates [3]. In the Biograph mCT, correction for the scanner's inherent dead time is applied in the process of data storage, and is done on the level of each LSO block.

2.5.4 Attenuation

It is estimated that 60% of the 511 keV photons in any given scan are not detected due to attenuation, the term given to the loss of radiation photon energy [37]. In the case of photons of this energy, attenuation occurs through Compton scattering and the photoelectric effect. The loss of energy through attenuation into a patient is the largest concern regarding patient safety. Attenuation also plays a role in reducing image quality; factors such as the depth through an object the photon has to travel and the density of regions within the object will result in either photons being completely absorbed or reaching the detector with an energy below the window established for legitimate singles measurements.

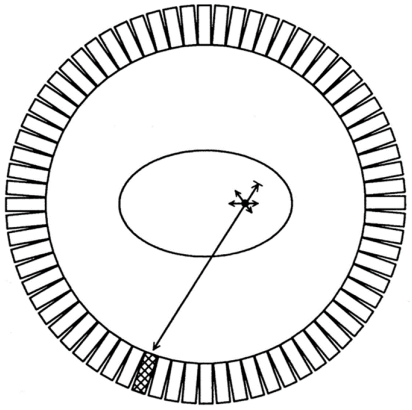


Figure 2.10: An illustration of how a single photon detection may be indicative of attenuation in the subject medium [5].

This loss of singles is accounted for with another applied correction factor. Attenuation is measured by either using X-ray CT scans, or by rotating a ^{68}Ge rod source around the object in the PET scanner - known as *transmission scanning* [40]. Using ^{68}Ge gives the benefit of directly using 511 keV photons to measure attenuation, a benefit not seen when using X-rays. The attenuation coefficients of the two are related by

$$\mu_{511\text{keV}} = \mu_E \cdot F \tag{2.11}$$

where F is described as 'the ratio of the mass attenuation for water at 511 keV to that at E keV' [41]. Here, the mass attenuation is simply the attenuation coefficient divided by the mass density, μ/ρ_m.

A map of the attenuation factors corresponding to every voxel, or μ-map, is created. This map, when compared to the PET data, will inform the reconstruction algorithm about the proportion of photons originating in each voxel that are likely to go undetected. Most attenuation correction algorithms will then proportionally add counts to the corresponding voxels. There are some drawbacks to this statistical process; for instance, the presence of metal objects within a patient may give artificially high uptake readings.

2.5.5 Normalisation

Emissions at different locations in the scanner have an inherently different probability of acceptance. Spontaneous emission and annihilation can produce collinear photons in a 4π solid angle, but the angle of acceptance for both of these photons in coincidence is much smaller and location dependent. Furthermore, deviations from perfect cylindrical geometry must be accounted for; this originates not just from the block-modular arrangement of the scanner, but also any small offsets or alignment errors in these blocks. Other efficiency factors, such as PMT efficiency/gain and the inter-ring plane efficiency, should be corrected [42].

One way of performing normalisation correction is to illuminate every single line of response with a known uniform activity and measure the response. Traditionally this is achieved by rotating a rod ^{68}Ge source around the circumference of the scanner, but can also be performed using a large cylindrical phantom in the centre of the scanner's axis [43].

As computational power has increased over time, the normalisation correction factors are typically calculated in a componentwise manner. This was necessitated by the advent of 3D PET, and the subsequent 10-fold increase in the number of LORs in any particular dataset [44]. Component-based normalisation reduces the number of coincidences needed to measure the correction factors. Several methods for component-based normalisation have been proposed [44, 45, 46, 47]. These techniques use different measurements (phantoms, rod sources, scanning line sources, etc.) in combination with similar, purely analytic, algorithms (e.g. fan-sum algorithm to calculate inherent detector efficiencies and the block profiles) to determine factors by which each LOR's count rate should

be multiplied by to give the same activity regardless of the position of emission.

2.5.6 Partial Volume Correction

The term *partial volume effects* (PVEs) refers to two distinct phenomena that contribute to image quality degradation: finite spatial resolution of the detector system, and voxelised image sampling [48]. The spatial resolution of a PET scanner is hampered by many of the effects highlighted in previous sections - effects emanating from uncertainties in the physical processes (non-collinearity, positron range), the finite detector element size or DOI effects [49]. The limited spatial resolution causes 'leaking' of intensity values between regions. An object of finite size will appear larger and less intense in the data, illustrated in Figure 2.11. This has the effect of the idealised ground truth data being multiplied by a detector-specific point-spread function, or PSF [48].

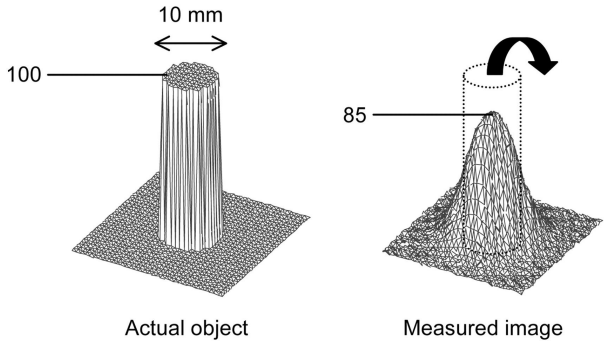


Figure 2.11: The 'leaking' effect of poor spatial resolution in 2D [48].

Image sampling causes a further leaking effect between regions. As each voxel has a finite size, often a single voxel contains multiple tissues with different uptake values. The total voxel intensity is the mean of the various intensities within.

PSF modelling is the method used to reverse this process. An approximation of the detector's PSF is applied to each image update before it is reprojected in MLEM/OSEM. Subsequent expectation maximisation will then correct for this blurring, and after an appropriate number of updates the resultant image should converge with clearer-contrasted boundaries, as if the detector had negligible in-

herent partial volume effects. However, while this improves the image quality and convergence over the course of reconstruction, it has been found to artificially affect the quantitative nature of the imaging process [50].

2.6 PET Image Reconstruction

The intensity measured at each LOR is plotted against the position of the LOR in a form of projection mapping called a *sinogram* - so named because a point source becomes mapped to a sine wave. This is also known as a Radon transformation. In 2D PET, the lack of inter-slice acquisition[8] resulted in sinograms being binned by slice, and therefore each slice reconstructed independently. In 3D PET, sinograms are binned by azimuthal projection angle.

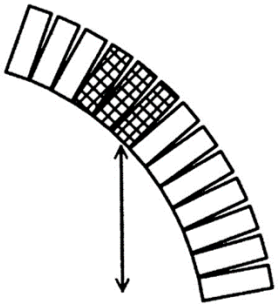


Figure 2.12: An illustration of the depth of interaction (DOI) problem, whereby an incident photon may be detected by one of a series of detector elements [5].

Each acquisition dataset is typically subject to two reduction factors in order to reduce the data storage requirements. A *mashing* factor can be applied which bins together counts from LORs with small differences in crystal spacing; at the same time, a *span* factor is applied which bins together inter-ring coincidence data within a certain azimuthal projection angle. Several sources of uncertainty allow these approximations to hold validity, particularly the non-collinearity of photons and the depth of interaction problem. Non-collinearity refers to the phenomenon by which the 511 keV photons from an annihilation event are actually

[8]This is a simplification, as the geometry of the septa allowed acquisition for a small span of rings. The data for each of these small-span inter-ring coincidences were usually rebinned into the most appropriate cross-sectional slice.

released with a slight angular deviation (around $\pm 0.25°$) from perfect collinearity; the wider the diameter of your PET rings, the more exaggerated the effect of non-collinearity becomes. The non-collinearity is caused by the conservation of momentum from the parent positron-electron pair. The depth of interaction problem, on the other hand, describes the scenario whereby a gamma photon may have enough energy to travel straight through the first detector element it reaches, and is measured instead by a nearby element, shown in Figure 2.12.

2.6.1 Analytic Reconstruction

In 2D PET, the analytic technique preferentially used to obtain the original image from the Radon transformation is filtered back-projection (FBP). Projections of an object at single angles, as seen in Figure 2.13, can sometimes destroy information about relative depth. By back-projecting these projections at a range of planar angles, it is possible to build up a complete image of the object. A ramp filter is applied to each projection before the back-projection process to prevent the inherent radial $(1/r)$ blurring [37].

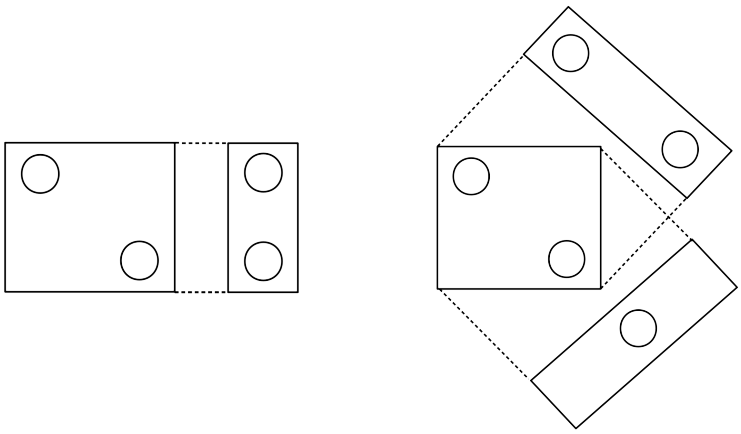


Figure 2.13: Projecting an object onto a plane gives little information about depth (LHS) unless further projections at a range of angles are taken (RHS).

2.6.2 Iterative Reconstruction

Analytic reconstruction techniques are very computationally expensive, and the back-projection stage can take a long time to perform with a large number of projection angles. It is usually preferential to use iterative reconstruction processes, particularly in 3D PET. One such algorithm is Maximum Likelihood Expectation Maximisation (MLEM). The algorithm predicts the intensity of each voxel using Poisson statistics based on the recorded LORs and the probability matrix $a(i, j)$, where

$$a(i, j) = \mathrm{P}(\text{event detected in LOR } j \mid \text{event emitted in voxel } i)\,, \tag{2.12}$$

which can be calculated from the PET detector geometry. A full description of the MLEM algorithm can be found here [51, 52].

In order to decrease computation time and expense, the MLEM algorithm can be run in parallel upon subsets of the recorded data. This implementation, Ordered Subset Expectation Maximisation (OSEM) is the algorithm mainly used to reconstruct 3D PET data [53]. It should be noted that the 'maximum likelihood' prefix is dropped, as while the computed image for each subset will be calculated in accordance to the maximum likelihood, the resultant image cannot explicitly be said to have maximised the global likelihood. However, it has been empirically shown that the OSEM-computed image, even for a large number of subsets, will have little or no resultant drop in quality to the MLEM computed image.

There are issues with iterative reconstruction. As the number of updates[9] is increased there is a resultant increase in convergence, but this is compromised by an increase in background noise in the image. This is generally caused by artifacts within the image estimation that become exaggerated as the number of updates continues.

[9] One update can be considered as one iteration of the MLEM algorithm; for OSEM, this corresponds to the number of iterations multiplied by the number of subsets, $N_i \times N_s$.

2.7 Quantitative PET and Radiomics

FDG PET performs exceedingly well as a metabolic imaging modality, with visual inspection alone enabling clinicians to identify cancerous material with ease in many cases. PET was, however, developed as a quantitative tool, and over time there has been an increased appreciation of the deeper objectivity of PET quantification [54]. Quantifying the uptake of FDG in a region of interest has proved useful over the years for earlier diagnosis and to determine how patients are likely to respond to therapy [54]. Multiple conventions exist for quoting uptake in a medical study, and these are detailed in Section 2.7.3.

Radiomics was developed to harness the power of the objectivity of quantitative PET. A radiomics extraction process is described in Figure 2.14. The key steps involve: delineating the ROI from the image (known as segmentation); discretising the image such that voxel intensities are allocated into gray level bins for texture matrix processing; and feature extraction - the mathematical calculation of statistics from this information. In addition to these basic principles, other pre-processing steps are required. For an individual patient, PET radiomics is often done in conjunction with other image-based processes (for instance, radiomics extraction performed on a companion CT scan); this requires resampling and interpolation of the images into the same matrix spacing[10]. For PET, images are required to be converted into voxel units of SUV (see Section 2.7.3) before discretisation and extraction.

Several free dedicated radiomics software are available for this research. The most popular[11] of these are pyradiomics [55], LIFEx [56] and CERR [57, 58]. The output of an extraction for a single ROI is a series of numbers describing first order, shape and texture features. Examples of these are described in this section.

[10]Image interpolation causes issues with segmentation - in practice, segmentation and interpolation are run in parallel.

[11]These are the most popular IBSI-compliant software; for details of what this entails, see Section 2.7.4.

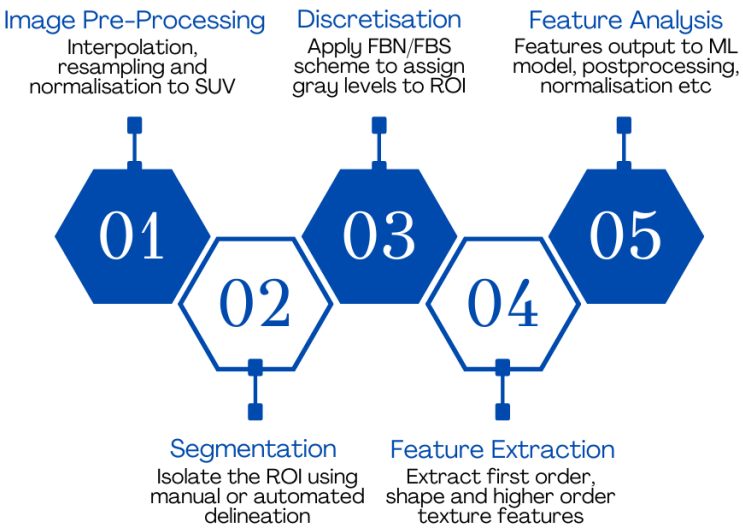


Figure 2.14: Flow diagram illustrating the processes associated with a radiomics feature extraction, from image input to statistical output.

2.7.1 Image Segmentation

The process of selecting the ROI in a medical image is referred to as segmentation. There are three methods used to segment an image: manual delineation, image thresholding and AI-based methods. This first step in a feature extraction is arguably the most crucial, as inaccurate segmentation results in either analysis of undesirable material or missing out crucial parts of the 'true' ROI. These outcomes would in turn not only cause the greatest shifts in our raw metric values, but would be harmful for a patient should these inaccurate ROIs be used to direct radiotherapy treatment.

Manual Delineation

Originally ROIs had to be drawn by a clinician due to the capabilities of contemporary available software. Typically the selection process would entail clicking around each relevant boundary pixel on a slice-per-slice basis. Such a process is hampered mostly by the length of time the process takes, making it inefficient

for regular and popular use. The potential upsides of this method are due to the clinician's specialist knowledge and ability to distinguish materials. Subjectivity and bias, however, remains the largest problem with manual methods. Even two expert clinicians may differ severely in selected ROIs for the same object, and in most cases the job of delineation may not be taken on by experienced staff due to time constraints. A single clinician may also delineate different voxels if shown the same image multiple times. This method is also highly dependent on the image display settings used, which is highly undesirable for standardising any process. This intra- and inter-observer variability is highly problematic, and the uncertainty of delineations causes serious patient safety concerns in image-guided radiotherapy [59].

Image Thresholding and Automated Segmentation

To address the observer dependence issue with manual delineation, more objective methods of segmentation are often used. These automated procedures generally involve thresholding; including all voxels in a local region that fall above (or below) a predetermined value. This method is fast and, to a certain extent, reliable and repeatable; any two clinicians will obtain the same ROI using the same threshold value on the same image, even when using different software.

The main drawback to a thresholding method lies in the selection of an appropriate threshold value. Thresholding may be marginally easier to standardise in CT imaging, where a range of Hounsfield Unit values for most tissues are well established, but in PET imaging it is required to know what activity concentration (or SUV) constitutes background over foreground. Common methods in FDG PET include thesholding to 40 % or 50 % of the local maximum, or thresholding to a choice of an appropriate background level, deduced from a region such as the liver [60]. Basic thresholding can lead to phenomena such as missing voxels in the resultant ROI. Techniques developing this include 'growing' the ROI, where an algorithm will step radially out from a chosen voxel, including voxels in the ROI until either a volume criteria is reached or an appropriate boundary is hit [60]. Adaptive methods, alongside AI-based segmentation, are generally successful, although there are concerns that more complex methodologies are less

effective on simpler geometries [61].

2.7.2 Intensity Discretisation

Another stage before full feature extraction is intensity discretisation. The discretised image is only used for texture matrix computation, and not for lower-order image features (such as mean, standard deviation). The discretisation into 'gray levels' requires the choice of using a fixed bin size (FBS) or a fixed bin number (FBN). There is no consensus as to which is the more appropriate method for discretisation, and this is a subject that prompts discussion on a study-specific basis. The choices involving discretisation in this work are discussed in Section 2.7.4. There is a consensus that standardisation and harmonisation is required, and as such further studies are required assessing and validating different discretisation protocols to determine best approaches across the field [62].

2.7.3 Image Features

Uptake & SUV

A PET image is typically presented with voxel units of activity concentration, usually in kBqml^{-1}. This can cause a problem in clinical comparisons as the activity given to any patient is subject to factors including the size and weight of the patient, as well as the investigation being performed. As a result, images are conventionally converted into units of the standardised uptake value, or SUV. The SUV in any voxel i is the ratio

$$SUV(i) = \frac{C(i)}{C_{tot}}, \tag{2.13}$$

of the activity concentration C divided by the total initial injected activity, C_{tot}. This is often approximated by

$$SUV(i) = \frac{C(i)}{A_{tot}/W} \tag{2.14}$$

where A_{tot} is the total decay-corrected injected activity, and W is the patient's body mass in kg. The comparable timescales of the half-lives of PET isotopes (110 minutes for ^{18}F, or 68 minutes for ^{68}Ga) are of the order of the typical duration of a scan, necessitating this decay correction. The time point used is typically halfway through the scan duration.

These values are voxel-specific. However, studies usually investigate the uptake of a whole organ, tissue or other ROI; these values are more useful for comparing values in intra- and inter-scan studies. There are many ways that the uptake of an ROI can be quantified:

- **SUV$_{max}$:** the maximum of the SUV voxel values within the ROI;
- **SUV$_{mean}$:** the mean of the SUV voxel values;
- **SUV$_{mean}^{x\%}$:** the mean of the SUV voxel values that are above a $x\%$ threshold (in literature typically $x = 40$ or 50) of the maximum SUV value;
- **SUV$_{peak}$:** a noise-corrected maximum SUV value, usually found by establishing the maximum of a fitted curve to the intensity distribution or finding the average SUV value in a small volume (e.g. 1 cm^3 proposed by Wahl [63]) around the voxel with the largest SUV value;
- **qPET:** the average of the highest-SUV voxel in the ROI & its 3 highest-SUV adjacent neighbours, divided by the average SUV value of some background reference. In clinical practice the liver is used as the reference [64];
- **TGV:** total glycolytic volume. This is a metric specific to FDG imaging, supposedly quantifying glucose metabolism. It is equivalent to the SUV$_{mean}$ multiplied by the volume, for an ROI delineating cancerous tissue.

Shape

Shape metrics, in this study, can be thought of as having a dual purpose. First and foremost, many have been used in their own right as biomarkers, diagnostic and prognostic indicators. Features extracted include descriptors of volume and

surface area, but also the sphericity, axis lengths and more complex shape parameters. The shape and size of tumours give clinically relevant information, and this is described briefly in the following sections.

Volume & Surface Area

One of the biggest, and most intuitive, predictors of a tumour's response to treatment is the tumour's reduction in volume. Metabolic tumour volume, the volume of a lesion as it appears on a PET scan, is cited in many studies, but the method of quantification is variable. The simplest method involves counting the voxels that may be delineated as lesion material. This results in the complication of volume measures with the delineation problems highlighted in Section 2.7.1. Not only the significant effects mentioned previously such as PVEs and image noise, but also image viewing parameters such as the provided contrast could result in independent manual delineations covering significantly different volumes. Differing methods of delineation have been shown to lead to poor repeatability of lesion volume measures between studies [65].

Sphericity

Sphericity, Ψ, is a shape metric defined by a ratio of a body's surface area and the surface area of a sphere with the same volume as the body, expressed by

$$\Psi = \frac{A_s}{A_b} = \frac{\sqrt[3]{36\pi V_b^2}}{A_b}, \tag{2.15}$$

where A_s is the surface area of the volume-equivalent sphere, and V_b and A_b are the volume and surface area of the body respectively [66]. There have been links drawn between the sphericity value of a tumour and the response to treatment; more irregular shapes could be a result of cell type variation or angiogenesis[12], and consequently may make the tumours more complicated to treat [67]. Hatt et al. (2018) demonstrated how the prognoses of a patient cohort could be stratified

[12]One of Hanahan & Weinberg's hallmarks, angiogenesis is the process whereby tumours may form their own blood vessels in order to enhance their growth to the detriment of surrounding tissue [33].

by the sphericity Ψ of their tumours from PET images [68].

Heterogeneity & Texture Matrices

The noteworthy aspect to radiomics is the implementation of textural analysis using a set of derived matrices known as *texture matrices*. However, textural analysis can be thought of as an extension of measuring heterogeneity of image voxel intensity in an ROI. At a basic level, areas of an image may be described as *locally homogeneous* if the pixel values are the same or similar. Texture matrices then develop this definition by determining the relationships between neighbouring voxels, as well as describing the runs and clustering behaviour of similarly-valued voxels.

The reasons for the clinical interest in measuring the heterogeneity of an ROI in a PET image are manyfold. Variations in glucose uptake across a tumour may be indicative of necrosis, or fragmentation of the tumour's main body. While many studies have found that heterogeneity in PET images are reliable predictors of prognosis or outcome, there is still limited knowledge from the biophysical perspective as to why this is the case [69].

Heterogeneity itself is, however, a relatively vague concept that can have different manifestations of quantification. Any metric that defines the variation of pixel values across an ROI could appropriately be described as a measure of heterogeneity; the statistical variance, or standard deviation, or even the range of pixel values, for instance, all have validity. However, spatial information is lost when considering these first-order metrics. Higher-order metrics can be used to retain some of the information about relative position of pixel intensities. Radiomics libraries typically contain many features which can appropriately define heterogeneity. From its beginnings in CT and MRI, there is a growing interest in the applications of radiomics to molecular imaging [15, 22, 17, 70].

There are doubts over the validity of heterogeneity measurements in PET images. The comparatively large voxel sizes (of the order of a few cubic millimetres) by far outscale the cellular-level changes in heterogeneity that may be of clinical importance [71]. This is complicated by the fact that PET images are inherently noisy even after some degree of correction is applied [22]; in addition,

post-reconstruction blurring filters that are applied to PET images may artificially remove some element of inherent heterogeneity in efforts to remedy the problem of noise elsewhere in the image. As such, few heterogeneity metrics have been found to be reproducible across studies, and studies into which of these may be reproducible have found dissonant results [72, 73, 74, 75]. Heterogeneity also has a complicated relationship to the with the ROI's shape and size. A definitive value of heterogeneity may be difficult to obtain, as a large ROI with a small variation of pixel intensities may have an equal heterogeneity value to a small ROI with a much larger variation of pixel intensities. According to Brooks, any given heterogeneity metric is not a valid quantification measure without prior attention given to the size and shape of the body being measured [71]. Indeed, in Brooks & Grigsby (2013), the heterogeneity metrics were believed to have been confounded by lower-order shape and volume metrics, and were as such not found to differentiate a patient cohort by prognosis [76]. Indeed, studies such as Hatt et al. (2011) show that tumour volume and length alone, when used in tandem with SUV measurements, can be used as prognostic factors for survival [77].

The effect of shape on heterogeneity metrics has also been investigated; for instance, O'Sullivan et al. (2005) were able to determine tumour progression from metrics combining information about the 2D tumour boundary with measured heterogeneity [78]. Brooks & Grigsby (2014) determined an estimate of a minimum comparison volume of 45 cm^3 for regions for which heterogeneity analysis may become valid [79]; many other studies only ignored lesions of sizes below 10 cm^3 [15].

This is not to say that the measurement of heterogeneity changes on this larger scale may not still be of some import. Tixier et al. (2011) used changes in co-occurrence matrices (as described later) to successfully distinguish between partial- and non-responders to chemotherapy in their patient cohort [80] In both shape-aware and shape-unaware comparisons, Brooks & Grigsby (2013) found that higher-order heterogeneity metrics enabled a mathematical ranking of a patient cohort by the heterogeneity of their tumours that matches rankings by experts in oncology and image analysis [81]. De Heer et al. (2018) used heterogeneity metrics to determine which patients were benefiting from treatment in a study of a melanoma patient cohort [82]. Success in studies such as these has

resulted in the continued pace of research in this area; the works mentioned here are only early examples of successes in a burgeoning field. It is apparent though that heterogeneity measurement in PET imaging is lacking a level of standardisation [83]. There is encouraging progress in the use of heterogeneity in PET, but discrepancy between studies in the features found to be useful, and the methods with which these features are collected, result in disagreeing conclusions as to the value of heterogeneity.

The features that can quantify heterogeneity have here been divided into 6 separate groups for purposes of explanation: the first-order voxel intensity parameters, cumulative intensity-volume histogram parameters, size-zone matrix features, co-occurrence-type matrix features, run-length matrix features, and dependence matrix features. It is important to state that not every radiomics package uses all of these features and matrices, but all of the features described apart from the cumulative intensity-volume histogram parameters are included in pyradiomics. This makes pyradiomics perhaps the most comprehensive package available, motivating its use in this work. The mathematical derivations of all features can be found in the pyradiomics documentation [55], and all texture features are listed in Table 2.2.

A) Intensity Parameters

These metrics are extracted from the undiscretised image. Such values include:

- Standard deviation of values;
- Coefficient of variation (CV), equal to the standard deviation divided by the mean;
- Range / interquartile range of values;
- Skewness of values.

These first-order features are included in every radiomics software, however are not analysed in this work.

B) Cumulative Intensity-Volume Histogram Parameters

This is a histogram with bins corresponding to intensity values, counting the number of voxels with intensity equal to or higher than the bin values. It is within reason that the heterogeneity of an ROI will be characteristic to the shape of this histogram. Studies using this method have taken normalised metrics from the area under the curves of these histograms [84], and while these metrics have shown positive test-retest repeatability in the few studies undertaken so far, there is little literature surrounding errors and uncertainties on these parameters [85]. Features from these histograms are not included in many radiomics software, and are not explored in this work.

C) Size-Zone Matrix Features

Each image, visualised as a 3D matrix of voxel values, can be expressed as a size-zone matrix. The original matrix can be said to consist of 'zones' of identically-valued intensities. In this new matrix, the rows i correspond to the zonal intensity value, and the columns j correspond to the number of voxels in the zone [86]. For example, a 2D representative matrix

$$M = \begin{pmatrix} 1 & 2 & 2 & 2 \\ 1 & 4 & 4 & 3 \\ 2 & 2 & 3 & 3 \\ 1 & 1 & 3 & 4 \end{pmatrix} \tag{2.16}$$

would become

$$M'_{\mathrm{GLSZM}}(i,j) = \begin{pmatrix} 0 & 2 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix} \tag{2.17}$$

under this transformation. As with the other texture matrices, there are a series of features that describe this matrix. For example, the *High Gray Level Zone Emphasis* would have a higher value if there were more zones with high gray levels than if there were more zones with low gray levels.

D) Co-occurrence-Type Matrix Features

Two matrices defined in radiomics software have been grouped together here: the gray level co-occurrence matrix (GLCM) and the neighbourhood gray-tone difference matrix (NGTDM). While they ostensibly show different values and properties, they are no doubt closely related - one shows the likelihood that a neighbouring voxel will have the same voxel intensity, while the latter shows the difference in intensities between adjacent voxels. There is no explicit size-dependence to the matrix definition, hence these textural matrices are grouped apart from the size-zone matrix for this study.

The co-occurrence matrix of an image, sometimes called the Haralick matrix, is a square matrix that determines the frequency of instances of neighbouring voxel values. For instance, the matrix M from Equation 2.16 becomes

$$M'_{\mathrm{GLCM}}(i,j|\delta=1,\theta=0)=\begin{pmatrix}1&1&1&1\\1&6&1&0\\1&1&2&2\\1&0&2&2\end{pmatrix}. \tag{2.18}$$

Here, the number of times that an intensity value i appears at a separation of δ voxels from a voxel of value j at an angle of θ to the horizontal is counted in each element of the matrix. The total co-occurrence matrix is normalised over all possible values of (δ,θ) [87].

Another matrix can be defined that quantifies differences in adjacent voxel values - the neighbouring gray-tone difference matrix, or NGTDM [88]. For voxel gray levels i located at $M_{x,y,z}$, the weighted sum of the average surrounding voxels,

$$s_i=\sum^{n_i}|i-\overline{A}_i| \tag{2.19}$$

where

$$\overline{A}_i=\frac{1}{W}\left[\sum_{a=-\delta}^{+\delta}\sum_{b=-\delta}^{+\delta}\sum_{c=-\delta}^{+\delta}M_{x+a,y+b,z+c}\right] \tag{2.20}$$

is calculated. Here, W refers to the number of voxels in the 'neighbourhood' (extent of δ), and n_i the number of occurrences of gray level i.

Table 2.1 shows the NGTDM values for the matrix M from Equation 2.16,

i	n_i	p_i	s_i
1	4	0.25	4.400
2	5	0.3125	2.975
3	4	0.25	1.000
4	3	0.1875	4.250

Table 2.1: Deriving the NGTDM features s_i for the matrix M.

using a neighbourhood extent $\delta = 1$. The values seem logical when analysing M; the lowest s_i corresponds to a gray tone of 3, and in M the elements containing 3 are surrounded by elements of a similar value. As with the size-zone matrix, these matrices must be derived using gray level, or gray tone, integer values.

E) Run Length Matrix Features

A fourth matrix described in radiomics software is the gray level run length matrix, or GLRLM. These runs are given by the number of consecutive pixels (j) with the same value i along the angle θ; the value of $M'(i, j|\theta)$ representing the number of such runs. The matrix M from Equation 2.16, looking only at runs across the horizontal ($\theta = 0$) becomes

$$M'_{\text{GLRLM}}(i, j|\theta = 0) = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 2 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}. \tag{2.21}$$

F) Dependence Matrix Features

One voxel i is dependent on its neighbour j if $|i - j| \leq \alpha$, with α a predetermined threshold level. The dependence matrix has a value at (i, j) equal to the number of occurrences where a gray level i has j dependent surrounding voxels at a distance of δ away. The matrix M from Equation 2.16, applying the thresholds of $\alpha = 0$ and $\delta = 1$, becomes

$$M'_{\mathrm{GLDM}}(i,j|\alpha=0,\delta=1)=\begin{pmatrix}0&4&0\\0&4&1\\0&2&2\\1&2&0\end{pmatrix}. \tag{2.22}$$

There are 75 listed textural features in pyradiomics; 24 GLCM features, 16 GLRLM and GLSZM features, 15 GLDM features, and 5 NGTDM features. These sit alongside 19 first order features and 26 shape features (both 2D and 3D) to provide 120 total. The texture features will form the focus of this work, and these are listed in Table 2.2.

2.7.4 IBSI Recommendations

Many different software have been developed to perform radiomics analyses. The lack of standardisation and harmonisation of algorithm implementation between these software packages is the subject of several active studies [89, 90]. The Imaging Biomarker Standardisation Initiative is a collaboration that was created to address this issue. As well as establishing agreed methods of feature derivation, the IBSI also suggests guidelines and nomenclature for reporting results of radiomics-based analyses [91]. Software that is intended for use in radiomics-based studies must declare its IBSI compliance. pyradiomics is IBSI compliant, and was selected for this work for its flexibility and ease of implementation for large datasets. The version of pyradiomics used in this work was v3.0.1, which was the most recent update at time of writing.

All data extraction performed for this thesis was compliant to the IBSI guidelines for independent radiomics studies. As a consequence, in this work:

- all images were converted into units of SUV;
- all extractions used a fixed bin number of 64 (FBN:64);
- default settings for pyradiomics were otherwise used.

FBN was chosen over FBS due to the comparison in this study between ROIs of differing size and shape, and therefore direct comparison of feature values could

be enabled [91]. In selecting the choice of bin number, work from previous studies was consulted; evidence showed that the bin numbers outside the limits of FBN:32 [92] and FBN:64 [15, 93] provide little informational advantage, or ease of computation. In addition to these considerations, various methods of discretisation were evaluated for an early experimental run that resulted in this selection. A summary of the findings from this testing can be found in Appendix I.

GLCM	NGTDM	GLRLM	GLSZM	GLDM
Autocorrelation	Coarseness	Short Run Emphasis	Small Area Emphasis	Small Dependence Emphasis
Joint Average	Contrast	Long Run Emphasis	Large Area Emphasis	Large Dependence Emphasis
Cluster Prominence	Busyness	Gray Level Non-Uniformity	Gray Level Non-Uniformity	Gray Level Non-Uniformity
Cluster Shade	Complexity	Gray Level Non-Uniformity Normalised	Gray Level Non-Uniformity Normalised	
Cluster Tendency	Strength	Run Length Non-Uniformity	Size-Zone Non-Uniformity	Dependence Non-Uniformity
Contrast		Run Length Non-Uniformity Normalised	Size-Zone Non-Uniformity Normalised	Dependence Non-Uniformity Normalised
Correlation		Run Percentage	Zone Percentage	
Difference Average		Gray Level Variance	Gray Level Variance	Gray Level Variance
Difference Entropy		Run Variance	Zone Variance	Dependence Variance
Difference Variance		Run Entropy	Zone Entropy	Dependence Entropy
Joint Energy		Low Gray Level Run Emphasis	Low Gray Level Zone Emphasis	Low Gray Level Emphasis
Joint Entropy		High Gray Level Run Emphasis	High Gray Level Zone Emphasis	High Gray Level Emphasis
Informational Measure of Correlation 1 (IMC1)		Short Run Low Gray Level Emphasis	Small Area Low Gray Level Emphasis	Small Dependence Low Gray Level Emphasis
Informational Measure of Correlation 2 (IMC2)		Short Run High Gray Level Emphasis	Small Area High Gray Level Emphasis	Small Dependence High Gray Level Emphasis
Inverse Difference Moment (IDM)		Long Run Low Gray Level Emphasis	Large Area Low Gray Level Emphasis	Large Dependence Low Gray Level Emphasis
Maximal Correlation Coefficient (MCC)		Long Run High Gray Level Emphasis	Large Area High Gray Level Emphasis	Large Dependence High Gray Level Emphasis
Inverse Difference Moment Normalised (IDMN)				
Inverse Difference				
Inverse Difference Normalised				
Inverse Variance				
Maximum Probability				
Sum Average				
Sum Entropy				
Sum of Squares				

Table 2.2: A list of the 75 textural features included in pyradiomics [55].

Chapter 3

Motivation & Methodology

The motivation for the work presented in this thesis is discussed in this chapter. The chapter begins by explaining what is meant by robustness, and details how it could be established. The experimental methodology underpinning the main data collection is then given in detail.

3.1 The Three Rs

A conflict of interest exists between producing clearer images and establishing accurate activity concentration uncertainties on a per-voxel basis. The sources of noise in a PET acquisition are numerous and confounding; correction methods for scatter, randoms, attenuation, partial volume effects, dead time, and post-reconstruction filtering all aim to give a more accurate representation of the underlying activity distribution, but consequently producing a picture of the voxel uncertainty becomes ever more complex.

There is a further lack of uncertainty provision in radiomics software. In addition to the reasons mentioned above, this is also due to the segmentation, discretisation and texture matrix manipulation applied to the images before features are calculated. Analytical uncertainty generation for features is not feasible. It also remains to be determined whether radiomics features could be reliably used to represent an accurate picture of an unknown activity distribution. For clinical medical imaging, this is problematic.

Examination of the reproducibility, robustness, and repeatability of these radiomics features, particularly those derived from texture matrices, becomes vital for harmonisation and standardisation of their use in future studies. These terms can be defined by

- **Reproducibility:** whether a feature maintains its value when the experiment is performed on a separate experimental setup;
- **(Test-Retest) Repeatability:** whether a feature is consistent when measured under the same conditions;
- **Robustness:** whether a feature's value is unaffected by perturbations in the experimental conditions.

A radiomics feature can be said to be reproducible, repeatable and robust if the value of the feature is dependent only on the underlying activity distribution. Determining this invariance is, however, fraught with difficulty when using clinical PET data. For instance, testing reproducibility in a meaningful way requires multi-centre studies on rigorously defined subjects, if not the same subject. Firstly, the ethics of repeatedly dosing the same patient with high levels of radioisotopes should be raised. Testing different patients requires accounting for differences in size and shape of the patient, and the size, shape and location of their tumour(s) - all of which we know affect the calculation of our texture features.

In order to determine whether the metrics themselves are truly reproducible, repeatable and robust, we need to determine their invariance on subjects where we do have prior knowledge of the underlying activity duration. This forms the motivation for this project. Two key methods are used to generate ground truth activity distributions: phantom scans, and Monte Carlo simulation. Using these two methods, we can perform multiple repeat acquisitions without ethical repercussions, and push the boundaries of the conditions that could be possible in a patient scan.

Three previous studies of robustness in radiomics features are cited as validation for performing this study [21, 94, 95]. In [21], GLCM features were examined for their robustness to *exposure*, defined there as the product of activity

concentration and frame duration, on a 125 cm^3 ROI in the centre of a ^{68}Ge-filled cylindrical phantom. It was found that the features generally showed poor robustness, although some degree of dependence on exposure was evident. The paper concludes that it may be possible to correct GLCM values to the levels of a 'plateau' obtained where exposure remains approximately constant [21]. In [94], radiomics features were tested for robustness to system, resolution and segmentation method by using custom-produced phantoms, consisting of a series of small heterogeneous cylinders placed inside the NEMA Image Quality phantom (see Chapter 4). The study concludes that only four features (and only one texture feature) are robust to all of the above, with mixed results for robustness to any one of the factors [94]. In [95] repeatability and reproducibility of radiomics features was examined using 3D-printed anthropomorphic tumour phantoms on different systems with differing reconstruction parameters and image frame duration. The three phantoms exhibited a degree of heterogeneity with relatively simple construction. Despite being a relatively comprehensive study, this paper did not consider the effects of robustness due to activity in the scanner and the potential noise differences that ensue. All three papers recommend further examination in this area, and while this work shall complement the work shown, the novel methodology that was used further examines some aspects of these studies in more detail. This methodology is explained in the following sections.

3.2 The Noise-Equivalent Count Rate

In 1990, Strother, Casey & Hoffman defined the Noise Equivalent Count Rate, or NECR, as a metric by which to measure the signal-to-noise features of an acquisition. The NECR is generalised by

$$NECR = \frac{T^2}{T + S + x \cdot R} \tag{3.1}$$

for the true (T), scatter (S) and random (R) rates as determined from the raw data [96]. The factor x by which the random rate should be multiplied is subject to the method of random rate estimation, but is within the range $1 < x < 2$. As an approximation, $x = 1$ should be used for random rates estimated from the

count rate at single detector elements, while $x = 2$ should be used for random rates estimated using a delayed time window [96]. The NECR emerges from the signal-noise ratio of the recorded PET data. As

$$\mathrm{SNR}_{\mathrm{data}} = \frac{T}{\sigma_T}, \tag{3.2}$$

and following that $T = P - S - R$ (P denoting 'prompt' coincidences, or all measured coincidences within the set coincidence window before corrections are applied) with P and R Poisson-distributed and uncorrelated, it can be shown that

$$\mathrm{SNR}^2_{\mathrm{data}} = \left[\frac{T^2}{T + S + x \cdot R}\right] \times \Delta t = \mathrm{NECR} \times \Delta t, \tag{3.3}$$

where Δt is the duration of the acquisition [97].

The NECR exhibits a characteristic curve shape when plotted against activity present in the field of view. NECR increases with count rate, but reaches a peak value before decreasing as activity increases further. Time resolution constraints of the scanner are responsible for this effect. The coincidence window, system time resolution and detector dead time effects result in the continued increase of random coincidences where true coincidences cannot be resolved[13]. The scatter fraction, $(S/(S+T))$, theoretically does not change as activity is increased. The curve shape is dependent on the geometry of the activity distribution in the field of view. In the NEMA NU-2 protocols, a set of standards is laid out for usage of the NECR as a scanner performance metric. A standardised phantom (a 20 cm diameter and 70 cm length cylinder with an off-centred line source) is recommended for this purpose; this is separate to the NEMA Image Quality phantom which is used in this work. When the NECR is used in this work it is therefore not equivalent to the NECR values quoted in scanner performance reviews, but is more akin to its use in a patient-specific context[14].

The relationship between $\mathrm{SNR}_{\mathrm{data}}$ and $\mathrm{SNR}_{\mathrm{image}}$ is complicated by image re-

[13]For the Siemens Biograph mCT the coincidence window is 4.1 ns, while the system time resolution is around 540 ps

[14]Studies have developed patient-specific noise equivalent counts by adjusting the standard definition of NECR to include geometric effects of different patients, with the aim of improving dose measurements to the individual [97, 98].

construction and data correction methods used. Despite this, image uniformity metrics have been found to correlate well with NECR for both filtered-back projection and OSEM, although some evidence to the contrary has been presented for the latter in certain conditions [99]. To illustrate, Figure 3.1 shows an image's *percentage integral uniformity*,

$$\%IU = 100 \times \frac{X_{max} - X_{min}}{X_{max} + X_{min}}, \tag{3.4}$$

for image voxel values X, against the phantom activity. X_{max} and X_{min} represent the maximum and minimum voxel value in the ROI. The nadir of this curve occurs at a comparable activity as the NECR-activity curve peak.

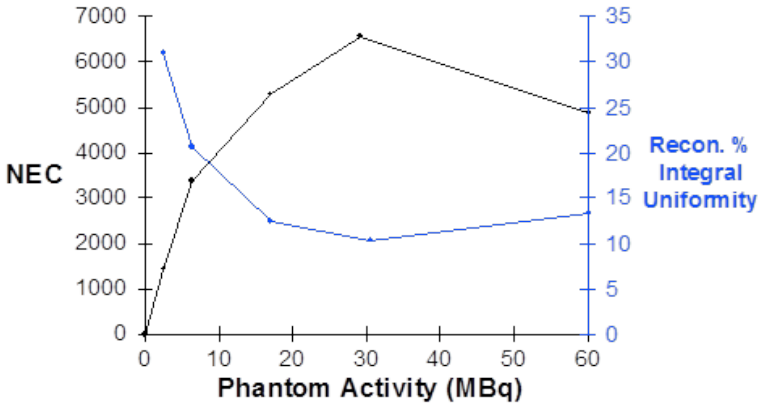


Figure 3.1: The NECR vs. activity curve (black) superimposed by the image's integral uniformity vs. activity curve (blue). Figure courtesy of P. Julyan et al., The Christie NHS Foundation Trust.

This work considered whether image heterogeneity metrics correlated with NECR. The characteristic nature of the NECR's behaviour could promote a discussion about the influence of noise on a radiomics texture-based feature's value in PET. Successful findings here would supplement work done in assessing robustness by testing for robustness to different parameters. If the findings are significant, they could also be used to suggest a framework by which to estimate uncertainties in these features, a mechanism that is yet to be implemented successfully anywhere in the field.

3.3 Experimental Setup

The experiment that produced Figure 3.1 scanned a phantom of a known initial activity of ^{18}F over several isotope half lives for a single bed position, separating into frames to give decreasing nominal phantom activity. Each acquisition was then reconstructed into an image and the metrics taken from this. A similar experimental setup was performed at The Christie NHS Foundation Trust. The 'global' nature of the NECR suggested that an initial experiment should be performed analysing a uniform cylindrical phantom, a simple geometry spanning the axial length of the scanner's FoV. Following completion of this pilot study, further phantoms were to be investigated to examine the effect of changing ROI shape and size on the outcome of any discovered correlation. The development and creation of these phantoms is laid out in Chapter 4.

The scanner used in this work is the Siemens Biograph mCT with TrueV, a conventional modern clinical PET-CT scanner, located at The Christie NHS Foundation Trust. Further specifications are listed in Table 3.1.

Properties	
Detector element dimension	4 x 4 x 20 mm
Detector elements per block	169 (13 x 13 array)
Blocks per ring	48
Number of element rings	52 (13 x 4 block rings)
Detector ring diameter	842 mm
Transaxial FOV	700 mm
Axial FOV	216 mm
Plane spacing	2 mm
Image planes	109
Coincidence window	4.1 ns
ToF Resolution	540 ps
Energy resolution (FWHM)	12%
Energy window	435 - 650 keV

Table 3.1: Selected key properties of the Biograph mCT scanner. Values taken from the Biograph mCT Specification Sheet [100].

Chapter 4

Creating Custom Phantoms

In this chapter, the process of creating a 3D-printed phantom insert will be discussed. After elaborating on the benefits of phantom scans, the following sections detail how identifying and segmenting patient geometries can be developed into a 3D model using 3D design software and fused-deposition modelling printing techniques. The merits of this printing methodology will be discussed. The chapter will summarise by showing the phantom inserts that were created for this work, and establishing the details of the scans that took place to provide the data to be analysed in later chapters.

4.1 Phantom PET Scans

A phantom scan is a powerful tool for nuclear medical imaging, as it can provide recorded data and images of previously known radioactivity distributions. The concept of internal medical imaging in the clinic implies a lack of knowledge about the radioactivity distribution being measured, meaning that improvements to the imaging equipment and techniques can be done with appropriate use of phantom scans. Recognisable industry-standard phantoms are vital for harmonisable, reproducible work; producing results that could be verified in similar conditions at different sites. Two such phantoms were used in this work; a cylindrical phantom and the NEMA Image Quality phantom.

The cylindrical phantom was an ideal candidate for a homogeneous volume.

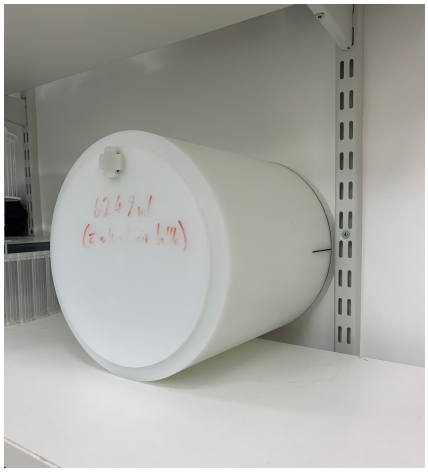


Figure 4.1: A photograph of the cylindrical phantom.

This was a perspex cylinder, 20 cm in diameter and 20 cm in length with a volume of 6.3 l, and is pictured in Figure 4.1. The axial FoV of the Siemens Biograph mCT was spanned by this phantom, and as such the acquisition could be taken in single bed position mode. This was important for reducing the complexity of the 12 hour experiment.

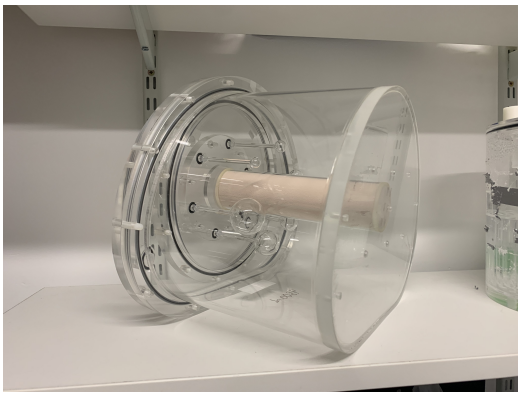


Figure 4.2: A photograph of the NEMA IQ phantom.

The NEMA Image Quality phantom, henceforth known as the NEMA or NEMA IQ phantom, is shown in Figure 4.2. It has a more complex geometry, designed to measure how well different scanner systems are able to acquire and reconstruct images on progressively smaller objects [101]. It is 20 cm long with a total volume of around 10 l. Its use as an imaging standard began around 2001, as more clinically-appropriate phantoms were desired for testing aspects of image

quality (such as contrast recovery factors, visibility of small lesions etc) in 3D PET scanners [102]. The phantom consists of a quasi-cylindrical shell designed to approximate the human torso, with six differently-sized spheres suspended inside by filling tubes. The spheres are sized to approximate a range of clinically relevant volumes. There is an additional optional 'lung insert', a perspex cylinder with polystyrene filling which can be fitted in a slot between the sphere inserts. A schematic for the phantom can be seen in Figure 4.3.

4.2 3D Printing Custom Phantoms

The standard phantoms detailed previously provide excellent examples of large homogeneous volumes and small regular geometrical shapes. These examples are without doubt useful, but having established that complex radiomics metrics are highly dependent on the shape and volume of the ROI, it is of limited use to study geometries which are extremely unlikely to manifest in a patient's scan. In order to create a more realistic example dataset, it was important to create a set of custom models which could mimic the size and shapes of tumours seen in real patient data.

A custom phantom dataset was to be subject to the same experimental conditions as the other two phantoms. A 12 hour scan again required a single bed position, and so to minimise wastage the NEMA phantom body was repurposed and a new 'baseplate' created. This baseplate, the schematic of which can be seen in Figure 4.4, featured supports for up to four new inserts. These new tumour inserts were to be 3D printed and adapted to be filled with radioactivity, and suspended in the NEMA phantom body (also filled with a solution of a known 'background' activity).

4.2.1 Isolating Geometry From Patient Data

Anonymised PET image data was obtained for 40 patients through The Christie NHS Foundation Trust & The PET-CT Academy. Of these, eight had colorectal tumours, 15 had lung tumours and 17 had oesophagael tumours. This range of cases was chosen to provide a range of different sizes and shapes of tumour, due

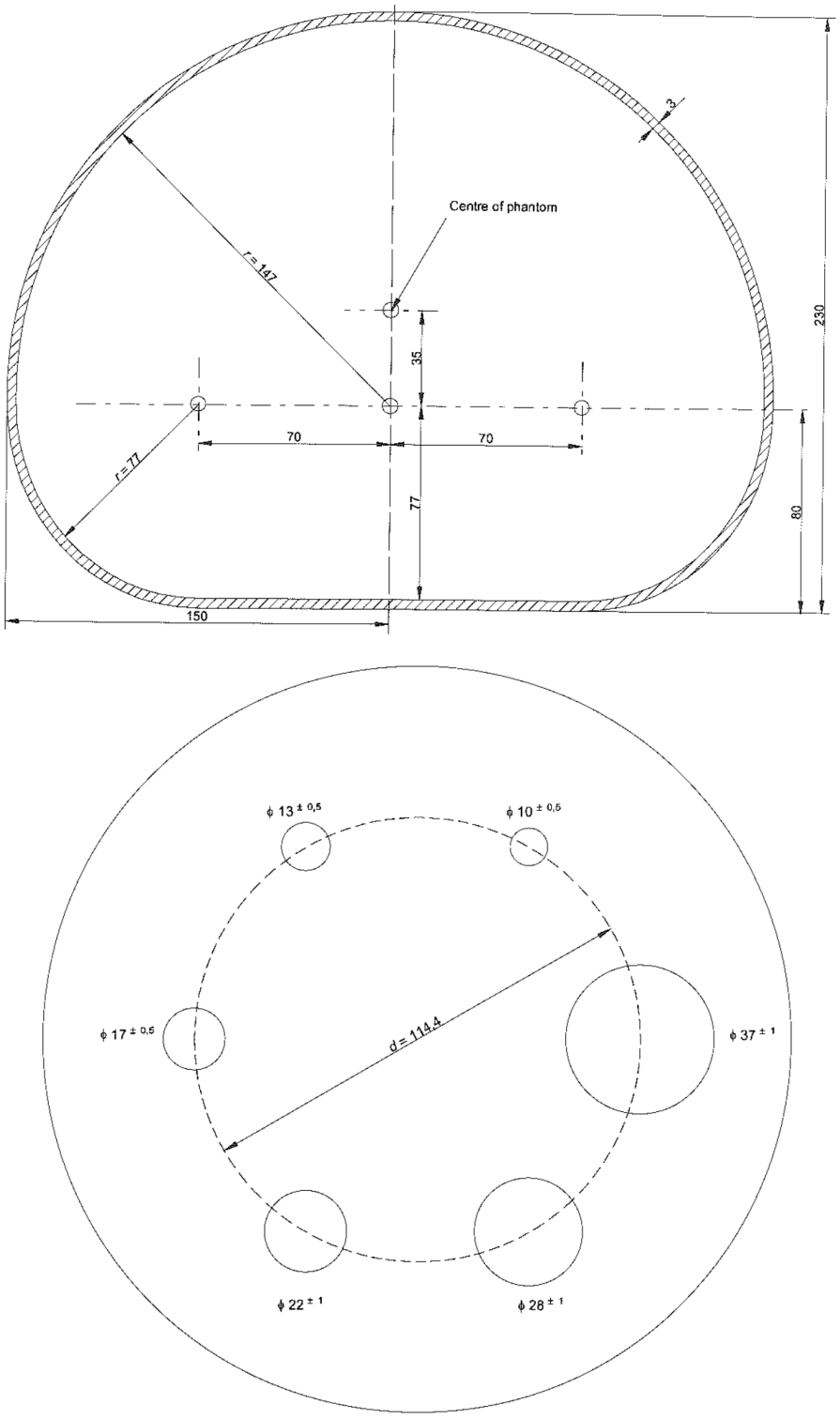


Figure 4.3: Schematic drawings for the NEMA Image Quality phantom body (left) and spheres (right) [101]. The dimensions shown are in millimetres.

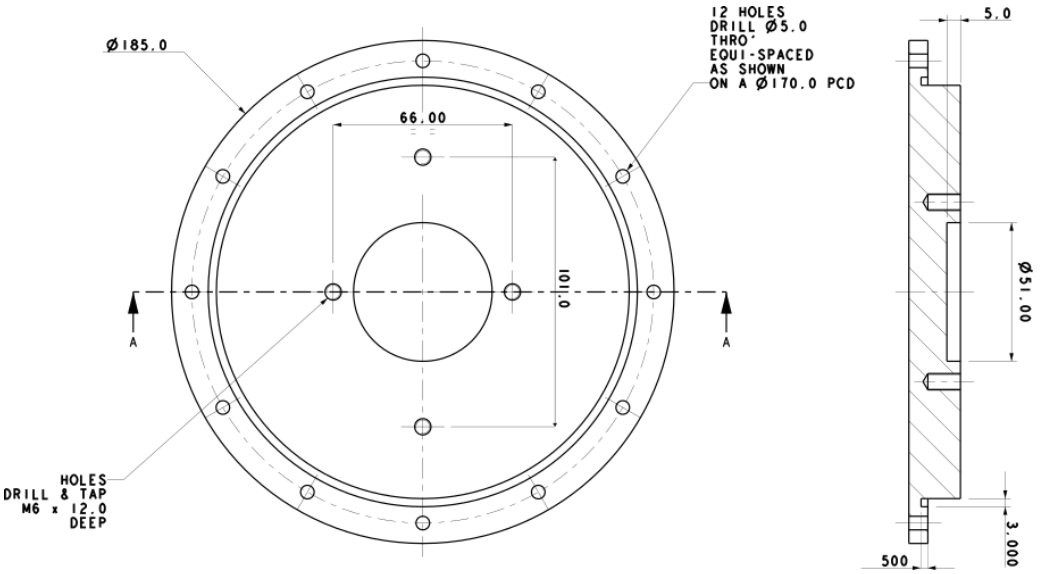


Figure 4.4: Schematic of the redesigned NEMA phantom baseplate for custom phantom inserts. The displayed measurements are in millimetres.

to characteristic differences typically exhibited between each disease type. One oesophagael case and three lung tumour cases were discarded due to restrictively small tumour volumes. The remaining 36 patients' tumour volumes were then considered as potential candidates for the custom phantom insert geometries.

The geometries were created by isolating relevant tumour material in a patient PET scan using LifeX. This was achieved by setting a broad spherical ROI around the region in question and thresholding down to 40 % of the maximum voxel value. These ROIs were exported as binary 'mask' images, which could then be opened and interpreted as 3D STL surface objects using ImageJ.

Section 2.4 explained that CT images have higher resolutions than PET images, with a CT voxel sized around $1 \times 1 \times 3\ \text{mm}^3$ compared to $4 \times 4 \times 3\ \text{mm}^3$ the PET voxel. Such a fine resolution for the CT data would appear to promote its suitability for isolating a patient's tumour geometry, but there were concerns over the practicality of producing 3D prints with the high frequency components that would be obtained by thresholding an ROI from a CT scan. The motivation for the experiment required, at this stage, objects of comparable volumes and shapes to tumours which may be obtained in a PET environment, and so any potential loss of resolution was not deemed significant. Consequently, the geometry

isolation was performed on the patient PET images.

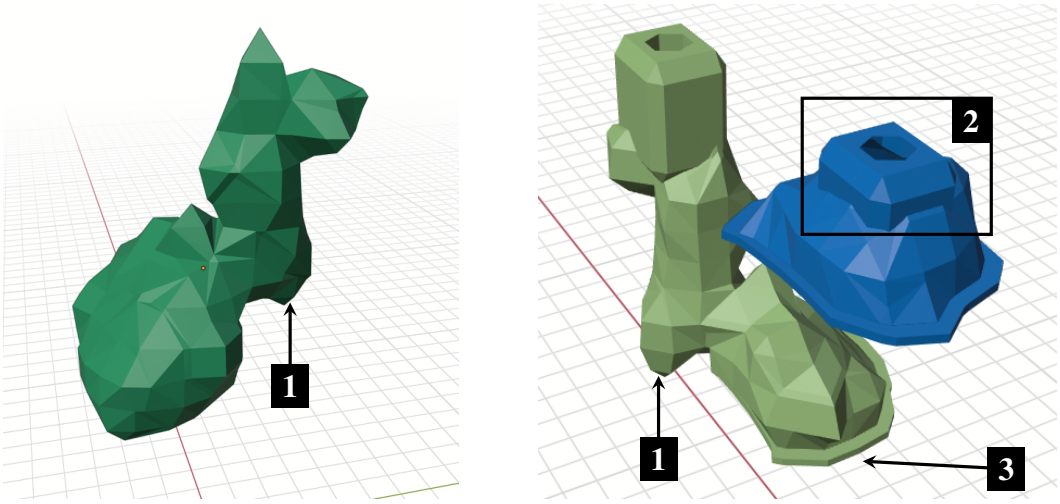
4.2.2 Adapting Regions in 3D Design Software

It was required to manipulate the obtained tumour STLs into a suitable form for printing, filling and mounting within the phantom. The STLs were modified and adapted using Blender, an open-source 3D design software [103]. There was no uniform method which performed equally well for every tumour geometry. In many cases, prototypes were created, a print attempted and subsequently adapted after print success or failure.

The tumour inserts were to be attached to the baseplate using M6 threaded supports and required some mechanism through which to fill them. The surfaces of the extracted STLs were blocky and angular, and therefore smooth surfaces had to be created on diametrically opposing sides of the insert. These surfaces were created using Boolean operations with separately designed 'platform' meshes[15]. However, FDM printers struggle with printing fine threaded structures. While the threaded holes themselves were omitted from the final designs, spaces were left for threaded nylon nuts and spacers to be attached post-print.

It was desired for each of the tumour phantom inserts to be created as one single printed object. This had advantages for some STL files, such as insert T1 (see Table 4.2), which contain more complex internal geometries. However, for ease of printing and removing the PVA support structures, the STLs could be split in two. The two sides of the structure were then joined together post-print. Figure 4.5 shows an example of how this was done, with the left hand side volume being split across a low-dimensionality boundary to create the right hand side STLs. To improve the adhesion, a 5 mm lip was extruded from both parts.

[15] A Boolean operation involves creating new objects based on the intersection or union of two existing ones. Here, separate objects were created for the tumour and for the filling port, and the two were joined in the most efficient way to create a stable platform with minimal adverse affect on the tumour's initial shape.



(a) Before (b) After

Figure 4.5: STL geometry of one lung tumour design viewed in Blender. Left and right show before and after the geometry is split into two, with filling & support structures added. This design was used to create phantom insert *T3*. Annotations: (1) shows the same point in both STLs for alignment purposes; (2) shows a developed filling platform; (3) shows the extruded lip.

4.2.3 Creating Phantom Inserts with 3D Printers

The method used for 3D printing in this work is *fixed filament fabrication* (FFF) or *fused deposition modelling* (FDM). These two equivalent terms refer to the process of extruding plastic filament through a high-temperature nozzle, forming the layers which construct the printed object. The printer used to create the phantom inserts was an Ultimaker S5, pictured in Figure 4.6, a dual extrusion system with a humidity-controlled cabinet for filament storage.

Due to the extrusion process there are a limited set of plastics which are suitable for FDM printing. The most widely used filament material, and the material used in this work, is polylactic acid, or PLA. Its low print temperature and low cost relative to the alternatives (listed in Table 4.1) have helped to create its status as the default material for prototyping 3D printed works. 'Dual extrusion' refers to the printer's capability to handle two filaments simultaneously[16], enabling a support structure to be constructed along with the main build. PVA is primarily used as the support structure material for PLA prints; PVA can be extruded at a similar temperature to PLA and is water soluble, meaning that support structures

[16]Only one filament is extruded onto the print bed at any one time.

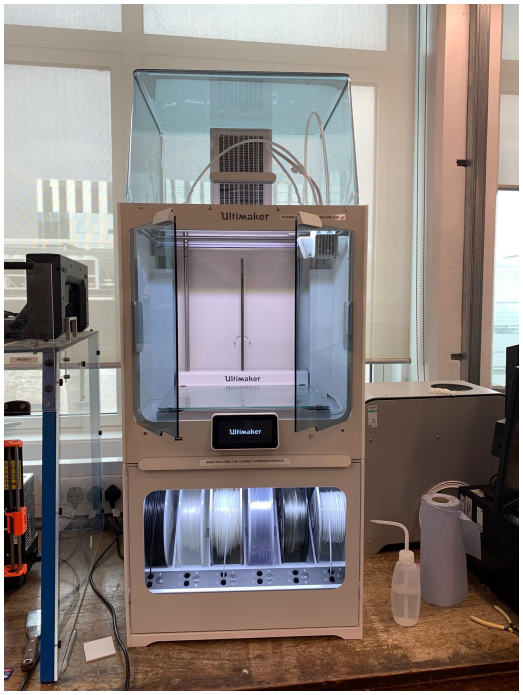


Figure 4.6: A photograph of the Ultimaker S5 3D printer used to produce the custom phantom inserts in this work. The printer comprises of two sections: the material bay underneath housing the plastic filament reels, and above (with doors open) the printing chamber. Objects are printed onto the glass bed in the chamber, which is raised and lowered in relation to the nozzles which are attached to a 2D frame at the top of the chamber.

can be printed quickly and easily alongside the main print and dissolved after the print is complete.

4.2.4 Finishing the Insert Prints

The nylon threaded nuts and spacers were affixed to the hexagonal spaces left on the prints using Araldite epoxy adhesive. The same adhesive was also used to join the two halves of the split geometry designs. The printed PLA structures were not watertight. To allow for fine detailing in the printing, the wall thickness for the prints was set to 1 mm. The FDM process results in relatively weak bonds between PLA layers, and the complicated and blocky structure of the insert geometries resulted in weak points where leakages could occur. There was no wall thickness that would have resulted in watertight prints while preserving the volume and fine structural details of the tumour insert. Two methods are

Material	**Melting Temp., °C**	**Cost /** 750 g	**Other Properties [104]**
PLA (polylactic acid)	145 – 160	£32.50	Widely available, strong, and comparatively easily-stored material. Filament prone to snapping, and has low heat resistance.
ABS (acrylonitrile butadiene styrene)	225 – 245	£37.50	High heat resistance, although results prone to shrinkage post-print. Releases harmful fumes during prints.
CPE (co-polyester)	> 100	£40.50	Negligible warping and smooth prints, but material is prone to wear.
Nylon	185 – 195	£53.50	Tough and flexible material. High maintenance storage (e.g. airtight, low humidity) required.
PVA (Polyvinyl Alcohol)	163	£83.50	Used almost exclusively as a support material. Dissolves in water at room temperature.

Table 4.1: A table arguing the advantages and disadvantages of the most common FDM filament materials. Cost is listed for a 750 g reel produced by Ultimaker available from RS Components (`uk.rs-online.com`) as of 27th June 2022. It should be noted that nozzle temperatures for printing are higher than the melting temperatures listed; for PLA, the nozzle should be set to $\sim 210\,^{\circ}$C.

commonly used to seal prints; gently melting the layers together with a heat gun, or coating in an epoxy resin. The resin sealant method was chosen for this application due to concerns over the wall size of the PLA being able to maintain strength and structure without reinforcement. The sealant comprised of bisphenol A epichlorohydrin polymer resin and a 2-ethylhexyl glycidyl ether 'hardener' mixed in a 3:1 ratio. This was then painted onto the surface of the print in a thin even layer in two coats, applied 24 hours apart. Once sealed, a leak test for the phantom inserts was required. No activity could be allowed to leak out of the insert, for safety whilst filling and for the purpose of experimental rigour. The completed and sealed phantom inserts were filled with coloured water and left for three days (see Figure 4.7). The prints were inverted on the second day; this was to ensure that the prints could be guaranteed watertight for at least twice the duration of the desired 12-hour scan series. A phantom insert was deemed ready if no colour was present on the surrounding paper, otherwise a second coat of epoxy resin sealant was applied.

4.2.5 Heterogeneous Phantom Design Principles

One aspect of phantom design that was unable to be completed was in-built heterogeneity. One example of how this could be achieved can be found in [95]. The authors achieved this in two ways, by bisecting the 3D design to create two independent sections, and by creating a 'necrotic core' by creating an empty section inside the 3D design. The first method could enable filling with two different concentrations of radioisotope; the two sections would be stuck or nested together with different filling ports. One ROI would then encapsulate two different base concentrations. The necrotic core model would enable background activity to be 'seen' within the confines of the ROI, achieving the same effect as the initial model.

4.3 Phantom Insert Selection

Four phantom inserts were selected for scanning; prefixed with a *T* label, these are listed in Table 4.2. In order to provide a variety of shape and size, and with the NEMA Image Quality phantom forming part of the dataset, tumour geometries were selected that were larger than the highest-volume NEMA IQ sphere. This meant selecting from among the largest subset of the extracted data. The four geometries were derived from lung and oesophageal tumour patients. Figure 4.8 shows two of these inserts (T1 and T3) attached to the custom baseplate.

Figure 4.7: The four selected phantom inserts undergoing leak testing. L-R: T4, T2, T3, T1.

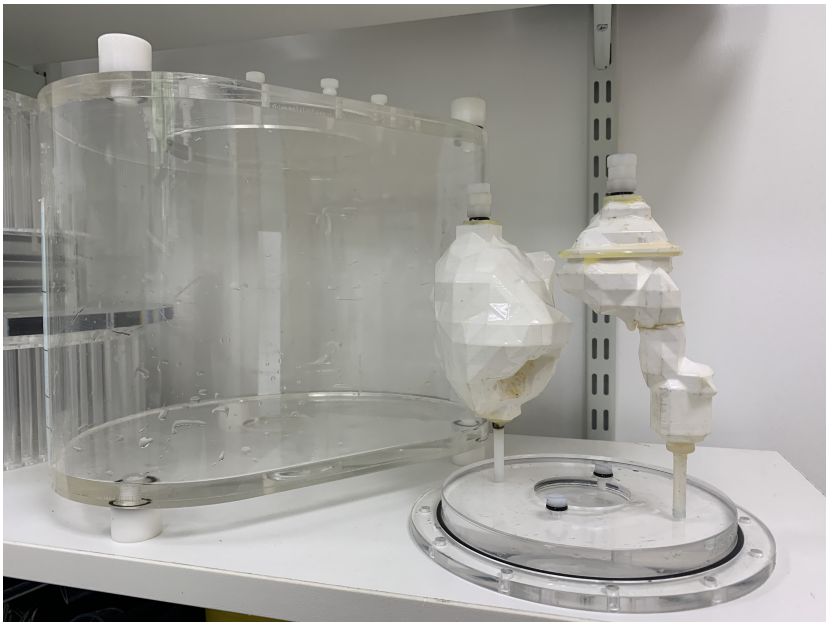


Figure 4.8: A photograph of T1 and T3 attached to the custom baseplate for the NEMA IQ phantom, itself pictured to the left.

4.4 Phantom Scanning

The phantoms were filled with ^{18}F in the form of FDG acquired commercially and diluted on site. The systematic uncertainty[17] in true activity measurement was related to the calibration of the field instruments used, with an upper estimate at the $\pm 5\%$ level [105]. The activity in each phantom scan is listed in Table 4.3.

Insert Number	Insert Volume, ml
T1	229 ± 2
T2	124 ± 2
T3	71 ± 2
T4	41 ± 2

Table 4.2: Volumes of the four selected tumour phantom inserts. Volumes were calculated from the mass of water used to completely fill the phantom insert without air gaps; such air gaps were unable to be completely eliminated when filling with radioactivity.

[17] Random uncertainty of instrument measurements is monitored by daily on-site Quality Control at The Christie NHS Foundation Trust using a ^{137}Cs source. This is found to be consistently $< 1\%$.

Phantom	Total Activity, MBq	Target		Background		Activity Conc. Ratio (BG:T)
		A. Conc., kBq/ml	Volume, ml	A. Conc., kBq/ml	Volume, ml	
Cylinder	612 ± 31	-	-	97 ± 5	6280 ± 20	-
NEMA IQ	553 ± 28	336 ± 20	48 ± 7	534 ± 27	9770 ± 20	1:6.15
T1+T3	741 ± 37	141 ± 7	302 ± 4	600 ± 30	9470 ± 20	1:7.34
T2+T4	728 ± 36	61 ± 3	167 ± 4	657 ± 33	9600 ± 20	1:5.33
Only3	339 ± 17	4690 ± 270	72 ± 2	-	-	-

Table 4.3: Table containing the activity at the start of scan for all phantom arrangements, along with the activity concentrations and volumes for the target and background regions.

Filled phantoms were centred on the bed of the Siemens Biograph mCT TrueV using the in-built laser alignment. All patient protocols begin with a planar X-ray tomograph and a CT image acquisition; this was followed for all phantoms, in order to centre the phantom in the PET axial FoV and to create the attenuation maps. Figure 4.9 shows the *T2+T4* phantom arrangement on the scanner bed before undergoing acquisition. The protocol used for the Cylinder and Tumour Phantom scans consisted of consecutive 5 and 25 minute frames, repeated 24 times for 12 hours scan time. A slightly altered protocol was created to test the NEMA Image Quality phantom. This adjustment was done to sample as frequently as possible around the peak of the NECR curve. The NEMA IQ protocol consisted of 24 five minute frames interspersed with 24 gap ('G-') frames of variable length. These are listed in Table 4.4.

Gap Frames	Frame Duration, minutes
G1-8	5
G9-11	10
G12-16	30
G17	40
G18-24	30

Table 4.4: Table listing the acquisition duration of the G-labelled frames for the NEMA IQ phantom scan series.

The resultant dataset consisted of 32 five minute frames, with the longer frames differing in duration. This was not explored further in subsequent scans to preserve the two definitive datasets ('long' 25 minute and 'short' 5 minute scans).

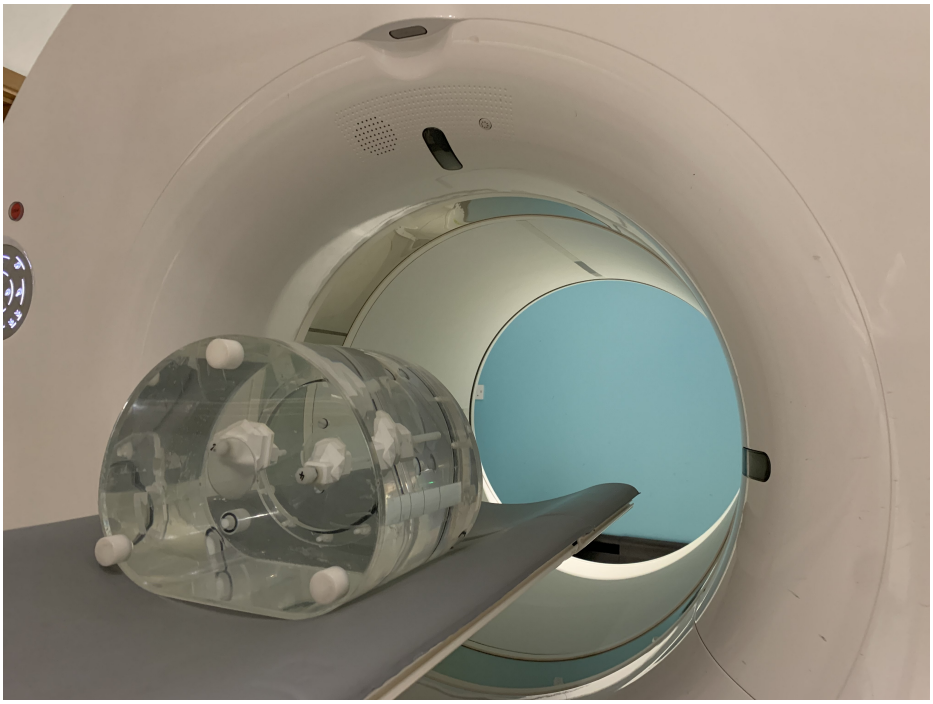
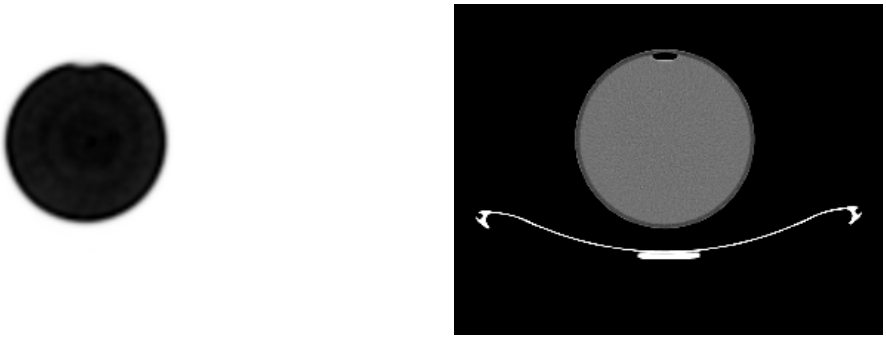


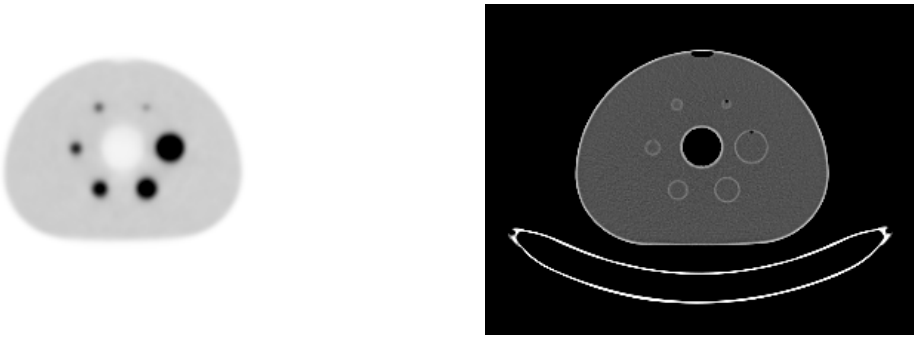
Figure 4.9: NEMA Phantom with custom inserts T2 & T4, filled and placed on the scanner bed.

4.5 Image Reconstruction

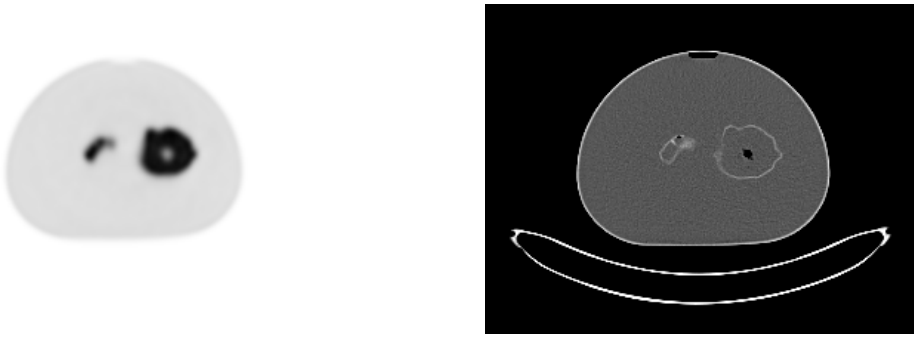
A clinically appropriate image reconstruction protocol was used to reconstruct all data used in this analysis. The reconstruction was performed on the scanner's associated computer system using the parameters listed in Table 4.5. The terminology and nomenclature used, such as UHD for the full point-spread function modelling capability, is specific to the Siemens architecture. Example PET and CT images from the first four phantom scan series can be seen in Figure 4.10.



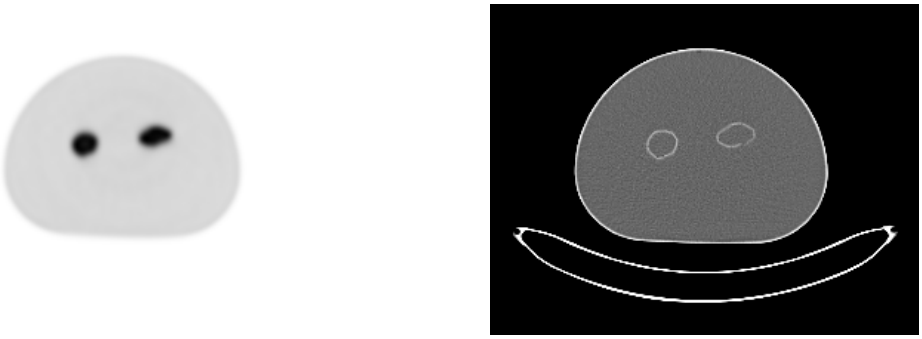
(a) A PET (left) and CT (right) image of the cylinder phantom.



(b) A PET (left) and CT (right) image of the NEMA IQ phantom.



(c) A PET (left) and CT (right) image of the T1+T3 phantom setup.



(d) A PET (left) and CT (right) image of the T2+T4 phantom setup.

Figure 4.10: Example PET and CT image slices from the four phantoms used in this work. From top-bottom: Cylinder, NEMA IQ, *T1+T3*, *T2+T4*.

Parameter	Value
Algorithm	OSEM + ToF + PSF (UHD)
Iterations	2
Subsets	21
Matrix Size	256 x 256 x 74
Post-smoothing filter	3D 5 mm Gaussian
Other notes	CT-based Bed Removal No zoom (set to 1)

Table 4.5: Details for the image reconstruction protocol used in this work. The algorithm used is known as UHD in Siemens nomenclature.

Chapter 5

Investigation of Radiomics Features with NECR

The following chapter details the investigation into the robustness of radiomics texture features to the global radioactivity level. The opening sections detail how count rate statistics were extracted from the raw data, along with a discussion on potential sources of uncertainty and an in-depth look at the reported scatter fractions. The image statistics are then analysed from the four main datasets in order; the cylinder, the NEMA IQ phantom, *T1+T3* and *T2+T4*[18]. Correlation with NECR is considered as a means by which these features can be more appropriately quoted for clinical use, and the difference between the two sub-dataset acquisition times (25 and 5 minutes) is evaluated.

5.1 NECR Measurements

5.1.1 Extracting the Count Data

The images obtained from the scanner are in DICOM format; a medical image standard whereby the data is exported in a series of files, generally corresponding with an image *slice* (usually axial) and containing the pertinent metadata. The

[18]It should be noted that *Only3* is included in Figures 5.3 and 5.5, yet is only introduced in the following chapter. Its inclusion is for illustrative purposes.

header contains all of the image information, each statistic addressable with a tag of the format (xxxx,xxxx) [106]. These tags are standardised and can be found here [107]. The scatter fraction, as estimated by the Siemens scanner software during scatter correction, can be accessed in the DICOM header with the tag (0054,1323).

The count rate information was not found in the image DICOM headers, but was listed in the raw data sinogram headers. This information was stored by the scanner, but had to be restructured using Siemens' offline image reconstruction platform, e7 tools. The sinogram header file contained information for the total number of prompt and random counts collected over the whole frame. Prompts, in this instance, refers to every collected coincidence, such that

$$P = T + S + R, \tag{5.1}$$

where P, T, S and R are the prompt, true, scattered and random coincidences. The randoms estimate is already given in the sinogram header, calculated using the delayed window method with the data streams directly from the scanner. Trues and scatters are not listed in the sinogram headers, and instead these are listed as all true *timing* coincidences, 'net trues', where

$$\text{net trues} = P - R = T + S. \tag{5.2}$$

The scatter fraction listed in the DICOM headers, f_S, refers to the fraction of the net trues which are designated as due to scattered coincidences,

$$S = f_S \cdot (P - R). \tag{5.3}$$

For this work, the quantities T, S and R were divided by the duration of each frame and used as rates from this point onwards.

The NECR was established using

$$\text{NECR} = \frac{T^2}{T + S + 2R}, \tag{5.4}$$

with the method-dependent 'x' factor equal to 2 corresponding to the advised

usage with a delayed window randoms estimation.

The NECR for the cylindrical phantom, plotted against the activity present in said phantom, is shown in Figure 5.1. The red curve represents a quartic fit to the data, chosen as the polynomial with the lowest order satisfying a minimum χ^2_ν requirement. The blue dotted line and shaded area correspond to the peak value, $\mathrm{NECR}_{\mathrm{peak}}$, and the corresponding activity at which this occurs. This value, and the associated uncertainty, was calculated using 10^6 iterations of randomly selected quartic fits generated from the original fit's covariance matrix.

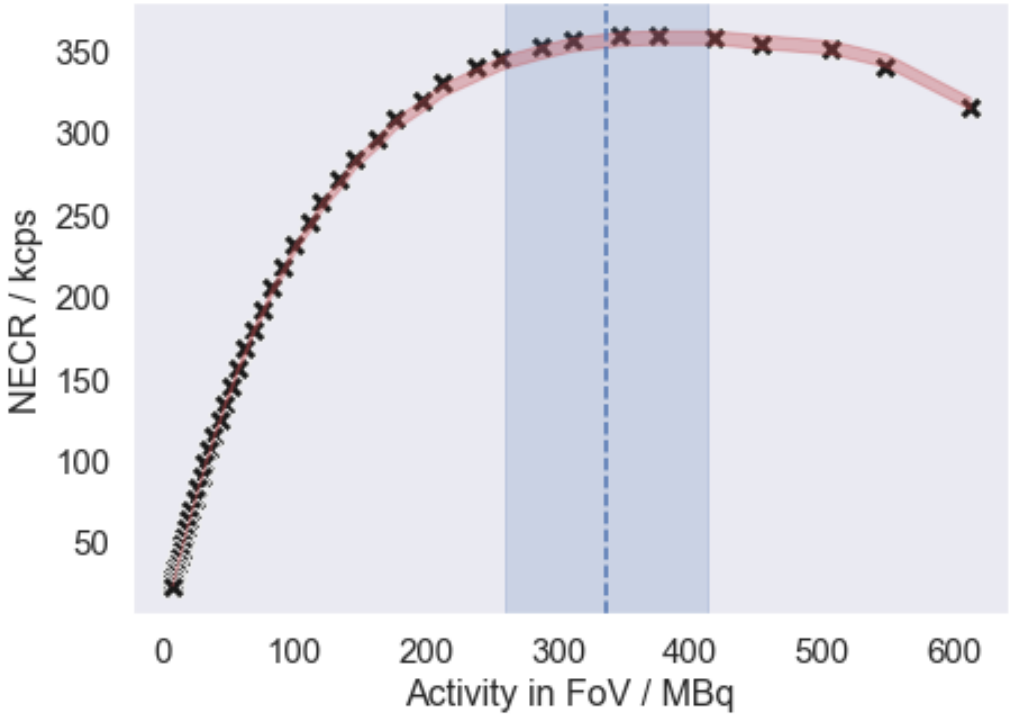


Figure 5.1: Scatterplot showing the NECR from all cylinder data. The blue dotted line represents the position of the peak NECR evaluated from quartic fitting with the shaded region representing the uncertainty. The red shaded region represents uncertainty in NECR given by the reported scatter fraction. The statistical uncertainty on any given measurement of NECR is negligible.

The count data, separated into trues, randoms and scatters can be found in Figure 5.2. As expected, the trues and scatters appear to increase approximately in a linear fashion, while randoms appear to increase quadratically over the same activity domain. Fundamentally this agrees with what we know of a random coincidence; the randoms rate can be said to be proportional to the single photon rate,

as each random coincidence requires a single photon from two distinct emissions. With each emission resulting in 2 photons, and therefore the potential to cause 2 randoms, the randoms can be modelled quadratically with FoV activity.

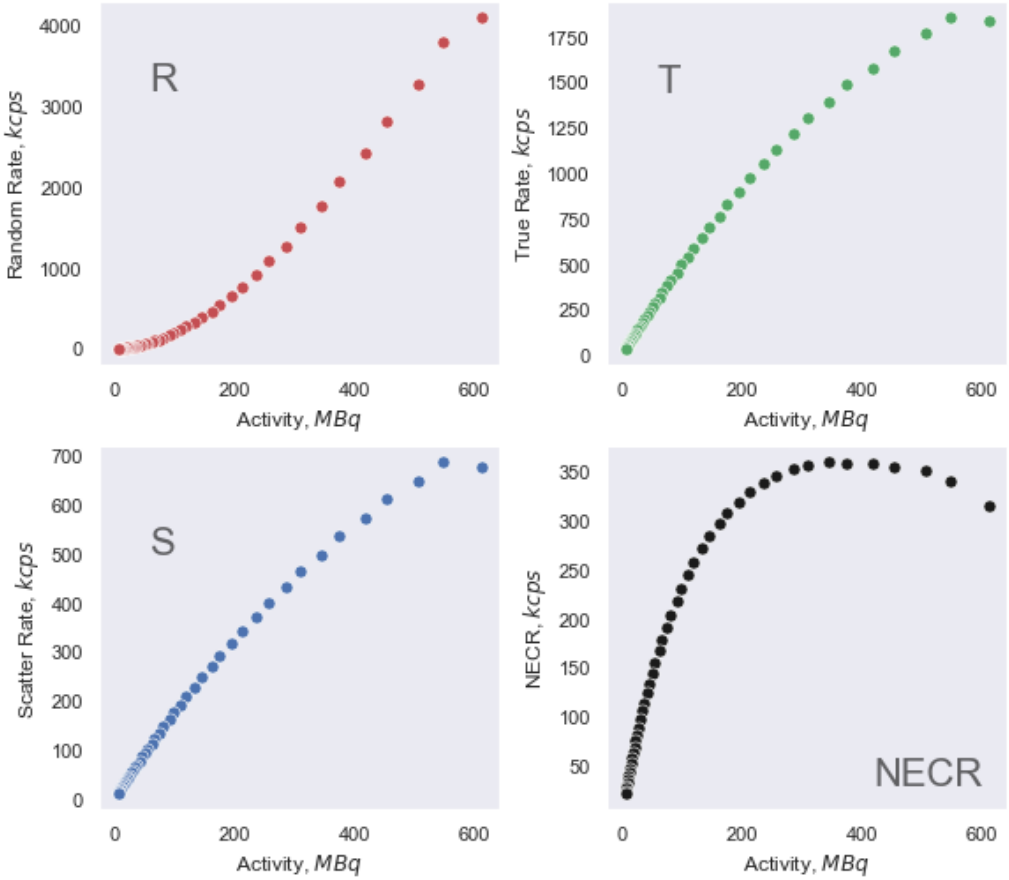


Figure 5.2: The count statistics for the cylinder data plotted against FoV activity level. Clockwise from top left: random rate R, true rate T, scatter rate S and the NECR.

The NECR plotted for all phantoms against the equivalent total activity within that phantom per frame can be found in Figure 5.3. This figure illustrates well the geometric dependence of the phantom and activity distribution on the NECR; all three phantoms based on the NEMA IQ phantom (with hot background) share a very similar NECR peak. Details of the peak NECR for each phantom can be found in Table 5.1. Example PET images from *T1+T3* can be seen in Figure 5.4, one early frame (high activity and high NECR) and one of the last (low activity, low NECR). The images, which are both SUV-normalised and use the

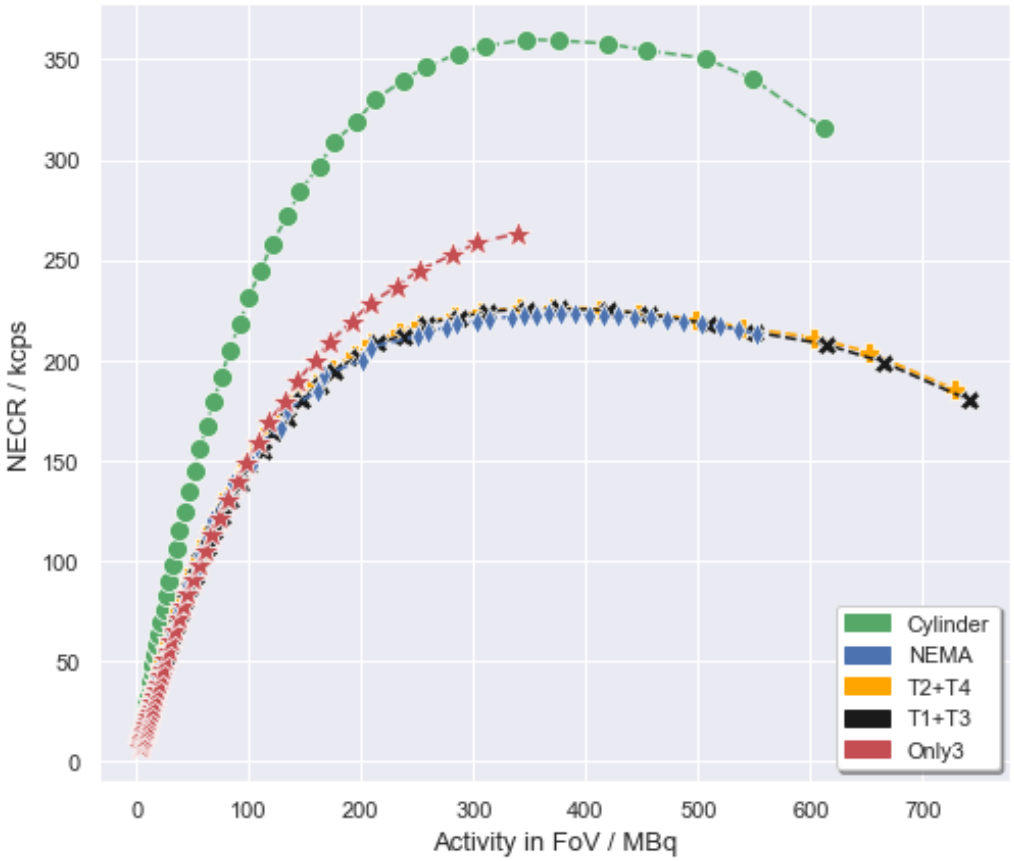


Figure 5.3: A scatter plot showing the NECR for all scans performed in this thesis. *Only3* is included here for illustrative purposes but is not explored in depth until Chapter 6. Likewise uncertainty boundaries as in Figure 5.1 are not shown for illustrative purposes and functional form differences in reported scatter fraction (see Figure 5.5). Connecting dotted lines are shown for visual effect.

same colour scale, show the impact of the lower count statistics on the resultant visual heterogeneity.

Experiment	$A(\mathbf{NECR_{peak}})$**, MBq**
Cylinder	334 ± 71
NEMA IQ	319 ± 121
T1+3	337 ± 91
T2+4	318 ± 82

Table 5.1: Table detailing the activity at which peak NECR is reached for the four main phantom acquisitions.

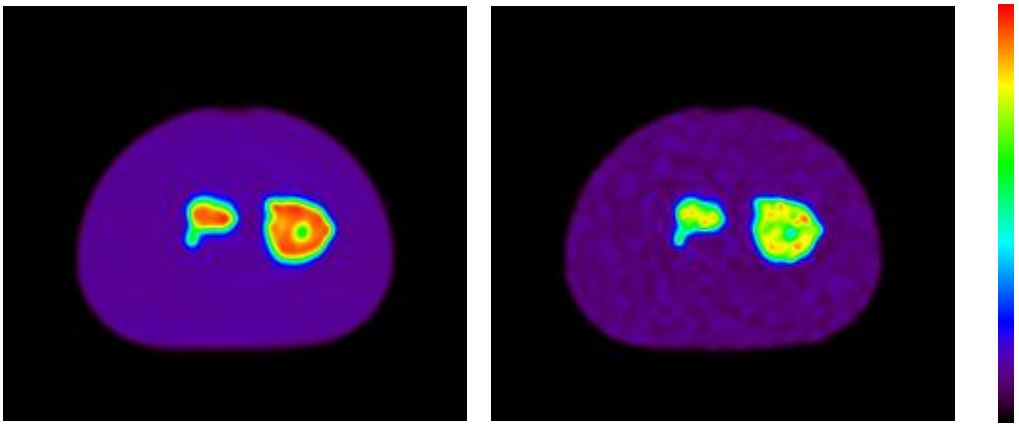


Figure 5.4: Slices from PET frames 2 (left) and 47 (right) from *T1+T3*. The images are SUV-normalised and the colour scale is equivalent in both images, demonstrating the increased visual heterogeneity due to noise in lower NECR scans.

5.1.2 Uncertainty on NECR, and the Scatter Fraction

The statistical uncertainty on any given measurement of NECR is negligible. Consider, for example, the lowest activity cylinder acquisition. The sheer volume of counts reduces the Poisson-approximated statistical uncertainty (σ) on each to the 10^{-4} level. Examples are detailed in Table 5.2. However, the source of error in our count information cannot be purely statistical. One consideration made was that of background radiation. A blank scan with no activity in the FoV, with a duration of 5 minutes, was taken in the Siemens Biograph mCT TrueV; the

Count Type	Measured Counts	σ	Percentage Uncertainty
Prompts	81227490	9013	0.011 %
Randoms	5649236	2377	0.042 %
Trues	54370165	7374	0.014 %
Scatters	21208089	4605	0.022 %

Table 5.2: Statistical uncertainty from the measured counts for the lowest activity (7.03 MBq) cylinder acquisition. Other information provided: scatter fraction 0.280611, net trues 75578254. σ is the Poisson square-root uncertainty of the measured counts.

corresponding count statistics are listed in Table 5.3. The background counts are caused by the presence of ^{176}Lu in the LSO crystals of the detector. The 980 ± 2 prompt coincidences per second of background forms a 1.8 % contribution to our lowest rate of prompts measured and 0.015 % to the highest rate of prompts measured in the cylinder dataset.

Prompts	294142
Delayed	291227
Net Trues	2915
Scatter Fraction	0.0466323
Prompt Rate	980.5 cps
Random Rate	970.8 cps
True Rate	9.3 cps
Scatter Rate	0.5 cps

Table 5.3: Count information extracted for a 5 minute blank scan performed on the Siemens Biograph mCT. Recorded counts are due to radioisotopes of lutetium in the detector crystals.

Equation 5.4 can be simplified to

$$\text{NECR} = (1 - f_S)^2 \frac{(P-R)^2}{P+R} \tag{5.5}$$

utilising Equations 5.2 and 5.3. We have assumed negligible uncertainties on P and R. The scatter fraction, f_S, is theoretically a constant value, depending only on the geometry and physiology of the scan subject, however notable patterns emerged when the scatter fractions for the five phantom datasets were visualised across the scans, shown in Figure 5.5.

The first remark is the positive linear gradient tended to by all datasets. It is known that the scatter estimation algorithm relies on the use of counts detected

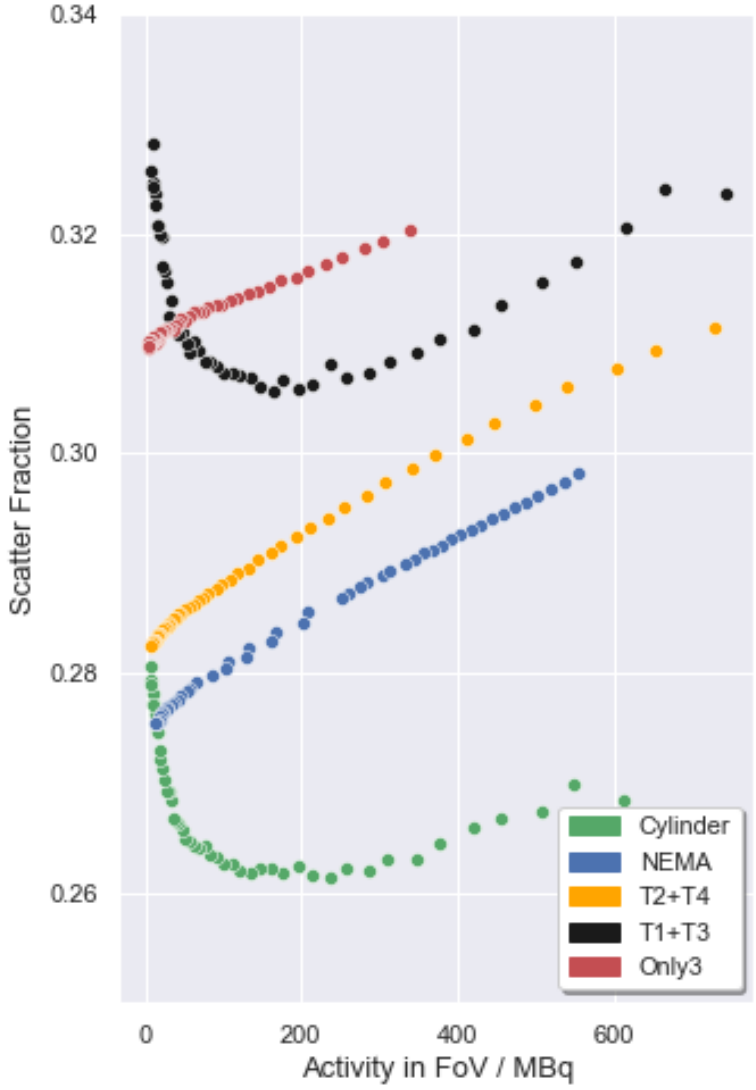


Figure 5.5: The scatter fractions reported by Siemens for all scan data in the experiment. As in Figure 5.3, *Only3* is included for illustrative purposes.

outside the main 'body' of the subject, as determined by the attenuation map (μ-map). While this is a generally successful estimation technique, it is inevitably difficult (or impossible) to distinguish scattered and random coincidences outside the body. As activity in the FoV is increased, and the randoms increase at a faster rate than scatters, a higher proportion of randoms will be included in the scatter estimation algorithm. As such, the reported value should be considered as an 'observed' scatter fraction, as it is unlikely that the value will be equivalent to the true proportion of scattered coincidences.

The second remark is the behaviour of the scatter fractions for the cylinder and T1+T3 phantoms. While at activities $\gtrsim$ 150MBq the familiar positive linear trend is observed, at lower activities the fraction appears almost hyperbolic. After ruling out background counts as a causal factor, the creators of the scatter estimation algorithm were contacted for comment. The most likely theory is that this is an artifact of the time recording in the scan raw data, but at the time of writing no satisfactory solution has been found and is still under investigation.

5.2 NECR & Relationship with Texture Features

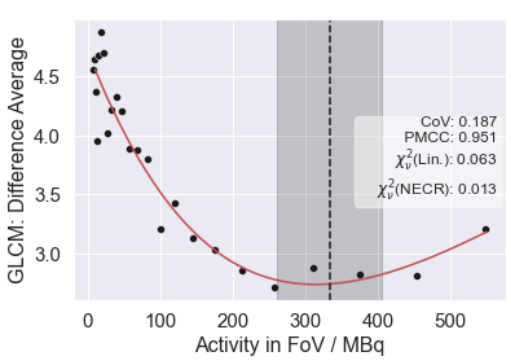
The 25 and 5 minute data, while considered simultaneously for NECR analysis, were to be considered separately; it can be shown that

$$\mathrm{SNR}_{image} \propto \sqrt{S \times A \times T} \tag{5.6}$$

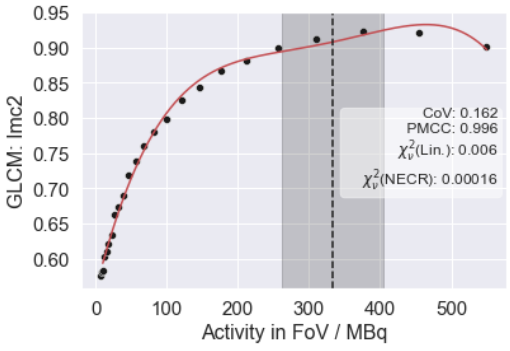
where S, A and T denote the scanner sensitivity, activity scanned, and duration of scan respectively [108].

5.2.1 25 Minute Cylinder Data

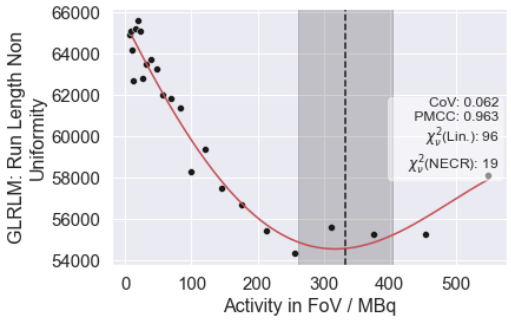
The features from the 25 minute cylinder data were each plotted against activity. For many features, a resemblance to the characteristic NECR-activity curve shape bore out; either directly or in an inverted fashion. Some examples can be found in Figure 5.6.



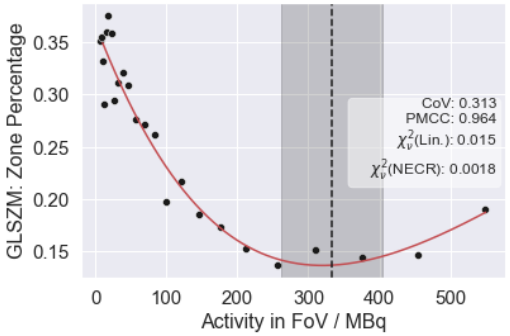
(a) Difference average, from the GLCM



(b) Informational measure of correlation (2), from the GLCM



(c) Run length non-uniformity, from the GLRLM



(d) Zone percentage, from the GLSZM

Figure 5.6: Four plots of texture features against FoV activity level for the 25 minute cylinder data. These are selected examples of features that correlate strongly with NECR. Statistics quoted in the superposed boxes are: CoV, absolute Pearson product moment correlation coefficient with NECR (PMCC), and the reduced chi-squared statistic χ^2_ν for the feature values against NECR and against activity ('linear'). The black dotted line and gray shaded area corresponds to the estimated activity at which the peak of the NECR occurs, and is included as a visual aid. Also included is a quartic fit to the data, illustrated with a red line, to aid visual comparison to those on the NECR curves in Figure 5.1.

An initial investigation calculated coefficients of variation (CoV) and reduced chi-squared statistics (χ^2_ν) for a linear fit; the initial hypothesis proposed that strong linear and/or constant behaviour over the activity domain would represent a robust metric. There were, however, some problems to consider for the use of these values. In this case, the ROI under investigation is large (80416 voxels) and homogeneous, and we expect to see some degree of variation in the feature values across the activity domain. It is unclear however if this noise variation represents a significant change in possible values of the feature when considering all possible distributions. Secondly, as the scale of the features is different on a

case-by-case basis (evidenced by comparing Figure 5.6b to Figure 5.6c), a comparison using these statistics is difficult to perform without prior normalisation. In addition the CoV, while a more useful statistic to measure feature variation, does not account for any functional form of the data.

Pearson product-moment correlation coefficients (PMCCs) were calculated between each textural feature and the NECR. Only the absolute value of this statistic was considered; strong correlations in both positive and negative senses were equally of interest. Many features gave very strong absolute PMCC. A threshold of $|\text{PMCC}| \geq 0.9$ was set, and 32 of the 75 features surpassed this threshold; lowering the threshold to 0.8, 42 of the features pass the threshold. The distribution of |PMCC| is shown in Figure 5.7. These strongly-correlating features form a set to investigate further, as their 'NECR-dependence' suggests that they are heavily influenced by data noise. This is not in itself a problem for their continued use, provided that we can successfully model the NECR.

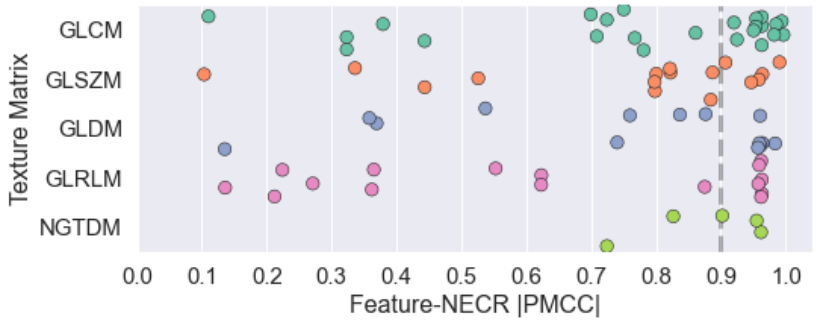


Figure 5.7: A 1D scatterplot showing the NECR |PMCC| of the 75 texture features for the 25 minute cylinder data. Jitter is applied in the vertical direction to enable all data points to be seen.

The peak NECR, $\text{NECR}_{\text{peak}}$, was estimated from quartic fits to the data, occuring at an activity of 334 ± 71 MBq within the field of view for the cylinder data. This activity is much higher than a typical level expected at any time within the FoV for a patient scan; ARSAC advice for ^{18}F-FDG whole body tumour imaging gives a National Diagnostic Reference Level (NDRL) 4.5 MBq/kg for a patient, meaning that a 70 kg patient would be injected with only 315 MBq total[19], which is left to decay for a period prior to imaging [8]. This would be

[19]NDRLs are only guidelines advising safe upper limits, and The Christie NHS Foundation Trust currently use 3.5 MBq/kg for whole body tumour imaging.

then imaged over multiple bed positions, while our data is measured from a single bed position. It is estimated that, even despite activity contributions partially outside the FoV of the scanner, it is unlikely that activity over 100 MBq would be considered typical of a clinical environment. From the NECR behaviour (Figure 5.1) it is deduced that a patient scan is performed with a much lower SNR than here. By way of accounting for this, 'compensation' or 'correction' factors were calculated for the strongest-correlating features. An illustration of this process is shown in Figure 5.8. These factors would in theory enable a clinician to estimate the 'true' value a feature could be expected to take at the highest possible SNR for a given activity distribution. Such factors for the ten highest-correlating metrics can be found in Table 5.4. The uncertainty on the correction factors listed in the table is on average ± 1.9 %.

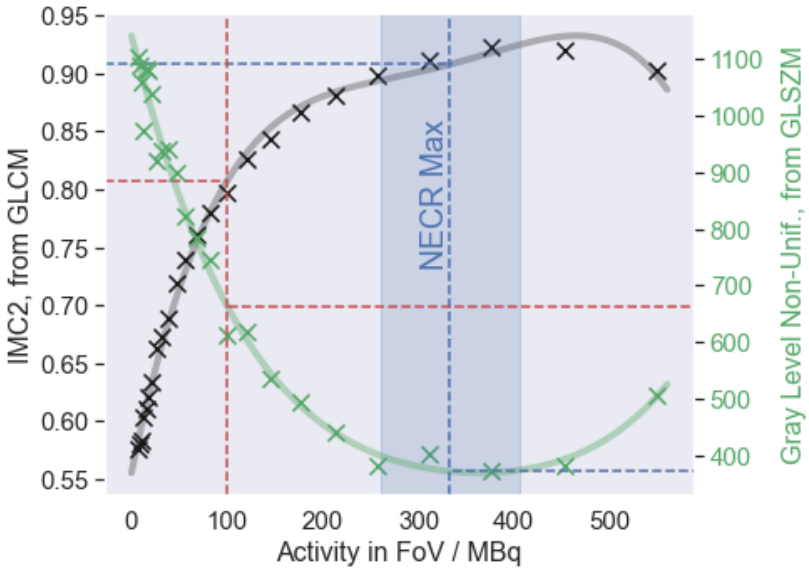


Figure 5.8: A diagram showing how feature compensation factors could be calculated. The feature value is targeted to be corrected from $A(\mathrm{NECR}_{\mathrm{max}})$ (blue line) to a reasonable clinical level (approximated to 100 MBq - red line).

5.2.2 Comparison to 5 Minute Cylinder Data

The 25 minute scans that have been considered are far longer than a typical patient scan is expected to take. Estimates using the Flow mechanism for continuous

Feature	|PMCC|	p-value	Compensation Factor, 100 MBq $\rightarrow$ A(NECR$_{max}$)
IMC2 (GLCM)	0.996	3.3×10^{-25}	1.125 ± 0.019
Correlation (GLCM)	0.994	1.5×10^{-22}	1.168 ± 0.036
Gray Level Non Unif. (GLSZM)	0.991	1.8×10^{-20}	0.656 ± 0.018
IMC1 (GLCM)	0.985	2.2×10^{-18}	1.679 ± 0.086
Dependence Entropy (GLDM)	0.984	6.5×10^{-18}	1.045 ± 0.010
MCC (GLCM)	0.982	1.8×10^{-17}	1.204 ± 0.036
Small Dependence Emphasis (GLDM)	0.964	3.8×10^{-14}	0.683 ± 0.004
Zone Percentage (GLSZM)	0.964	3.8×10^{-14}	0.594 ± 0.008
Run Length Non Uniformity (GLRLM)	0.963	4.7×10^{-14}	0.958 ± 0.005
Inverse Variance (GLCM)	0.963	4.7×10^{-14}	1.216 ± 0.008

Table 5.4: A table showing the ten radiomics features that correlate most with NECR for the 25 minute cylindrical phantom data alongside the respective compensation factors.

bed motion on the Siemens Biograph mCT TrueV give an approximate 3 minute duration in any single bed position [109]. If the 5 minute cylinder data is more representative of a clinical setting, it requires equally as rigorous an examination. It was observed that |PMCC| between NECR and the 75 texture features dropped significantly on average. Using the same $|\text{PMCC}| \geq 0.9$ threshold, only 7 features are now categorised as strongly NECR-correlated. These features are listed in Table 5.5 and a distribution similar to Figure 5.7 for the 5 minute data can be seen in Figure 5.9. The ten highest-correlating features, listed in Table 5.4, show a mean decrease in |PMCC| of (11.5 ± 6.6) %.

The plots in Figure 5.10 demonstrate the improved correlations as frame duration is increased for two selected features. There is evidence to suggest that, by increasing the scanning time, it is possible to achieve much better and clearer

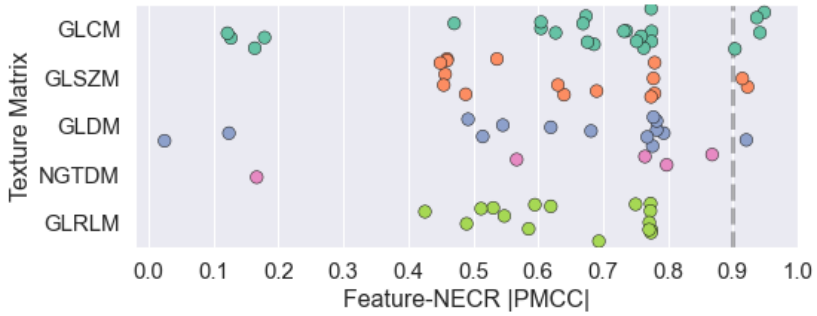
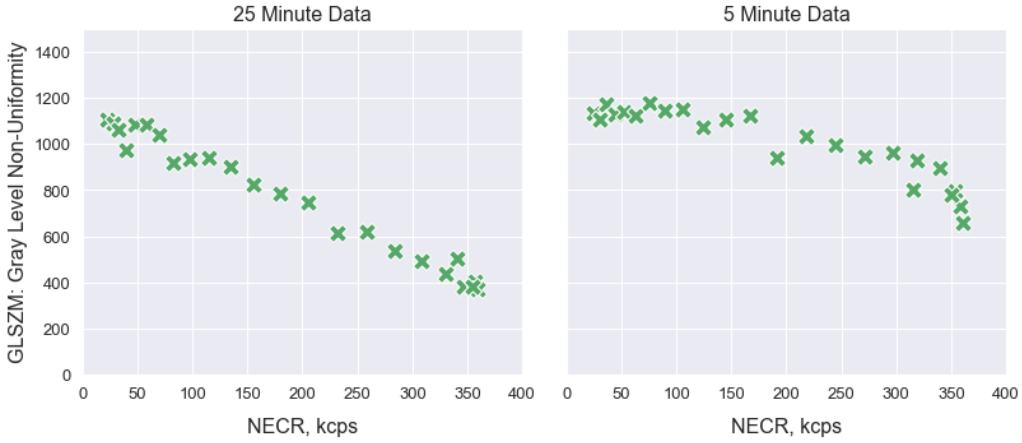


Figure 5.9: A 1D scatterplot showing the NECR |PMCC| of the 75 texture features for the 5 minute cylinder data. Jitter is applied in the vertical direction for visual aid.

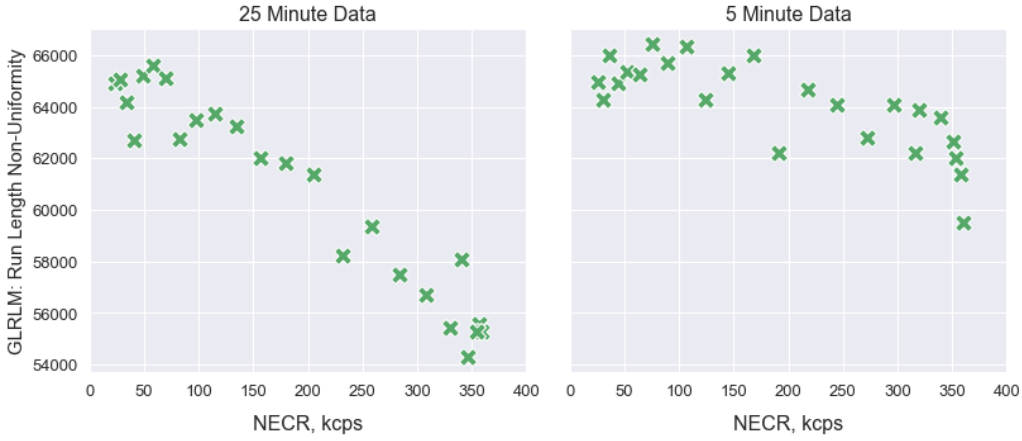
Features with NECR |PMCC| $\geq$ 0.9			
(25 Minute Data)	**|PMCC|**	**(5 Minute Data)**	**|PMCC|**
IMC2 (GLCM)	0.9965	IMC2 (GLCM)	0.9485
Correlation (GLCM)	0.9939	Correlation (GLCM)	0.9419
GrayLevelNonUniformity (GLSZM)	0.9905	GrayLevelNonUniformity (GLSZM)	0.9148
IMC1 (GLCM)	0.9853	IMC1 (GLCM)	0.9033
DependenceEntropy (GLDM)	0.9834	DependenceEntropy (GLDM)	0.9209
MCC (GLCM)	0.9822	MCC (GLCM)	0.9370
SmallDependenceEmphasis (GLDM)	0.9640		
ZonePercentage (GLSZM)	0.9640		
RunLengthNonUniformity (GLRLM)	0.9633		
InverseVariance (GLCM)	0.9633		
Idm (GLCM)	0.9632		
RunLengthNonUniformityNormalized (GLRLM)	0.9631		
Id (GLCM)	0.9631		
ShortRunEmphasis (GLRLM)	0.9623		
RunPercentage (GLRLM)	0.9622		
Strength (NGTDM)	0.9620		
DependenceNonUniformityNormalized (GLDM)	0.9607		
DependenceNonUniformity (GLDM)	0.9607		
DependenceVariance (GLDM)	0.9595		
LongRunEmphasis (GLRLM)	0.9593		
SizeZoneNonUniformity (GLSZM)	0.9587		
RunVariance (GLRLM)	0.9579		
LargeDependenceEmphasis (GLDM)	0.9573		
Coarseness (NGTDM)	0.9553		
DifferenceEntropy (GLCM)	0.9545		
Idn (GLCM)	0.9538		
DifferenceAverage (GLCM)	0.9506		
ZoneEntropy (GLSZM)	0.9469	ZoneEntropy (GLSZM)	0.9228
Idmn (GLCM)	0.9249		
Contrast (GLCM)	0.9202		
GrayLevelVariance (GLSZM)	0.9072		
Busyness (NGTDM)	0.9022		

Table 5.5: The highest NECR-correlating features for the cylinder dataset, listing all texture features with NECR |PMCC| greater than 0.9. Features are listed in descending order for the 25 minute dataset, and aligned on the right hand side for comparison.

noise modelling for apparently non-robust features by using NECR. However, this is not a generic solution. If a feature does not correlate well at 'faster' image acquisition times, it appears that such NECR correlation cannot be gained by increasing the frame duration (see Figure 5.11).



(a) Gray Level Non-Uniformity (GLSZM) against NECR



(b) Run Length Non-Uniformity (GLRLM) against NECR

Figure 5.10: Two examples of features with strong NECR correlation. (a) 25 min $|\text{PMCC}| = 0.9905$; 5 min $|\text{PMCC}| = 0.9148$ (b) 25 min $|\text{PMCC}| = 0.9633$; 5 min $|\text{PMCC}| = 0.7746$

Whether each feature appears in a similar population location for NECR correlation between each dataset is an interesting question. Both datasets have 17 features with $|\text{PMCC}| \leq 0.5$, but only have seven of these features in common[20]. While there are zero features that have $|\text{PMCC}| \leq 0.5$ in one dataset and

[20]These are Sum Squares (GLCM), Gray Level Non-Uniformity (GLDM), Gray Level Non-

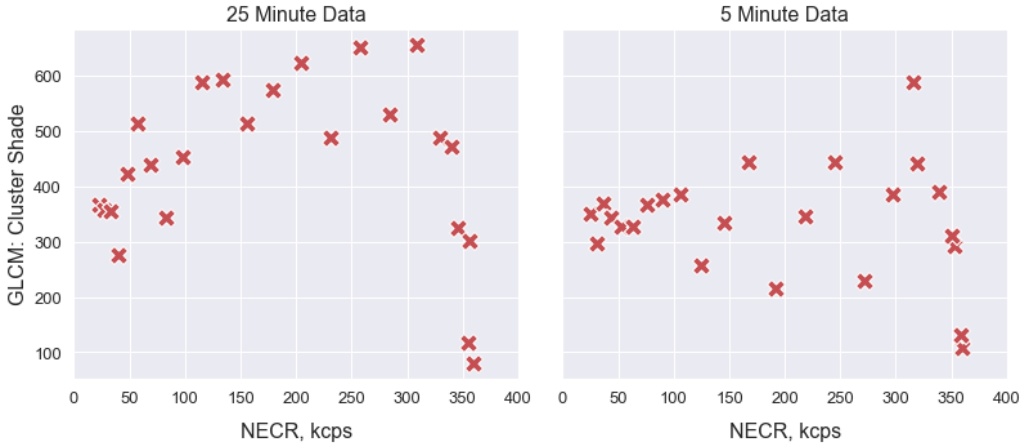


Figure 5.11: Scatterplots showing Cluster Shade (GLCM) against NECR for the cylinder datasets. 25 min $|\text{PMCC}| = 0.1102$; 5 min $|\text{PMCC}| = 0.1267$

Features with Lowest NECR Correlation			
(25 Minute Data)	**|PMCC|**	**(5 Minute Data)**	**|PMCC|**
SmallAreaLowGrayLevelEmphasis (GLSZM)	0.3360	ZoneVariance (GLSZM)	0.4546
SumAverage (GLCM)	0.3232	LowGrayLevelZoneEmphasis (GLSZM)	0.4495
JointAverage (GLCM)	0.3232	LongRunLowGrayLevelEmphasis (GLRLM)	0.4257
ShortRunHighGrayLevelEmphasis (GLRLM)	0.2710	MaximumProbability (GLCM)	0.1786
ShortRunLowGrayLevelEmphasis (GLRLM)	0.2241	Contrast (NGTDM)	0.1665
LongRunLowGrayLevelEmphasis (GLRLM)	0.2120	JointEnergy (GLCM)	0.1635
LowGrayLevelRunEmphasis (GLRLM)	0.1358	ClusterShade (GLCM)	0.1267
LowGrayLevelEmphasis (GLDM)	0.1353	LargeDependenceLowGrayLevelEmphasis (GLDM)	0.1238
ClusterShade (GLCM)	0.1102	JointEntropy (GLCM)	0.1215
LowGrayLevelZoneEmphasis (GLSZM)	0.1034	SmallDependenceHighGrayLevelEmphasis (GLDM)	0.0245

Table 5.6: The ten features with the lowest correlation with NECR for 5 and 25 minute data.

$|\text{PMCC}| \geq 0.9$ in the other, there are three that have $|\text{PMCC}| \leq 0.5$ in 5 minute data that have $|\text{PMCC}| \geq 0.8$ in 25 minute data; these are Large Area Low Gray Level Emphasis (GLSZM), Large Area Emphasis (GLSZM), and Zone Variance (GLSZM).

Considering 25 and 5 Minute Data Together

Equation 3.2 states that the square of the SNR_{data} is equivalent to the product of the NECR and the duration of acquisition of the image, Δt. NECR is used throughout this study as a proxy for the signal noise ratio, and the subjects of

Uniformity Normalized (GLRLM), Small Area Low Gray Level Emphasis (GLSZM), Long Run Low Gray Level Emphasis (GLRLM), Cluster Shade (GLCM) and Low Gray Level Zone Emphasis (GLSZM).

the study are measures of heterogeneity that are affected by noise. By analysing feature changes with respect to $\text{SNR}^2_{\text{data}}$, the datasets could be combined.

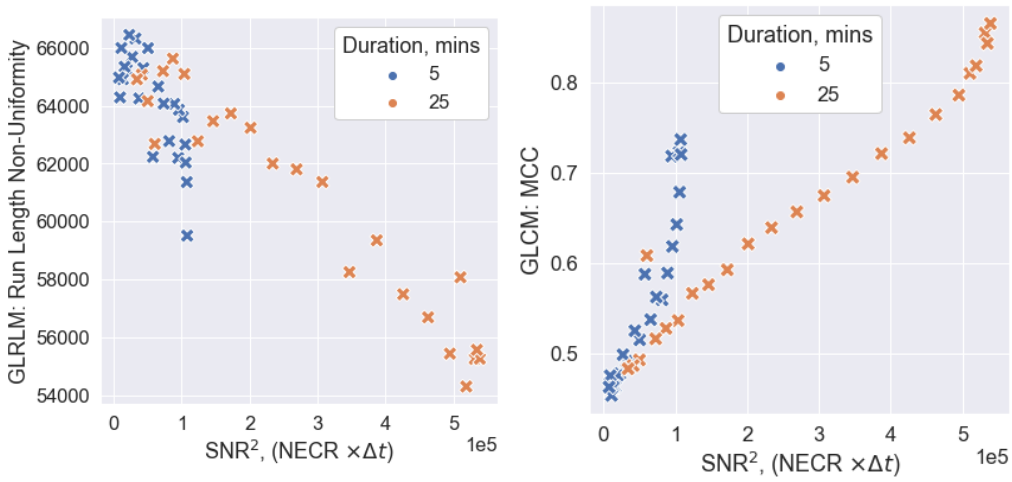


Figure 5.12: The GLRLM Run Length Non-Uniformity (left) and the Maximal Correlation Coefficient (MCC, right) plotted against the square of the SNR_{data} for all cylinder images.

Figure 5.12 shows two examples of features plotted against $\text{SNR}^2_{\text{data}}$. Both features correlated strongly with NECR when the differing scan duration image subsets were considered separately. By combining the image datasets, any found correlations would be more reliable due to the greater number of data points. The behaviour of the two chosen features show contrasting outcomes in attempting this comparison. The Run Length Non-Uniformity plot suggests strong linear behaviour across the combined dataset, which is supported by a strong |PMCC| of 0.9342. This behaviour is also seen when the two subsets are separately analysed, as seen in Figure 5.10b. The behaviour of the MCC, however, suggests that the comparison across all images might not work for all features. The two sets appear drawn from different distributions, and certainly do not achieve acceptable inter-set linearity.

Figure 5.13 shows how the variance of voxel values in the ROI for the cylinder images changes with $\text{SNR}^2_{\text{data}}$. The variance, a first order measurement, shows a continuous function common to both 5 and 25 minute datasets. There is hence evidence that, when performing the image reconstruction on the 5 and 25 minute dataset, there is a decoupling of the SNR_{data} from the higher-order texture matrix features. There are many potential sources of this, such as the potentially variant

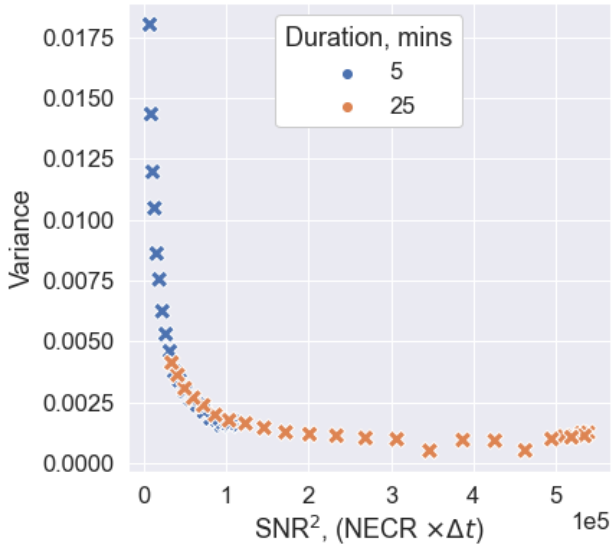


Figure 5.13: The variance of voxel values plotted against the square of the SNR_{data} for all cylinder images.

efficiency of running the scatter estimation algorithms on a dataset with lower numbers of counts. This is an important discovery for this work, and motivates analysing image datasets with differing acquisition durations separately. Further work should seek to investigate whether changing image reconstruction protocols could preserve the feature-SNR_{data} trend continuity between the 5 and 25 minute datasets.

5.2.3 NEMA IQ Phantom Data

The NEMA IQ phantom data was planned in order to maximise sampling over $NECR_{peak}$. While this provided many more images at a single duration (31 instances of a 5 minute duration acquisition), this protocol did not provide the two well-defined and comparable image subsets as seen in the other acquisitions. The 32 five minute frames were taken as a comparable set to the 5 minute cylinder data. The standard image reconstruction protocol, listed in Table 4.5, was followed. The ROIs for the six spheres were established using a spherical mask created with the known volume for each.

Figure 5.14 shows the |PMCC| against NECR for the ten features listed in

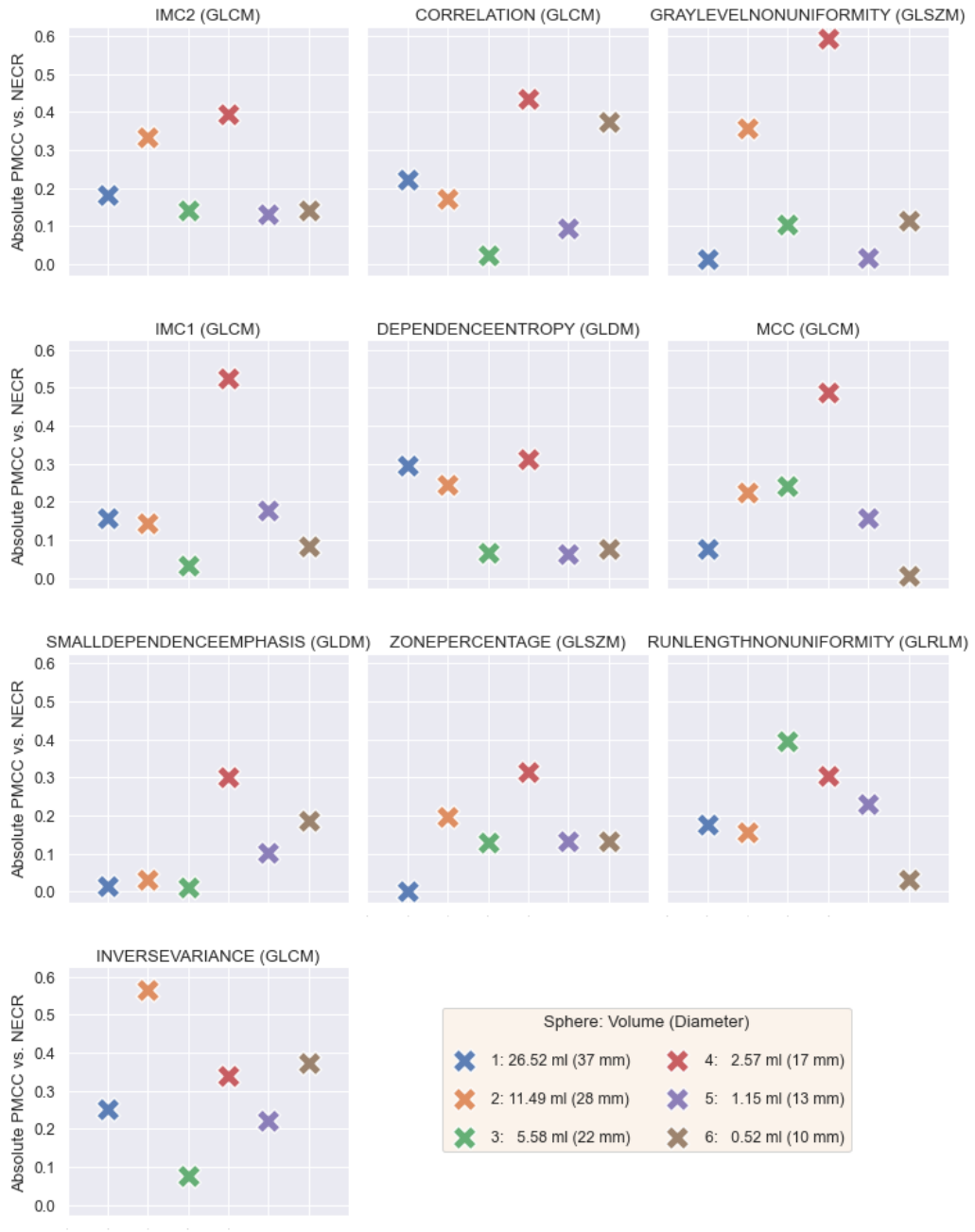


Figure 5.14: One-dimensional scatterplots showing the correlations for the six NEMA spheres, considering only the ten highest correlating features from the 25 minute cylinder dataset as listed in Table 5.4, labelled above.

Sphere (Diameter)	Mean Drop in NECR-|PMCC| for Features of Interest	
	Raw |PMCC|	Percentage
1 (37 mm)	-0.73 ± 0.12	-84 ± 12 %
2 (28 mm)	-0.62 ± 0.16	-72 ± 19 %
3 (22 mm)	-0.74 ± 0.15	-86 ± 15 %
4 (17 mm)	-0.47 ± 0.08	-54 ± 10 %
5 (13 mm)	-0.73 ± 0.13	-84 ± 9 %
6 (10 mm)	-0.72 ± 0.16	-82 ± 15 %

Table 5.7: Drop in NECR correlation from the 5 minute cylinder data for the six NEMA spheres for the 5 minute acquisitions, averaged over the ten features listed in Table 5.4.

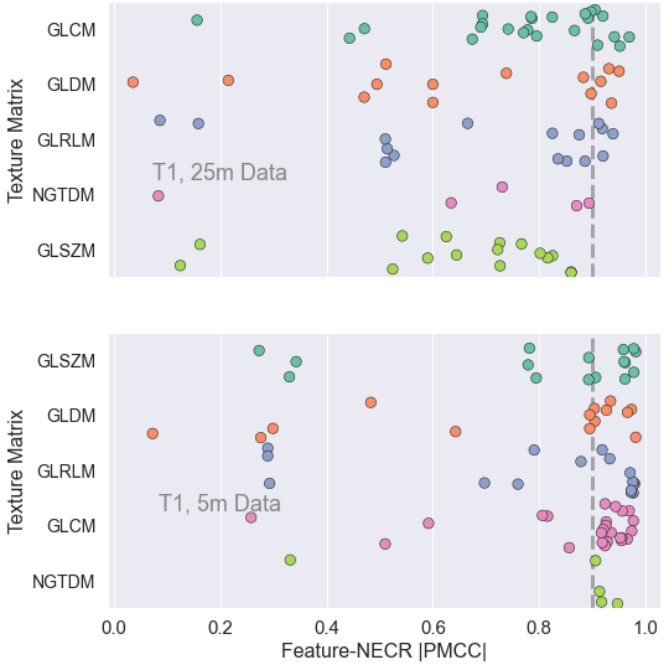
Table 5.4 for each of the six spheres. There is a notable decrease for all features, with none of the ten surpassing $|\text{PMCC}| = 0.6$ for any of the spheres. There is no apparent dependence on ROI volume for the drop in feature-NECR correlations. This information can be summarised by examining how the feature-NECR correlations for the ten features detailed in Table 5.4 drop for the spheres, and can be found in Table 5.7.

5.2.4 Custom Tumour Phantoms

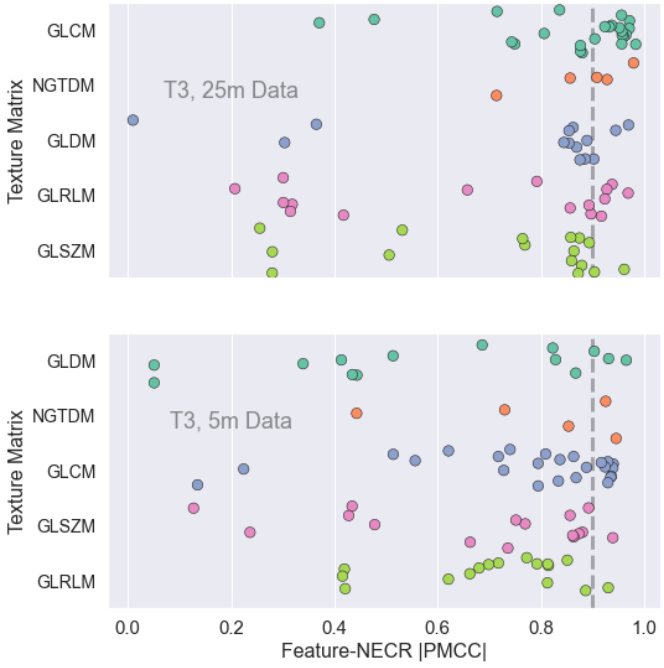
The custom tumour phantom scans were performed with the same protocol as the initial cylinder data. Images were reconstructed similarly and ROIs determined using known volumes for the two 'T' inserts, with an additional background ROI determined using a cylinder with a 428 ml volume.

T1+T3

Figure 5.15 shows the NECR |PMCC| for every textural feature for the two defined 'tumour' ROIs in the T1+T3 phantom scan. The NECR correlations were expected to be poor for the localised regions; instead, features show generally strong correlations with global NECR for both 25 and 5 minute data. While for T3 the general trend in correlation strength decreases when decreasing frame duration, the same cannot be said for T1, where the 5 minute data show features generally show a more obvious correlation.



(a) T1 feature NECR correlations for 25 and 5 minute data.



(b) T3 feature NECR correlations for 25 and 5 minute data.

Figure 5.15: Feature-NECR correlations shown for all datasets in *T1+T3* scan. Strong correlation criterion of $|\text{PMCC}| = 0.9$ shown as dotted grey line.

T2+T4

Figure 5.16 shows the equivalent to Figure 5.15 for the second custom phantom scan, using T2+T4. The results for the T2+T4 scan are more varied; the distribution of feature correlation strength is wider, and generally located at lower correlation coefficient values. Curiously, as in T1, the correlation coefficients are generally higher in the 5 minute data than the 25 minute data. Table 5.8 shows the number of texture features that fit the definition of strong correlation used in the cylinder data; there are many in the *T1+T3* scan series data, and few in the *T2+T4* scan series data.

		Number of Features with $\|$PMCC$\| \geq 0.9$	
Region	**Volume**	**25 Minute Data**	**5 Minute Data**
T1	229 $\pm$ 2 ml	14	46
T3	71 $\pm$ 2 ml	27	15
T2	124 $\pm$ 2 ml	0	2
T4	41 $\pm$ 2 ml	0	0

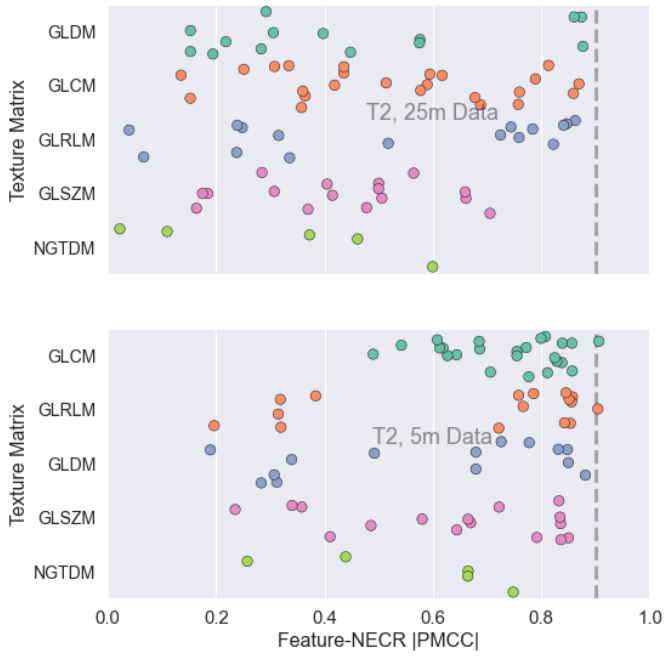
Table 5.8: Table listing number of strongly NECR-correlated texture features for each dataset

Backgrounds

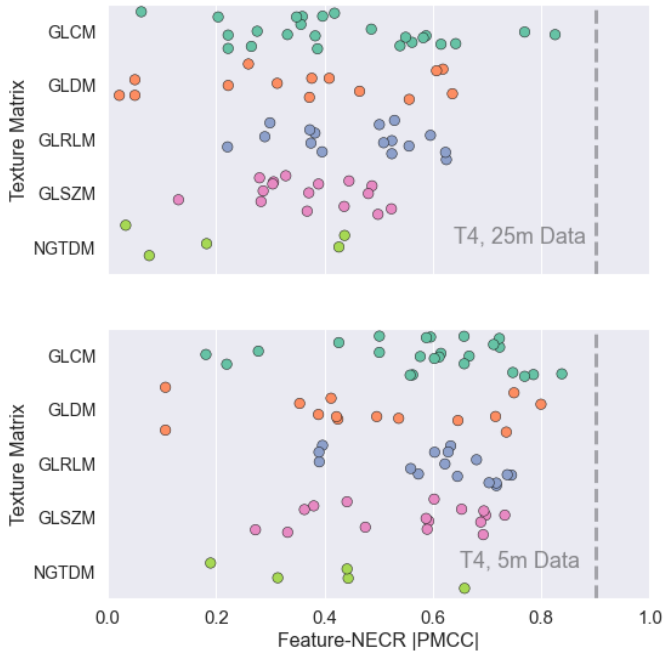
Similar plots are shown for the background ROIs for *T1+T3* and *T2+T4* scans in Figure 5.17. Interestingly, although these ROIs are more similar in shape and volume to the initial cylinder data, the general texture feature NECR correlations are much worse, with no features surpassing a |PMCC| of 0.7 for either dataset.

Initial observations of the two scan series (with six ROI datasets each with two frame duration subsets) promote several questions.

1. Are the same features the strongest-correlating with NECR across all datasets?
2. Are the same features the weakest-correlating with NECR across all datasets?
3. Is there a link between feature correlation with NECR and the volume of the phantom insert used?
4. Why do the background ROIs exhibit weaker NECR correlation in texture feature values than the inserts and the cylinder datasets?

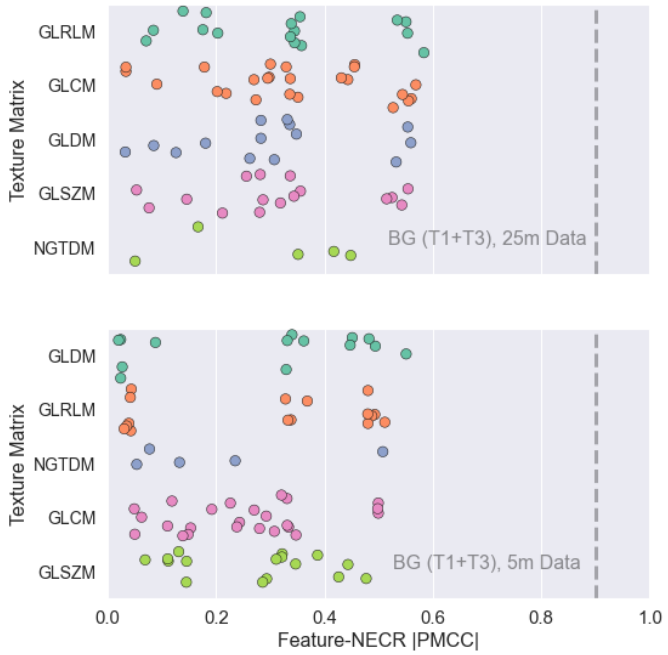


(a) T2 feature NECR correlations for 25 and 5 minute data.

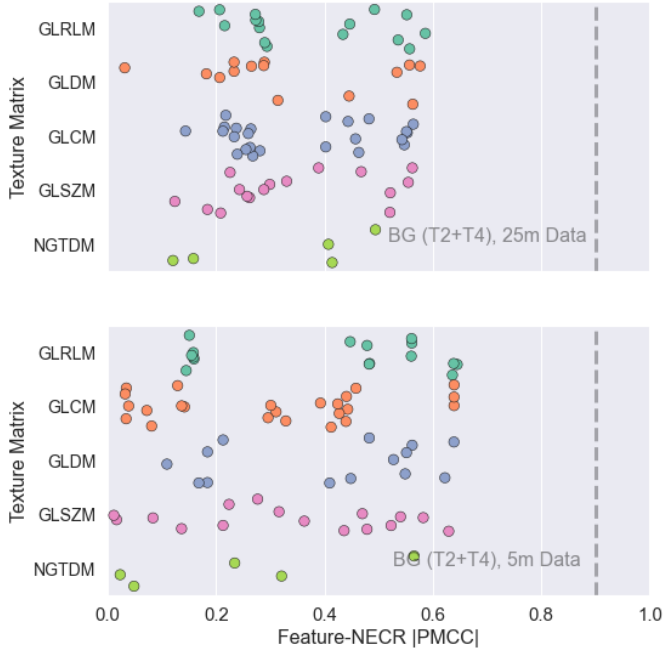


(b) T4 feature NECR correlations for 25 and 5 minute data.

Figure 5.16: Feature-NECR correlations shown for all datasets in *T2+T4* scan. Strong correlation criterion of $|\text{PMCC}| = 0.9$ shown as dotted grey line.



(a) Background of T1+T3 feature NECR correlations for 25 and 5 minute data.



(b) Background of T2+T4 feature NECR correlations for 25 and 5 minute data.

Figure 5.17: Feature-NECR correlations shown for background regions in *T1+T3* and *T2+T4* scans. Strong correlation criterion of $|\text{PMCC}| = 0.9$ shown as dotted grey line.

The first two questions can be addressed in tandem. The list of features for each ROI can be thought of as a ranked list, rated by their correlation with NECR. If the same features occur in similar locations in these ranked lists for all ROIs across all datasets, then it could be possible to begin to define a group of features that are more suitable for NECR-led correction. The *intraclass correlation coefficient* was derived and computed with `pingouin`, a statistical Python package, using the features as 'targets', the ROIs (for the cylinder and all 'T' scans, separated into 25 and 5 minute data to give 14 total) as 'raters' and the NECR |PMCC| as the 'rating' [110]. The ICC3K statistical test was used; here, the statistic (in the domain $[0,1]$) is closer to 1 if the average ratings of a fixed set of k raters are reliable. Such a test, with a corresponding high value, will indicate whether the ratings are reliable, and whether features are consistently highly NECR-correlated or weakly NECR-correlated across every dataset. The value obtained is $ICC(3,k) = 0.77740$, with $p \sim 0$ and a 95 % confidence interval of $[0.70, 0.84]$. The results show that there is evidence that features are similarly ranked for each ROI, although it is unlikely that any features maintain the same ranking. Further post-hoc analysis must be done to establish which features consistently appear poorly-correlated, which features appear strongly-correlated, and which features (if any) are correlated strongly or weakly depending on the dataset and scan.

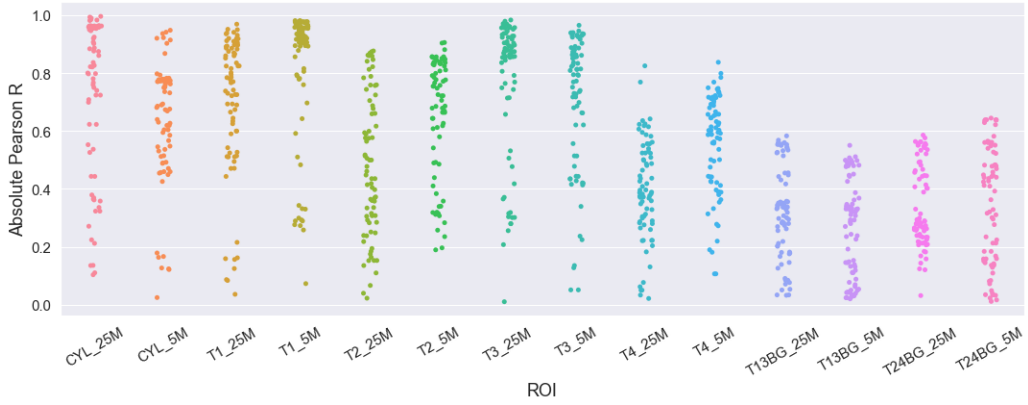


Figure 5.18: Each feature's |PMCC| plotted for each ROI in each dataset, illustrating the spread of correlations across all collected data.

When comparing the individual rankings, no features appear commonly in the

top and bottom twenty for all ROIs. Even when discounting the background ROIs for the 'T' scans, only one feature appears in the top twenty for the remaining ROIs (GLCM IMC1), and one commonly for the bottom twenty (GLSZM Small Area Low Gray Level Emphasis). Figure 5.19 shows the rank averages across

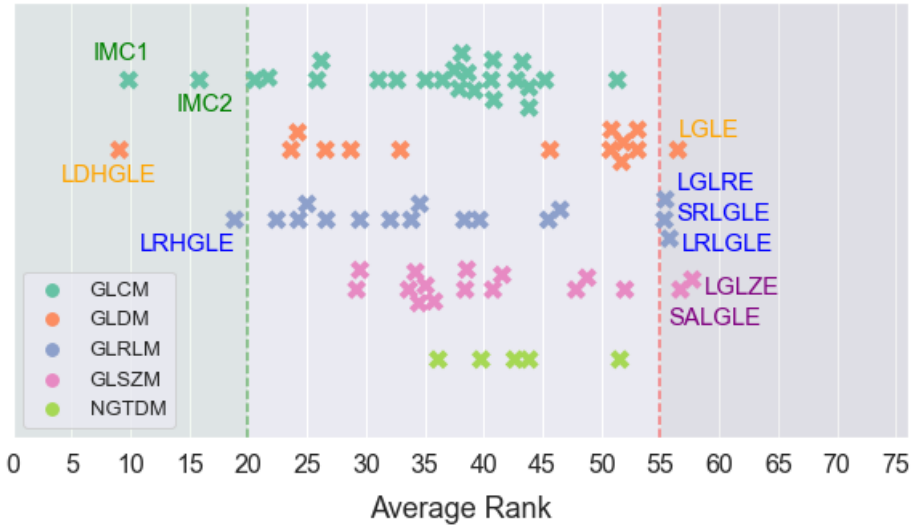


Figure 5.19: The average rank for all features when ranked by NECR correlation strength, averaged over the 14 ROIs.

all 14 datasets, with the top and bottom twenty features highlighted. Consistently highly-correlating features include the IMC1 and IMC2 from the GLCM, the Low Dependence High Gray Level Emphasis from the GLDM and the Long Run High Gray Level Emphasis from the GLRLM. Features with consistently weak correlation include the Low Gray Level Emphasis from the GLDM, the Low Gray Level Run Emphasis, Short Run Low Gray Level Emphasis and Long Run Low Gray Level Emphasis from the GLRLM, and the Low Gray Level Zone Emphasis and Small Area Low Gray Level Emphasis from the GLSZM. The spread of feature ranks is, however, localised around the mid-range of the rankings. It should be noted that this does not mean that features that are mid-ranking are consistently uncorrelated with NECR; we observe (in Figure 5.18) that the rankings are often skewed towards high correlation, meaning that even mid-ranked features display strong correlations in some datasets.

Features are calculated on discretised 'homogeneous' objects. When discretising the images SUV_{min} and SUV_{max} are used as the range with a fixed bin

number. It is known that SUV_{max} is an unstable measure for test-retest repeatability [111]. The range boundaries are therefore sensitive to image noise, and the bulk of image intensities are likely to lie in the mid-range for homogeneous ROIs. The weakest-correlating features, those consistently low-ranking in the feature NECR-correlations as in Figure 5.19, are *low gray level*-based features. For ground-truth homogeneous objects, it is expected that the high and low gray level emphasis features are likely to be unstable due to the low statistics in the fringe intensity bins. If the NECR is a useful proxy for SNR in the data, it stands to reason that the approximation works better in regions with a greater proportion of the useful signal; it is not that the features are unaffected by the level of image noise, only that the effects of this increased data noise are not as easily remedied.

Figure 5.18 demonstrates that the link between volume and general texture feature NECR correlation is weak, although the circumstances surrounding each ROI in particular should be taken into account. Firstly, the background regions are larger than any of the tumour ROIs, yet exhibit consistently weak NECR correlations. Part of the reason behind this is in the very fact that these are background regions, with fewer counts and hence a lower signal component. There appears to be no general trends in ROI volume for feature correlation, however it is notable that correlations are consistently weaker for ROIs smaller than T4 (volume 41 cm^3). While no definitive conclusions can be made owing to the small number of geometries under investigation, it is noteworthy that this is comparable to the 45 cm^3 volume at which Brooks & Grigsby's paper suggests that GLCM-based features may no longer predict accurate measures [15, 79].

5.2.5 Discussion

The experiment began by examining images of a cylinder, filled uniformly with ^{18}F and utilising a region of interest spanning the entire object. A 'ground truth' image of such an object would be completely homogeneous, but noise is imposed onto the resultant image due to counting statistics. What we measure, in theory, when we observe the texture of our cylinder image, is the texture provided solely by noise. While it is not uncommon for image artifacts to manifest due to activity diffusion patterns in phantoms, steps were taken to ensure full mixing of the ^{18}F

and water prior to imaging; the possibility of diffusion artifacts can be eliminated.

Such high correlations with the NECR for many texture features can be seen as an indication of poor robustness. However, should we be able to correct for the imaging conditions (in this case, initial activity) then there is no reason why the measured value of such a feature should be considered unreliable. NECR forms a useful metric which can approximate the noise that we see in our image, and hence an appropriate platform to begin such a discussion.

We cannot discuss measures of correlation without discussing the associated p-values at the same time. The p-value can be considered as the probability of finding the value of the |PMCC| should the two variables in reality be completely uncorrelated. It is observed in Table 5.4 that the highest-correlating features have p-values calculated so small that they are virtually zero. These raw values are not particularly useful due to the small size of the dataset and the omission of uncertainties - which we are trying to establish in the first place. It is however important to consider the associated p-values when lower |PMCC| is measured. Figure 5.20 demonstrates that as lower |PMCC| is measured, it becomes more probable that the feature is uncorrelated to the NECR; an unremarkable statement, but it is noteworthy that for a p-value around 0.5 (indicating weak correlation) it is unlikely that such a value would be measured without any relationship between the feature and NECR.

We have observed that many radiomics texture features are affected by noise, with large volume, long time acquisition phantom scans exhibiting many texture features that correlate strongly with NECR. This appears to confirm our prior assumption that image noise does affect measured feature values, while promoting optimism that well-modelled noise can enable compensation of these features from noisy but attainable conditions to those which are more optimal (higher SNR). We know that the correlations between features and NECR (specifically the NECR) become more confused when considering local ROIs within a hot background (the NEMA IQ spheres and the custom tumour phantom inserts). However we have been limiting ourselves to considering only the NECR as our noise metric, which is de facto a global measure. If we were to establish some way of quantifying the noise on a local level, correlation with this metric could enable correction of a texture feature from any given ROI. This shall be explored

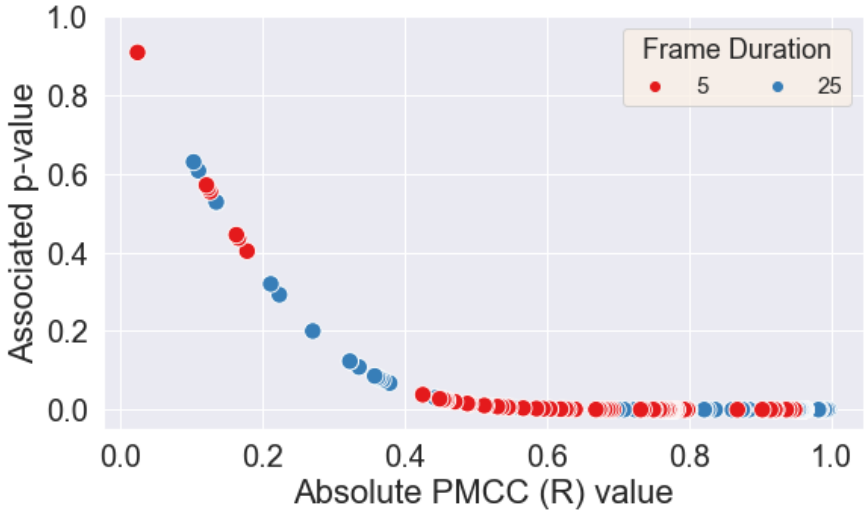


Figure 5.20: A scatter plot showing the associated p-values for measured |PMCC| between the 75 texture features and NECR in the 5 and 25 minute cylinder datasets.

in the next chapter.

5.2.6 Assessing Robustness using Kruskal-Wallis

In this phantom data, we have determined that we cannot eliminate the possibility that NECR (or an NECR-like local metric) can be used to predict how the value of a radiomics texture feature may change with increasing base activity levels. However, what is unclear is that on a case-by-case basis whether this change would lead to a mischaracterisation by some potential future classifier. It is important to state that the textural profile of a ROI will be defined by many features in combination as opposed to relying heavily on single values[21]. To simplify the problem, we need to determine whether the change that an increase in image noise may impose on a feature's value is enough to confuse it with the value said feature may take on a completely different ROI. Figures 5.21 and 5.22 illustrate the spread of values for the three highest and lowest correlating features for the thirteen ROIs in the dataset.

A statistical test is required to determine whether the samples for each ra-

[21]Many machine learning classifiers assign 'importances' (or weights) to each feature included in the input data, and so strictly the importance of reliably quoting the value of some features may be greater than others depending on the model used.

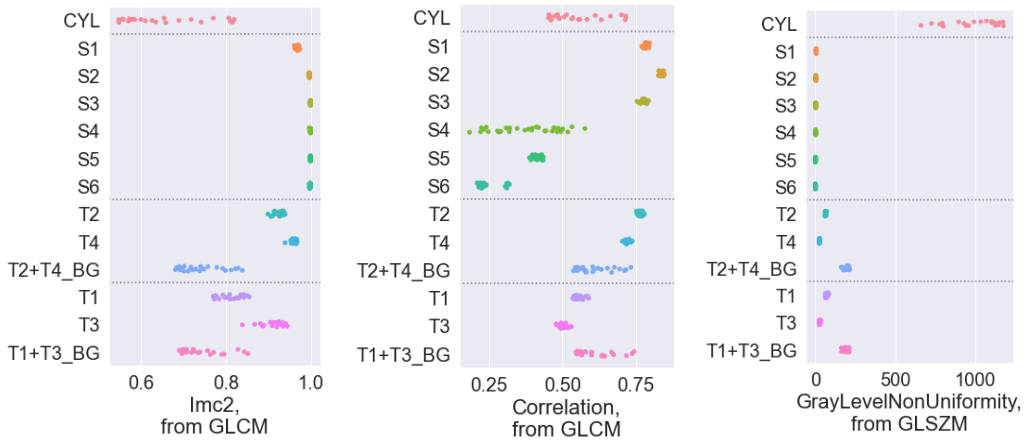


Figure 5.21: Three features which are consistently highly NECR-correlated, plotted for the 5 minute datasets for all ROIs included in this work.

diomics feature measurement could possibly be drawn from the same distribution, despite the different ROIs on which they are measured and the spread of measured values, which are suspected to be caused by changes in NECR. If the radiomics statistics are truly *robust*, they should be independent of the activity present in the FoV, but if the values can be corrected then the feature is still useful. We shall treat our ROIs as categorical data; while the ROI volume could be interpreted as an ordinal, the varying shapes of the ROIs and prior knowledge of how ROI shape affects statistics imply that many covariates would need to be considered. Treating the data as categorical samples from each ROI would broaden the statistical test. The starting point considered was an ANOVA (**an**alysis **o**f **va**riance) test, designed to determine whether the means of groups within a population can be significantly distinguishable. However, there is an underlying assumption in ANOVA that variance in a group can be assumed to be normally distributed, which measurements in our groups cannot be said to be drawn from. A variation of ANOVA, the Kruskal-Wallis test, is non-parametric, meaning that no assumptions need to be made on the underlying distribution from which the data is drawn. The hypotheses should be set up thus:

- *null hypothesis,* the median across all groups is equal;
- *alternate hypothesis,* the medians from each group are not equal.

The Kruskal-Wallis test calculates the H statistic by ranking all data and com-

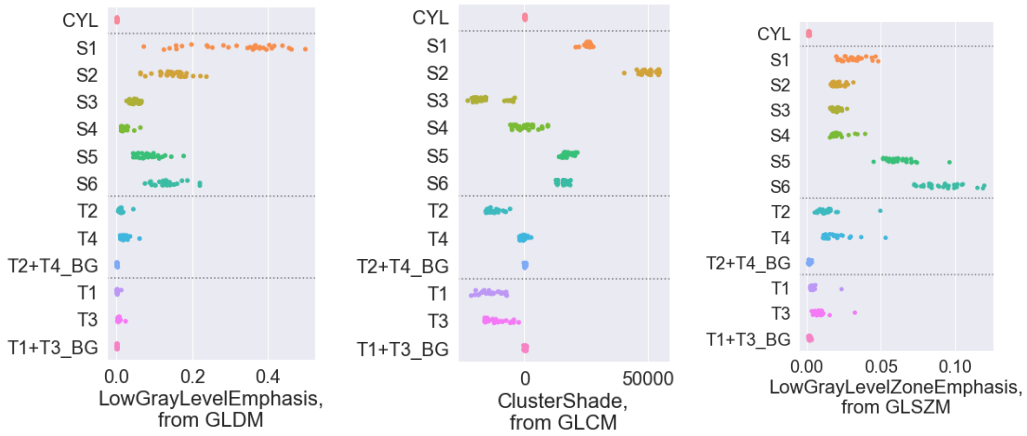


Figure 5.22: Similar to Figure 5.21, three features which are consistently weakly NECR-correlated, plotted for the 5 minute datasets for all ROIs included in this work.

puting H such that

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{C} \frac{R_j^2}{n_j} - 3(n+1) \tag{5.7}$$

where n is the total number of measurements, C the number of groups, n_j the number of measurements in the group j and R_j the average rank of the measurements in group j. The statistic is then compared to the χ^2 statistic for $C-1$ degrees of freedom; if $H > \chi^2$ then the null hypothesis is rejected [112]. The Python package *pingouin* was used to carry out Kruskal-Wallis tests for each of the 75 features. Listed in Tables 5.9 and 5.10 are the outputs of the Kruskal-Wallis test for the top ten and bottom ten features that correlated best with NECR for the cylinder dataset.

In all cases, there are 12 degrees of freedom due to the 13 ROIs used. The χ^2 value at the 0.001 significance level is 32.91 [113]. This value indicates that if H is greater than 32.91, there is a 0.1% probability that the data is drawn from the same distribution. For all radiomics features tested over the 13 ROIs, the null hypothesis can be rejected, as all H values comfortably exceed this value. It can be concluded that there is evidence to suggest that despite the spread of values for each texture feature, the ROI groups are distinguishable.

There are some stipulations to the results of this test, however. The assumptions made when setting up the test demand independent observations; that is, each measurement taken in any particular group is not dependent on any other

Degrees of Freedom	Kruskal-Wallis H Statistic	Two-Tailed p-value	Feature	Matrix
12	336.87	0.0000	IMC2	GLCM
12	335.60	0.0000	Correlation	GLCM
12	347.93	0.0000	Gray Level Non-Uniformity	GLSZM
12	344.72	0.0000	IMC1	GLCM
12	340.51	0.0000	Dependence Entropy	GLDM
12	339.25	0.0000	MCC	GLCM
12	331.94	0.0000	Small Dependence Emphasis	GLDM
12	334.83	0.0000	Zone Percentage	GLSZM
12	349.79	0.0000	Run Length Non-Uniformity	GLRLM
12	322.69	0.0000	Inverse Variance	GLCM

Table 5.9: Results of the Kruskal-Wallis test for an example set of suspected highly NECR-correlating features. The features chosen are the features from Table 5.4.

Degrees of Freedom	Kruskal-Wallis H Statistic	Two-Tailed p-value	Feature	Matrix
12	307.77	0.0000	Small Area Low Gray Level Emphasis	GLSZM
12	315.25	0.0000	Sum Average	GLCM
12	315.25	0.0000	Joint Average	GLCM
12	296.04	0.0000	Short Run High Gray Level Emphasis	GLRLM
12	333.96	0.0000	Short Run Low Gray Level Emphasis	GLRLM
12	335.64	0.0000	Long Run Low Gray Level Emphasis	GLRLM
12	334.52	0.0000	Low Gray Level Run Emphasis	GLRLM
12	334.91	0.0000	Low Gray Level Emphasis	GLDM
12	328.03	0.0000	Cluster Shade	GLCM
12	319.60	0.0000	Low Gray Level Zone Emphasis	GLSZM

Table 5.10: Results of the Kruskal-Wallis test for an example set of suspected highly NECR-correlating features. The features chosen are the ten lowest-correlating in the cylinder 25 minute data, listed in Table 5.6.

measurement, and that there are no repeated measurements. The validity of this assumption can be questioned by the proposal that the value of the texture feature can be said to be dependent on the NECR. It can easily be argued that each measurement for each ROI is independent, as the each measurement is taken with a different level of starting radioactivity. The underlying ground truth distribution is unchanging, but the changing activity means that no measurements are strictly repeated. The Kruskal-Wallis test is also ambiguous, as it does not give any indication as to which groups can be distinguished from each other and in which direction. For the sake of this test, we are aware that the texture feature values are highly dependent on shape and volume, and these factors have been encapsulated

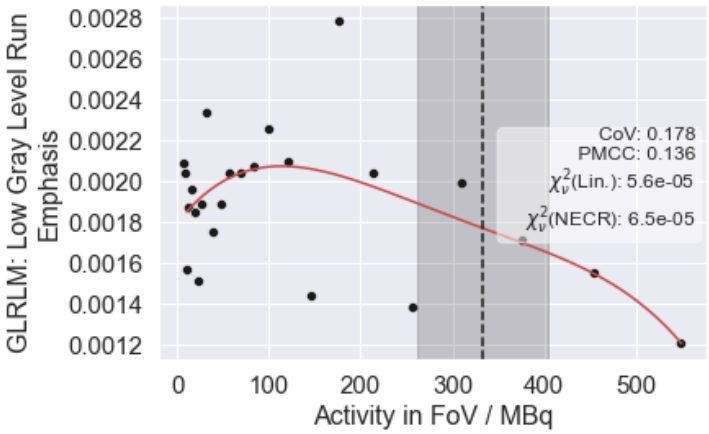
into our categorical variable.

This test provides evidence that despite the lack of robustness for each ROI, it could be possible to distinguish ROIs by the feature values for all texture features. This is unlikely to be the case when further heterogeneous distributions of different sizes are included however, and such tests should be repeated using many more sizes, shapes and heterogeneity of activity distribution for further verification.

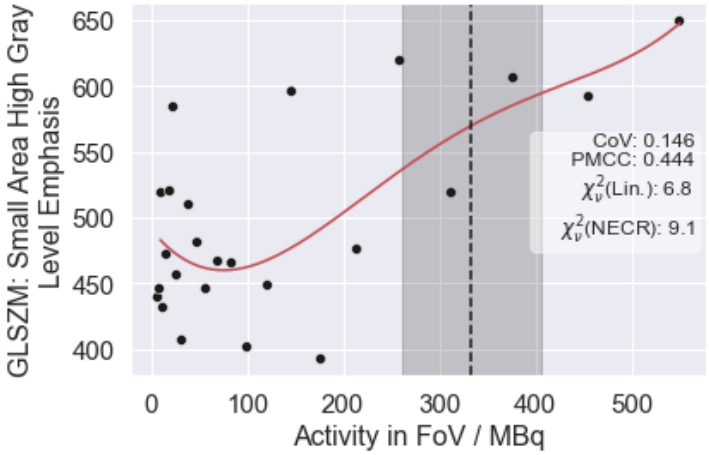
The use of heterogeneous phantoms, of which the possible designs are outlined in Section 4.2.5, would enable benchmarking of these feature values. An identical method to that used in this work would be used, except a dual-section design phantom would be filled with differing pre-determined activity concentration. This creates a new ground truth unseen in this work, with the discretisation taking place over a wider range of SUV and small inter-voxel perturbations of image noise taking lower significance. One expects the sizes of the zones and the runs in the GLSZM and GLRLM for a heterogeneous phantoms to be more varied than those found in the homogeneous versions presented so far. This experiment could be repeated with differing ratios of activity concentration between sections to provide new benchmarks. This extra data, provided potential feature-NECR correlation, would give a similar Kruskal-Wallis test enough diversity to draw significant conclusions.

5.3 Conclusions

This phantom study sought to determine the robustness of texture features to activity in the field of view. It can be suggested that the fact that many features correlate strongly with NECR demonstrates a lack of robustness, yet the strong possibility of correcting feature values by making use of this correlation promotes optimism that lacking classical 'robustness' need not impact their future use, and enable uncertainty estimation for these compliant features. For features that exhibit weak NECR correlation, the converse, that therefore these features *are* robust, cannot be said to be true from this data. Drawing a consensus is difficult for two reasons. The significantly different orders of magnitude in which these feature values lie make establishing definitive criteria (such as limits in CoV



(a) GLRLM Low Gray Level Run Emphasis



(b) GLSZM Small Area High Gray Level Emphasis

Figure 5.23: Examples of poorly-correlating texture features for the 25 minute cylinder data, along with example of the poor resultant quartic fitting (red line). |PMCC| against NECR is shown as 'PMCC' on the figures.

or reduced-χ^2 for linear behaviour with activity) for robustness invalid. This is demonstrated to some extent by the plots for two features in the 25 minute cylinder data, shown in Figure 5.23. Secondly, the manifold of geometries on which these tests are being done lacks variation beyond homogeneous (albeit with noise) objects. The Kruskal-Wallis test above determined to some satisfaction that the values between the geometries tested so far can be significantly clustered, but these are all nominally homogeneous objects. To more rigorously determine this second point, a much more developed and diverse set of activity distributions is required.

The NECR has been used until this point for pragmatism, as it is an easily calculable metric with a clear and easily explainable parallel to data signal noise ratio. Despite the strong analogy between image and data noise for analytical reconstruction such as FBP, previous work has shown that NECR is a weak predictor for image noise in modern iterative reconstruction [99]. The strong correlations displayed in the work demonstrate that the use of such a metric can be useful, as the ratio may still be indicative of the amount of correction performed on the data. The strong correlations motivated investigation into new definitions of noise metrics. The NECR in its current guise is not sacrosanct to defining the noise in the data, and it could be possible to define measures of noise more appropriate to localised regions, incorporating knowledge of global activity effects such as the dead time and randoms effects that cause the characteristic NECR curve. The work done into defining this new *tumour-specific noise equivalent count rate* is explored in the next chapter.

5.4 Summary

This chapter has demonstrated that there is some evidence of NECR correlation in radiomics texture features. This correlation is seen in large volume ROIs and long-duration acquisition frame data, but is not observed in all features. In examples of successful correlation, a method is proposed to enable correction of feature values from noisy low-activity clinical levels to the value expected at peak NECR. Assessing general robustness is difficult due to the lack of variety in the ground truth heterogeneity of the included phantoms. There is evidence that

Low Gray Level Emphasis features across GLDM, GLRLM and GLSZM matrices exhibit poor robustness, but this is complicated by the homogeneity of the phantoms used. The size of the regions appears to affect the NECR correlation, with smaller objects showing poor NECR correlation, yet this is complicated by the diminished feasibility of using texture analysis on small objects. The NECR correlation is encouraging despite the diminished effectiveness of using NECR as a proxy for SNR in the UHD reconstruction method, motivating the development of new tumour specific quantifications of local noise to better assess the effect of noise on feature values.

Chapter 6

Investigation of Tumour-Specific Noise Equivalent Counts

The modelling to this point has considered a measure of noise which applies on a global level. When comparing this metric to global texture for long duration acquisitions, even features that are not classically 'robust' can be compensated for. However, we know that our assumptions break down when considering small regions within the overall image, and that NECR is a weak proxy for image SNR when iterative reconstruction is used. Here we investigate whether a new metric measuring 'local noise' could be established. Correlations between this metric and a feature could enable a new model for correcting features for clinical data. It should be desirable for this noise measure to be easily implementable by a clinician or research scientist using easily accessible information.

(a) Frame 16 of *T1+T3* (L-R, T3 and T1)

(b) Frame 16 of *Only3*

Figure 6.1: Slices from PET images of *T1+T3* and *Only3* in coronal view[22].

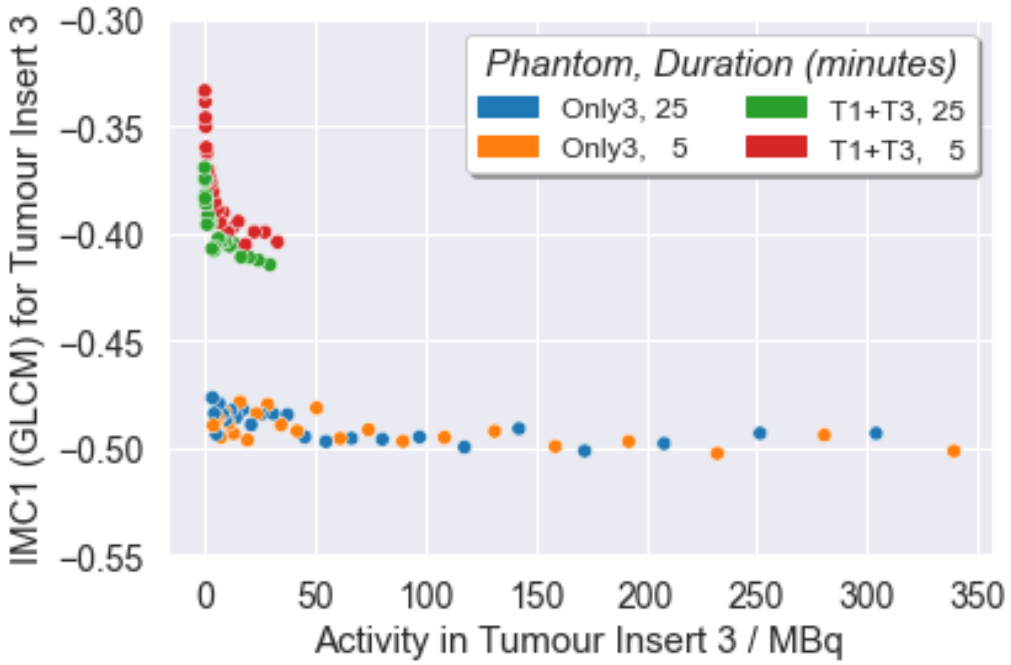


Figure 6.2: GLCM IMC1 for region T3 in both *T1+T3* and *Only3* scans plotted against the activity in the T3 region. 5 and 25 minute data are plotted separately.

Two scans were performed using the same phantom-insert setup. Using the *T1+T3* phantom setup but filling only *T3* with activity[23] a further 12 hour scan series was taken. Example PET images of these two configurations can be seen in Figure 6.1. Figure 6.2 shows one feature (the GLCM IMC1) for T3 across both scans, labelled *T1+T3* and *Only3*, considering only the activity contained in insert T3. For 25 minute *Only3* scan data there is evidence of NECR correlation for IMC1, with $|\text{PMCC}| = 0.7281$ (falling to 0.6540 when considering 5 minute data). Evidence of this NECR correlation may be obscured on the figure shown due to choice of y-axis display, however the figure also demonstrates a weakness with using NECR correlation to determine a feature's robustness. Compared to the values from *T1+T3*, the *Only3* IMC1 measurements appear more consistent, as any noise level correlation has a much smaller percentage effect on the value it takes. Whether this indeed makes the feature more robust depends again on the

[22]The *coronal* view corresponds to slicing the image as if going from the tip of the nose to the back of the head (y-axis by convention). Figure 5.4 shows examples of images sliced *transaxially* along the z-axis.

[23]This new scan can be described as a 'hot' insert on a 'cold' background.

possible values that the feature may take.

The relative consistency of the *Only3* feature values could be attributed to image reconstruction convergence. When using iterative reconstruction algorithms, images often converge after a number of updates, showing little appreciable change to the image as further updates are used. Should an excessive number of algorithm updates be used, the noise in the image will become amplified. It is therefore important in practice to use an appropriate number of updates to achieve acceptable convergence without unnecessarily emphasising the noise. The appropriate number of updates depends on the algorithm used, as some point-spread function modelling procedures can severely reduce the effect of noise amplification due to excessive updates. It also depends on the patient or subject being scanned. The feature values from *T1+T3* are taken from images where a similar total activity is distributed over a larger volume than the *Only3* equivalent. It is known that the OSEM algorithm converges faster for regions of high 'uptake' [38]. It could hence be argued that, as the two image sets are reconstructed using identical algorithms with identical numbers of OSEM updates, that the *Only3* images reach convergence faster and thus may exhibit more consistent values of noise-affected features across the set. This argument is dependent on whether it can be accepted that 42 OSEM updates (2 iterations of 21 subsets) is sufficient to reach acceptable convergence for the *T1+T3* image set. As these settings are recommended by the manufacturer, further work is required to determine the point of acceptable convergence. This would be achieved by reconstructing the same two datasets with a larger variety of protocols, and performing other scans with the other isolated phantoms *T1, T2,* and *T4*.

It can be assumed that the ground truth distribution for activity within T3 is equivalent between the two experiments[24]. Consequently the difference between the values of IMC1 in *T1+T3* and *Only3* can be attributed to noise contributions from activity outside *T3* itself in the *T1+T3* scan. In obtaining texture features, it should remain the goal to obtain as 'true' a value as possible, and evidence suggests that feature values become more reliable (or, at least, more robust with activity levels) when observing the hot region in isolation. Here we investigate the

[24]The ROIs for T3 had to be redefined between scans due to misalignment, but the exact process was replicated for ROI definition and the resultant ROIs contained an equivalent number of voxels.

possibility of obtaining a measure of 'local' noise, such that texture feature values could be corrected to more robust and desirable levels. Such a measure could be used to improve the use of texture features when looking at small ROIs, such as those used in oncology PET, where radiomics work seeks to make the biggest impact. One area of research of particular concern is Total Body PET and longer axial FoV scanners (see Section 8.3), where much higher-activity volumes such as the brain and bladder will be in the FoV at the same time as the lesion, introducing further sources of noise. The ideal measure can be thought of as a 'tumour-specific' noise-equivalent count rate, and shall be referred to as such henceforth. While this is a novel methodology, there is precedent in previous work which determined to establish a patient-specific NECR for dose optimisation [97].

As such, the whole image NECR is a weak indicator of lesion noise. The metric itself has great value due to its resemblance to a traditional definition of signal-noise ratio and, as seen in the previous chapter, exhibits strong correlations to some image heterogeneity measures. Nonetheless, its direct equivalence to noise as seen in an image is tenuous unless image reconstruction is done analytically and without corrections. In correcting the collected data, performing image reconstruction and post-hoc smoothing and/or adjustment, noise is not removed from an image but the level of noise will no longer relate as strongly to the ratio of false coincidences that are deduced by the scanner.

We require a measure of the level of noise within a local ROI, yet this will be affected in no small part by the activity present globally. A successful attempt should seek to match the values for T3 between *T1+T3* and *Only3* scans. The following sections describe methods employed to define the metric, beginning with the most naïve assumptions and adding complexity step by step. The models are evaluated and discussed in the final section of the chapter.

6.1 Scaled Method

The initial naïve attempt adapted the current definition of NECR (seen in Equation 5.4) to express the three terms T, S and R in terms of the activity local to the

insert, such that

$$T_{\text{insert}} = \frac{A_{\text{insert}}}{A_{\text{tot}}} \times T_{\text{tot}}, \tag{6.1}$$

$$S_{\text{insert}} = \frac{A_{\text{insert}}}{A_{\text{tot}}} \times S_{\text{tot}}. \tag{6.2}$$

and

$$R_{\text{insert}} = \frac{A^2_{\text{insert}}}{A^2_{\text{tot}}} \times R_{\text{tot}}, \tag{6.3}$$

This approach was labelled the *Scaled Method*. The values of this 'new NECR' will therefore be unchanged from standard NECR for the *Only3* scan. The values of this new Scaled NECR for *Only3* and T3 from *T1+T3* can be seen in Figure 6.3.

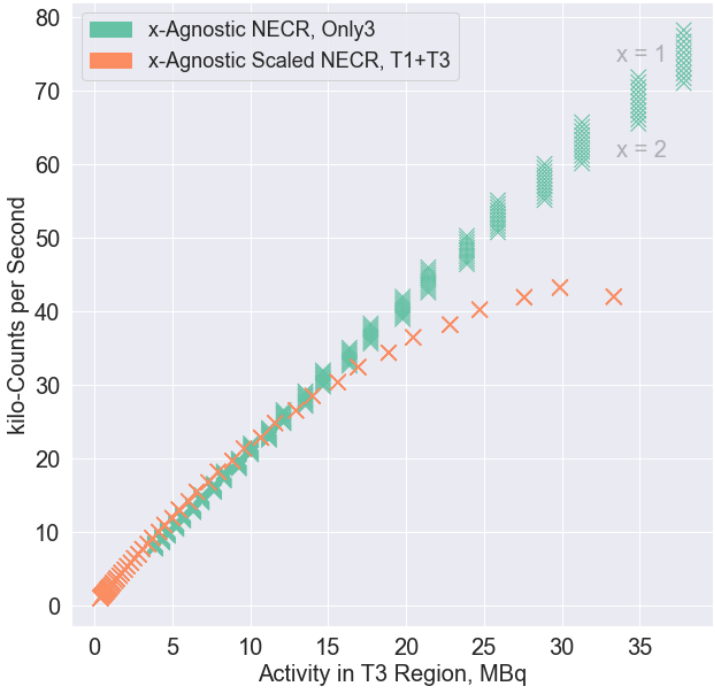


Figure 6.3: A plot comparing NECR from *Only3* against Scaled NECR from *T1+T3*. The y axis compares units of the two metrics in kilo-counts per second (kcps).

In Strother & Casey (1990) it is laid out that our standard NECR takes the form

$$\text{NECR} = \frac{T^2}{T + S + x \cdot R}, \tag{6.4}$$

where x is defined such that $1 < x < 2$ as a randoms estimation method-dependent constant [96]. Using a delayed window method, NECR is quoted as standard with

$x \approx 2$. For the new method, due to the ambiguity of the approach, the assumptions on the value x should hold were relaxed. The definition of NECR is a pragmatic one, with factors of $x = 2$ and $x = 1$ used for delayed window and singles-based randoms estimation due to the perceived effect this would have on the number of random counts returned (as discussed in Section 2.5.2). By considering further fractional apportionment of the random counts on a tumour-specific level, it was decided to loosen these assumptions and observe the effects of changing the x parameter. As such, the new method was 'x-agnostic', and the new Scaled NECR calculated for all x where $x \in \{1.0, 1.1, 1.2 \ \ldots \ 1.9, 2.0\}$. Figure 6.3 shows the x-agnostic calculation of the Scaled NECR for *Only3* and for T3 from *T1+T3*.

6.2 Developing Spatially-Aware Methods

Consideration must be paid to whether spatially-dependent considerations should be given to R, S and T. For the true coincidences T, the argument is less convincing. By definition a true coincidence requires the line of response to be triggered regardless of location and matter present; the naive model therefore fits this well. The arguments behind spatially modelling S are more complex. One approach could be to determine the likelihood of particular LORs being triggered in relation to the underlying attenuation map, using CT scans or μ-maps already available from the PET-CT scan. The complexity, however, originates from where the LORs from scattered coincidences appear in relation to where the scattering happens. The formula for Compton scattering,

$$E_{\gamma'} = \frac{E_\gamma}{1 + \left(\frac{E_\gamma}{m_e c^2}\right)(1 + \cos(\theta))}, \tag{6.5}$$

relates the photon energy pre- and post-scattering event $(E_\gamma, E_{\gamma'})$ to the angle of scatter θ. The energy window of the Biograph mCT is set at $435 - 650$ keV, and part of the reason the window is set as such is to reduce the proportion of scattered counts that are accepted. Despite this, using the lower bound of the energy window, any scatter up to an angle of $\sim 146^\circ$ would still be considered legitimate. This does, albeit by a small proportion, reduce the pool of potential

LORs that could be triggered by scattered coincidences, preventing illumination of LORs between closely neighbouring detector elements. The resultant distribution of scatter in projection space will result in characteristic 'tails' of triggered LORs outside the scanned object - alluded to in Section 2.5.1 - which is used to construct the whole scatter distribution. By masking this distribution in projection space using the ROI sinogram as before, it should be possible to determine the amount of scatter noise within the ROI.

In practice this information was problematic to obtain. It was, however, possible to reconstruct images without scatter correction, using sinograms taken only after randoms subtraction. Reconstructing the image with exactly the same protocol but with scatter correction disabled, and then subtracting the scatter-corrected image from this, provided a 'map' of the origins of scattered counts; or, more precisely, the counts classified by the scatter correction algorithm as likely to be due to scatter. This scatter subtraction image was then radon-transformed and masked with the sinogram of the T3 ROI. This process is illustrated in Figure 6.4 as it was performed for the first 25 minute frame of *T1+T3*. Using the relative sums over the matrices, it was shown that $\sim 11\%$ of the total counts determined to be scatter triggered LORs intersecting the ROI of T3. By comparison, the fraction of activity $A_{\mathrm{T3}}/A_{\mathrm{tot}}$ was calculated to be ~ 4.5 %, indicating that the Scaled and initial Spatially Aware models possibly underestimate the level of scatter noise.

The fraction of scatter estimated from this image-based scatter estimation method was incorporated into the the Scaled method, and the x-agnostic version of this new metric, labelled *Spatially Aware* is plotted in Figure 6.5.

6.3 Model Evaluation

The success of Tumour Specific NECR models depends on the ability of the model to obtain the values of radiomics features as they would appear in an isolated version of the ROI. This aim is complicated by a requirement to extrapolate beyond the data obtained. It is proposed that, should a feature correlate well with the Tumour Specific NECR and exhibit a peak or plateau as observed in those curves, the correction could be performed on this feature value. Consequently it was important to establish whether feature correlations are improved by using

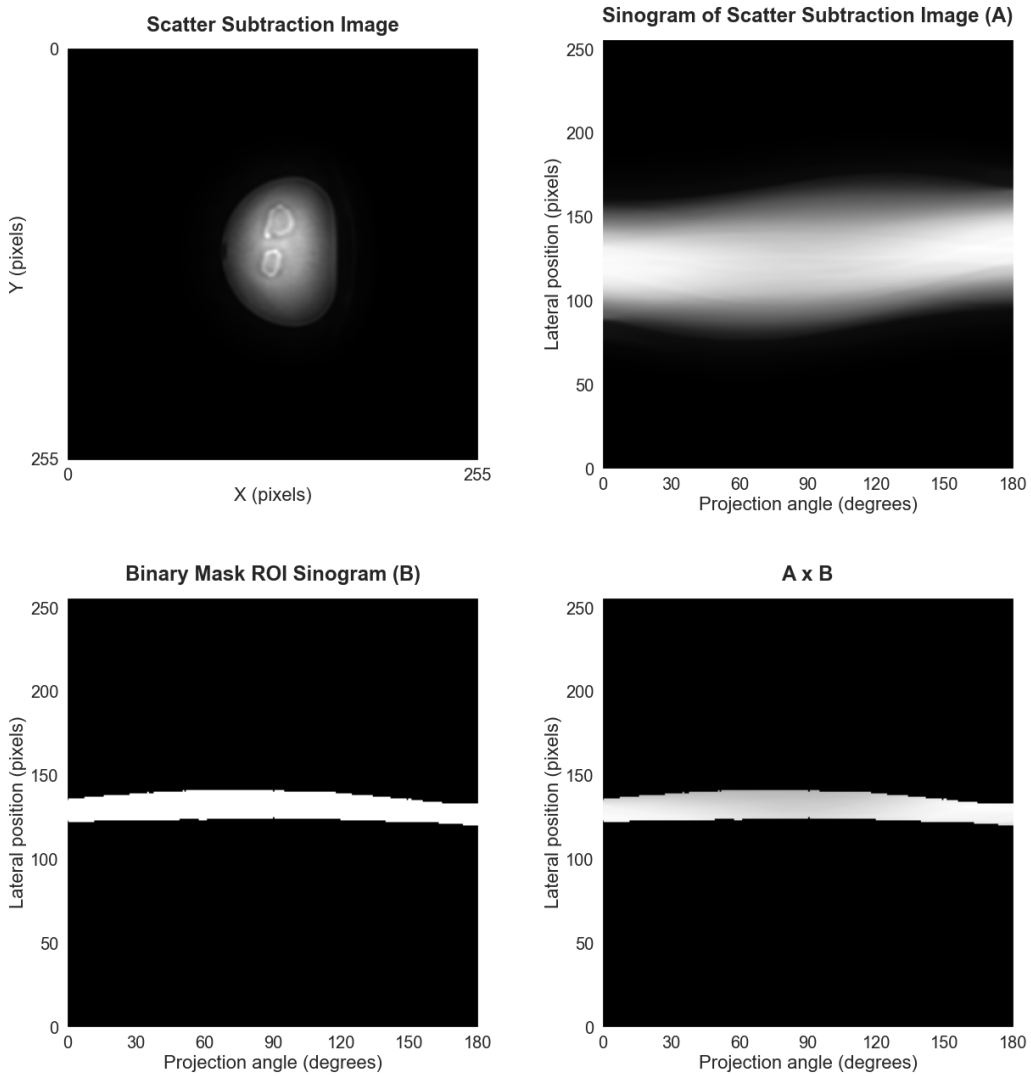


Figure 6.4: Figures showing the method of local scatter estimation. Top left: the scatter subtraction image. Top right: the sinogram of the scatter subtraction image. Bottom left: a mask of the T3 ROI sinogram. Bottom right: the product of the mask sinogram and the scatter subtraction sinogram.

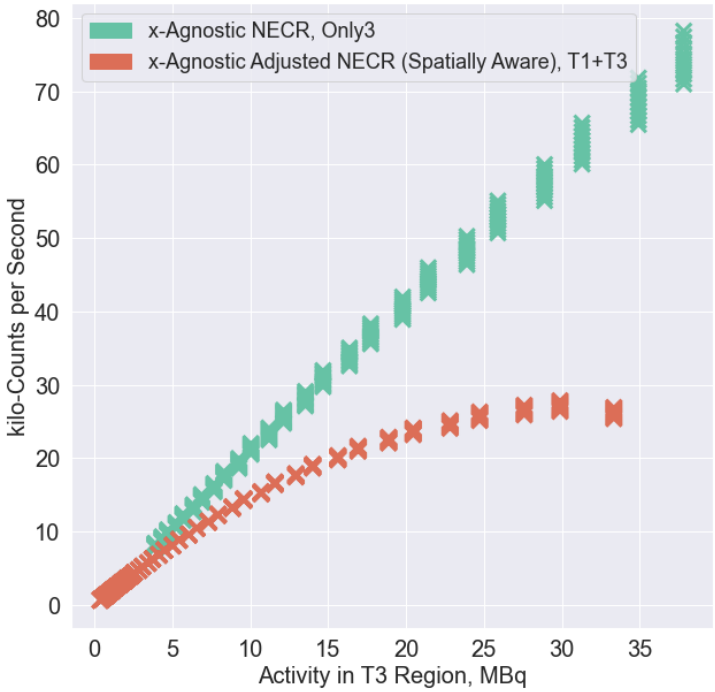


Figure 6.5: A plot comparing the adapted spatially aware method, correcting R and S, against *Only3* NECR.

the Tumour Specific NECR over the global NECR. |PMCC| was calculated for all texture features against the three models proposed in the previous sections; *Scaled*, and *Sp. Aware* (with sinogram-based S adjustment). The data used was the T3 ROI from the 25 minute acquisitions of *T1+T3*. The results for features with prior weak NECR correlations are shown in Table 6.1.

For the 14 features included in this subset, 9 show improved |PMCC| using the developed models when compared to general NECR. The spatially aware models showed better correlations than the naive scaled models for 9 of the 14 features, justifying the decisions made in this work. The features selected were those which previously correlated weakly with NECR, showing promise that using image-based methods could help improve models for localising PET noise. The models do not significantly improve texture |PMCC| over 0.5. These are still generally poor correlation strengths, and associated p-values indicate that there is still a likelihood of no correlation.

The models generally do not improve correlations for previously strongly-

Feature	Absolute PMCC for 25 minute data			Do Models Show Improvement?
	Global NECR	Scaled	Sp. Aware	
JointEnergy (GLCM)	0.48	0.28	0.26	No
RunLengthNonUniformity (GLRLM)	0.42	0.26	0.25	No
IDM (GLCM)	0.37	0.48	0.48	Yes
SmallDependenceLowGrayLevelEmphasis (GLDM)	0.37	0.50	0.51	Yes
ShortRunLowGrayLevelEmphasis (GLRLM)	0.32	0.46	0.47	Yes
LowGrayLevelRunEmphasis (GLRLM)	0.32	0.46	0.47	Yes
LowGrayLevelEmphasis (GLDM)	0.30	0.45	0.46	Yes
LongRunLowGrayLevelEmphasis (GLRLM)	0.30	0.45	0.46	Yes
ShortRunEmphasis (GLRLM)	0.30	0.15	0.13	No
LargeAreaLowGrayLevelEmphasis (GLSZM)	0.28	0.12	0.10	No
SmallAreaLowGrayLevelEmphasis (GLSZM)	0.28	0.42	0.43	Yes
LowGrayLevelZoneEmphasis (GLSZM)	0.26	0.40	0.41	Yes
RunLengthNonUniformityNormalized (GLRLM)	0.21	0.05	0.04	No
LargeDependenceLowGrayLevelEmphasis (GLDM)	0.01	0.17	0.18	Yes

Table 6.1: Table showing the correlation of texture features to Tumour-Specific NECR models. Features chosen are the features with global NECR $|\text{PMCC}| \leq 0.5$.

correlating features. Only 12 features out of 75 exhibit enhanced correlations using these models. The three not listed in Table 6.1 are listed in Table 6.2. These three features show potential benefits of using these local image-based techniques on even previously strongly-correlating features.

Feature	Absolute PMCC for 25 minute data			Best Model
	Global NECR	Scaled	Sp. Aware	
Correlation (GLCM)	0.72	0.90	0.91	Sp. Aware
InverseVariance (GLCM)	0.74	0.78	0.77	Scaled
MCC (GLCM)	0.75	0.85	0.85	Sp. Aware

Table 6.2: Table showing the correlation of texture features to Tumour-Specific NECR models for three high-performing examples of successful model implementation.

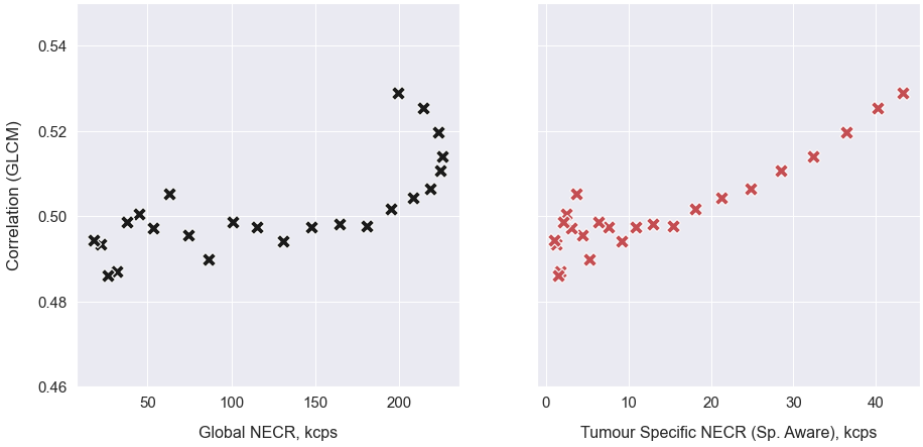
Figure 6.6 shows three example texture features plotted against global NECR and the *Sp. Aware* model of tumour specific NECR. The figure shows the benefits of enhanced linearity (GLCM Correlation) and diminished linearity (GLDM Small Dependence High Gray Level Emphasis) when using the tumour specific model. However, the GLDM Small Dependence Low Gray Level Emphasis is an example of a feature where 'enhanced correlation' with a tumour specific model of NECR may be disadvantageous. The feature exhibits arguably robust behaviour with increasing NECR, with a value that could be said to be constant at 0.006 ± 0.002. Applying tumour specific modelling may draw out false correlations to already-robust features, which not only may be redundant but also

misleading. There is a percentage uncertainty of 31 % in the value of the GLDM Small Dependence Low Gray Level Emphasis, and conclusions on the feature's robustness can only truly be made when using more heterogeneous activity distributions as part of the examination. In the homogeneous phantoms, it is expected for the low and high gray levels to be sparsely populated, and the low statistics may lead to discrepancies in the results of extracted feature values.

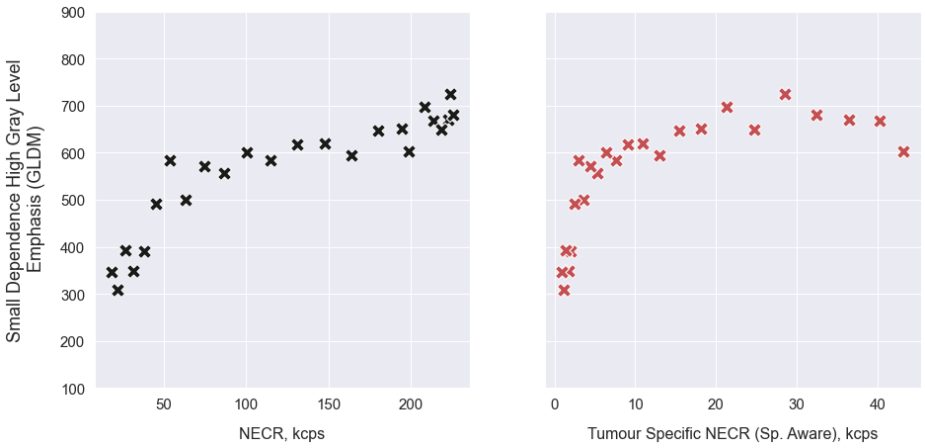
6.4 Conclusion

This chapter has detailed how models could be used to estimate localised noise in PET images. There is evidence to suggest that these models could improve correlation with texture features in comparison to regular global NECR. The model definitions are simple and could be easily incorporated into future image analysis with little additional information needed. However, the models developed to this point are not sufficiently sophisticated to improve poorly-correlating features to a level at which correction could be achievable. This is for two competing reasons. Fundamentally these local models do not account for the effects of activity in the entire FoV, which are inextricable from the noise experienced locally. More complex and considered models are required in order to attribute the level of noise into more localised regions. These models not only need to consider the fractional components of T, R and S but also the effect on each of those rates due to the environment in which they are scanned, doubtless incorporating some manner of detector dead time modelling. In addition, if a feature exhibits poor robustness with activity, with randomly-distributed values, it is unlikely that any manifestation of a tumour-specific NECR could lead to value correction. It is important to state that poor values of |PMCC| in this work are not necessarily indications of poor robustness to activity level, but rather an indication that their robustness is difficult to define without further examples of heterogeneous distributions included in the study. Nonetheless, this work has formed foundations that could be built on with further modelling; the evaluation is yet to investigate whether value correction using this method is achievable or reliable, as this would require many more scans of isolated tumour models. Further work should aim to repeat studies such as this on a wider phantom set.

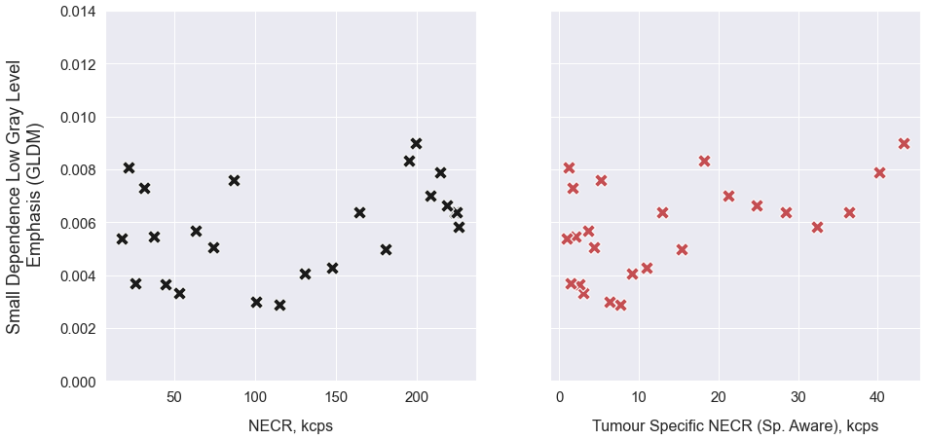
Further development of a localised NECR is likely to become more necessary with the advent of Total Body PET and long axial FoV scanners. With more activity in the scanner from sources such as the brain or bladder, it is more likely for ROIs to be affected by noise caused by activity in these regions. For more on the effects that Total Body PET may have on this work, refer to Section 8.3. Furthermore, as future AI-based image reconstruction and denoising becomes more widely used, the separation between SNR_{data} and SNR_{image} will only become more significant.



(a) GLCM Correlation: LHS $|\text{PMCC}| = 0.7153$; RHS $|\text{PMCC}| = 0.9113$



(b) GLDM Small Dependence High Gray Level Emphasis: LHS $|\text{PMCC}| = 0.8895$; RHS $|\text{PMCC}| = 0.7925$



(c) GLDM Small Dependence Low Gray Level Emphasis: LHS $|\text{PMCC}| = 0.3653$; RHS $|\text{PMCC}| = 0.5079$

Figure 6.6: Scatterplots demonstrating the differences in modelling three radiomics texture features against global NECR and the adjusted NECR using the *Sp. Aware* model.

Chapter 7

A Monte Carlo Simulation Approach

There are many advantages of supplementing experimental data with simulated data. The definition of an object in simulation software provides an exactitude that is not trivial even with physical phantoms. The extra data is of a benefit especially where scanning time is at a premium – at The Christie, the scanner can only be used for research when not in clinical use. Furthermore, and most notably when simulating to mimic a patient scan, the ethical concerns regarding repeated scanning or unnecessary scanning of human subjects do not apply to the simulations. Simulation is widely used in medical imaging in both research and clinical fields. Any new detector technology must first be rigorously modelled before beginning the highly expensive process of development, while in the clinic there is already widespread adoption of simulation techniques for scatter correction and radiation dose modelling.

This chapter details work undertaken in establishing and validating PET simulations of the Siemens Biograph mCT TrueV, henceforth referred to as the mCT. This is followed by a simulation validation study, demonstrating difficulties associated with reconstructing simulated image data to the degree of accuracy required for implementing into reliable image studies. This validation study was performed earlier in the PhD in order to validate the use of offline image reconstruction, and was presented to the British Nuclear Medicine Society in 2021.

7.1 PET Scan Simulation

7.1.1 Defining a GATE Simulation of the Siemens Biograph mCT

Simulations were developed using GEANT4 software (**GE**ometry **AN**d **T**racking) on the GATE (**G**EANT4 **A**pplication for **T**omographic **E**mission) open-source framework [114]. A simulation of the geometry of the mCT system was created, implementing the dimensions and important coincidence-processing criteria used on the real scanner. Some of the most important features are highlighted in Table 3.1.

GATE provides a number of templates by which a PET geometry can be described, two of which are appropriate in this context: *cylindricalPET* and *ECAT*. The two templates differ only in the syntax used to describe all of the detector gantry parts, and in the output formats provided. The syntax difference is most evident in the labelling of the crystal detector elements, as each element is described by the hierarchy used to define it. In cylindricalPET, the crystal element is labelled by its radial sector (rsector), the module within that sector, the block within that module and finally the crystal within that block. The ECAT template, however, allocates each crystal element just a global block number and a crystal number within each block.

ECAT takes its name from the line of PET scanners originally built by CTI, the US-based company subsumed by Siemens in 2005. The legacy of this line of PET scanners remains in the file format used by sinograms from subsequent Siemens PET scanners, and the benefit of the ECAT description in GATE is the ability to give output in this sinogram format. In addition, both cylindricalPET and ECAT enable list-mode output in ROOT, CERN's own data platform for visualisation and storage, while cylindricalPET provides a further plain text list-mode output option. The mCT was recreated using both of these templates.

A visualisation of the mCT PET detector geometry simulation can be seen in Figure 7.1, here described using the cylindricalPET template. In reality, the scanner comprises of additional elements. The scanner bed, the largest non-subject

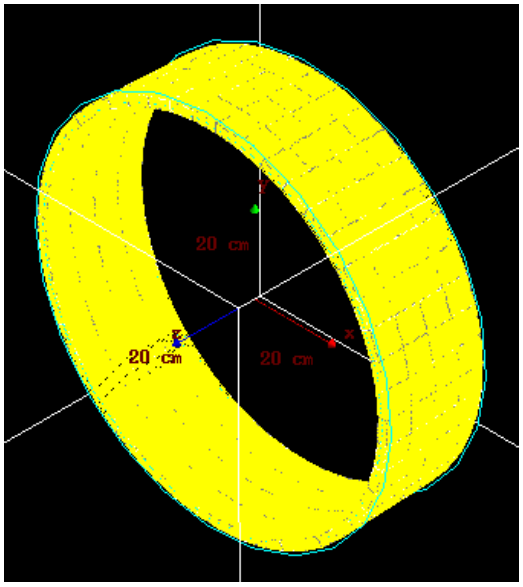


Figure 7.1: A visualisation of the GATE-simulated mCT PET detector gantry. The crystal blocks (yellow) can be seen, demonstrating the modular structure. The white gridlines bound the extent of the scanner, and axes illustrating dimensions can also be seen.

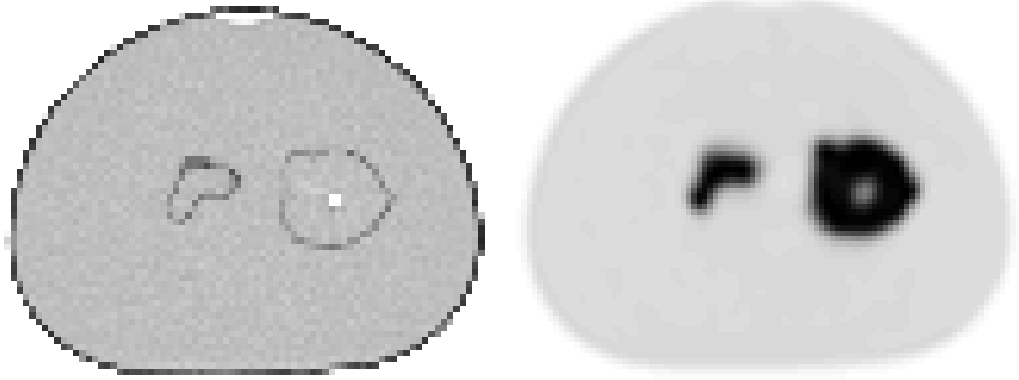
source of attenuation and scatter, was included in the simulation by describing the attenuation map from a CT scan of the entire bed. In the real scanner, there are two rings of lead forming septa, 'capping' the cylinder. This is to prevent scattered photons from entering the detector from outside the FoV. The inclusion of these septa was debated, as scatter from outside the FoV due to the bed was possible, yet the simulations only considered activity distributions defined entirely within the scanner's axial FoV. For the purpose of simulation speed, these were omitted. Also omitted were the plastic detailing and casing structures around the PET ring, as the additional attenuation was assumed to be small. Additional features such as the CT gantry, which serve no purpose to the simulation of PET acquisition, are also omitted in these simulations. These are features which can be added later should the project require a fully realistic clinical recreation of a PET scan, but carry little relevance to the acquisition of coincidences required at present.

Simulations were run using the cluster computing facilities in the Nuclear Physics group at the University. It became apparent that sinogram output from the ECAT-defined structure would not be compatible with running the simula-

tions on this framework, so the cylindricalPET system was selected for further development.

7.1.2 Creating GATE Phantoms and Sources

In GATE, phantoms and sources can be defined using either standard geometrical definitions used in Geant4 or image-based methods. While the former enables faster running time for simulations, the basic geometric shapes are inadequate to perform realistic simulations of clinical scenarios. Using image-based or 'voxelised' definitions, the materials describing the sources and phantoms can be implemented on a level approaching realistic granularity.



(a) CT (b) PET

Figure 7.2: A transaxial slice of the real *T1+T3* image in PET and CT, discretised and cropped. The CT image was used for the GATE phantom definition, and the PET image for source definition.

Phantoms and sources were defined for the *T1+T3* and *T2+T4* insert arrangements using the CT scan and the first 25 minute PET frame for each scan series. First, these images were registered to the same matrix size. Using ImageJ, all images were discretised, converted to 16-bit and cropped to include only the extent of the phantom's outer shell using the same coordinates on both PET and CT. All images were used in Interfile format, with headers formatted in a style compliant with GATE's required definitions.

In order to simplify the simulation, air, water and 'plastic' were the only three materials defined for the phantom, with the physical properties of the latter taken

from those defined as standard in GATE for Perspex. The source was defined with only two activity concentrations, that of the background and the insert. The ratio between these was kept identical to that in the physical scan, and the exact per-voxel concentration levels defined to provide a total FoV activity. The simulation was to be run with $A_{\text{tot}} \in \{50, 100, 150, \ldots, 700\}$ MBq, therefore the 14 activity reference files were defined separately and were accessed using GATE's macro-based commands.

7.1.3 Utilising GATE Output Data

The GATE output data was given in ROOT format. ROOT files store data in a hierarchical structure of *trees* and *branches*, and was developed by CERN in the 1990s [115]. Coincidences as detected by GATE are stored in a *Coincidences* tree in listmode format. Source position in (x,y,z) and detected (global) position in (x,y,z) are recorded for each photon in the coincidence, along with the labelled crystal and block element of the detected single photon as described by the cylindricalPET protocol. Also stored is the detected time and energy of each photon, and whether either photon underwent Compton or Rayleigh scattering between emission and detection.

This raw count rate information is incredibly important, and validation of this output will enable interpretation of whether the GATE infrastructure can be reliably used to mimic real detectors, yet obtaining reliable coincidence information is only the first stage of an elaborate multi-step process to recreating reliable data. Performing the various data correction and image reconstruction processes upon the collected data in the exact same way that equivalent processes would run on the scanner is not a solved problem, owing to the multi-levelled complexity brought about by the various scanner manufacturers' proprietary algorithms and software, which are being constantly updated to remain on the cutting edge of the field. External to the manufacturers, a number of different collaborations have instigated open-source software for image reconstruction intended for research purposes. The most popular of these open-source software are STIR (Software for Tomographic Image Reconstruction) [116] and CASToR (Customizable and Advanced Software for Tomographic Reconstruction) [117]. Two approaches

become viable for reconstructing simulated image data: manipulating the GATE ROOT output into a format that can be readable by the scanner manufacturer's software, or replicating the performance of the scanner manufacturer's software using open-source tools.

Siemens provide an offline version of their image reconstruction software, traded as *e7-tools*. Included in this software are all of the tools required to take raw data from the scanner and reconstruct using the same algorithms on a Windows desktop in a separate research facility. The principle issue with implementing ROOT data into the e7-tools framework lay in the restructuring of the raw data. Siemens use a strategy of 'virtual crystals' for data storage with the mCT, creating fake crystal element labels for inter-block and inter-ring gaps[25]. In addition, the use of ToF information is not well understood outside of scanner-specific convention, evidenced by the inclusion of ToF in only the most recent updates of both STIR and CASTOR.

Time was spent over the course of this PhD working with members of the team at STIR in efforts to implement the virtual crystal methods for the mCT GATE simulation into the framework. This process was ultimately successful, with the mCT becoming included as part of the STIR-GATE Connection [118] software package. Further work should examine this software and its suitability for reconstructing images with ToF, as part of a wider validation study reconstructing images with algorithms approximating those used in Siemens UHD reconstruction.

7.1.4 Validating the Simulation

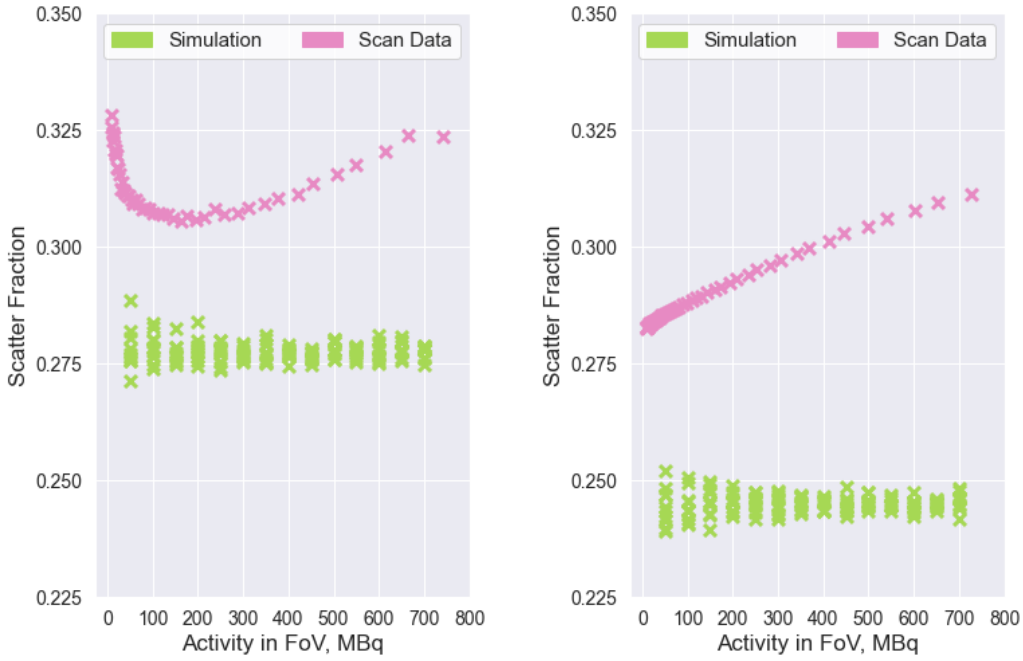
Simulations were run for *T1+T3* and *T2+T4*, replicating one second of simulated time. These were repeated ten times each, with each second of simulation lasting 17 hours of computation time for the 50 MBq acquisition. Count rate statistics were established directly from the ROOT files, categorising a coincidence as scattered if at least one of the Compton entries in the entry was non-zero, and categorising a coincidence as random if the source position of the two photons

[25]These gaps are formed unintentionally when attempting to fit together the crystal blocks with the attached electronics and framework.

was not equal. From here, the scatter fraction was calculated using

$$f_S = \frac{S}{T+S} \tag{7.1}$$

to replicate the quoted value as appears in the Siemens DICOM header files. Figure 7.3 shows the values for the two simulations against the values determined from the physical scans.



(a) *T1+T3* Scatter

(b) *T2+T4* Scatter

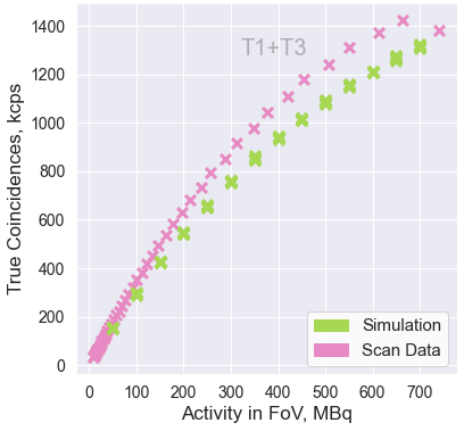
Figure 7.3: Plots showing the scatter fractions calculated from GATE simulations of *T1+T3* and *T2+T4* alongside the corresponding physical data.

By the definition of what is expected from a scatter fraction, the value should be constant regardless of the activity level, as it should depend solely on the attenuation map and geometry of the subject. This is reflected by the simulation, giving values of (27.8 ± 0.2) % and (24.5 ± 0.2) % for the *T1+T3* and *T2+T4* simulations respectively. It is already known, however, that there is a contribution to the estimated scatter from randoms for the scatter fraction established from the scanner software, principally evidenced by the linear gradient tended to by the scatter fraction for the physical scan data. This simulation demonstrates,

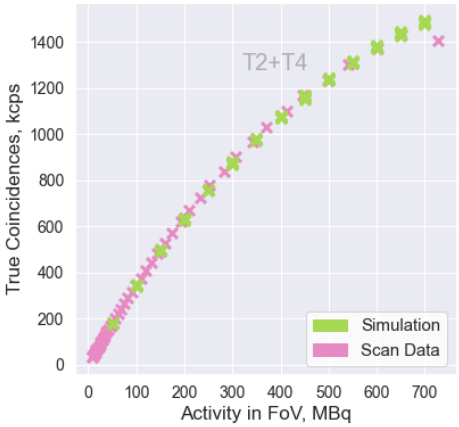
however, that there is an consistent overestimation of the scatter due to randoms categorisation even at low activities tending towards zero.

Figure 7.4 shows the count rates T, S and R for the simulated and real datasets for the *T1+T3* and the *T2+T4* scans. Subfigures 7.4a and 7.4b show perhaps the clearest indication of the accuracy of the simulation in terms of simple architecture. It is clear that the *T2+T4* simulation models received true coincidences remarkably well for all FoV activities, while the *T1+T3* attempt routinely underestimates received trues. Most enlightening are the S and T rates; using the raw simulation data, the scatter is routinely drastically underestimated, while the randoms appear grossly overestimated. There are several potential reasons for this. Firstly, as has been established priorly, the physical scanner will conflate a proportion of randoms into the genuine scattered counts. Secondly it is possible that dead time modelling in the simulation may not have been well aligned to the scanner's system. This is illustrated by the simulations appearing to match the scattered and random coincidences well for low activities but diverge as further activity is added.

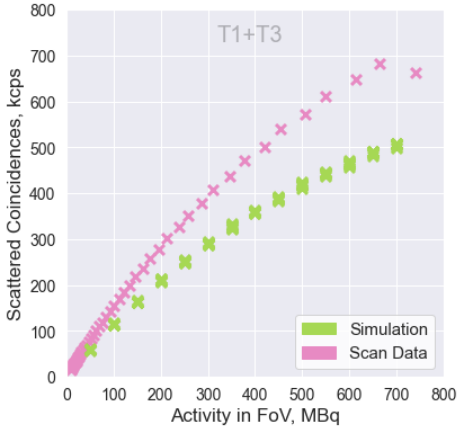
While the differences between real and simulated systems cannot be ignored, it should be evaluated whether the differences between R and S complement each other. The standard definition of NECR was calculated from the count rates, setting $x = 1$ for the simulation data due to the direct calculation of randoms, and the values are shown in Figure 7.5. This figure shows the difficulty in using the standard definition of NECR to validate the simulation, due to the pragmatic nature of the original NECR definition. Nevertheless, at lower activities there is a good match between simulated and scanner NECR using these definitions. This suggests that the principle issue is in overestimation of the randoms, as the simulations show poor agreement once the randoms exceed beyond the scale of the trues and scatter. This reinforces the prior assumption that dead time may not be appropriately modelled. In order to improve this work, it is imperative that simulated data be processed in the same manner as the physical data. This was made difficult by the lack of time of flight provision until recently at the time of writing, yet the simulation definition and development will provide useful material for further work in this area.



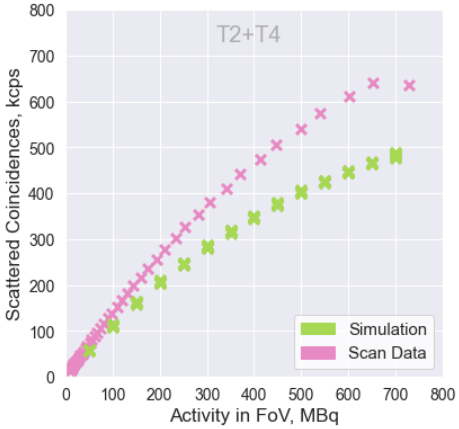
(a) *T1+T3* Trues



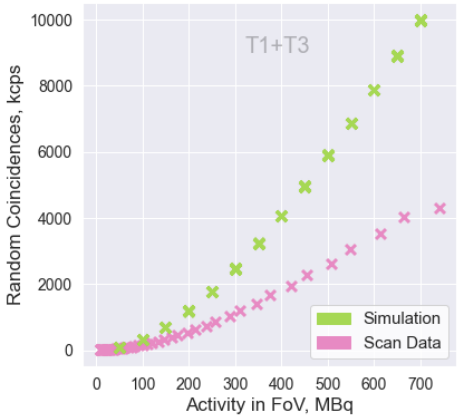
(b) *T2+T4* Trues



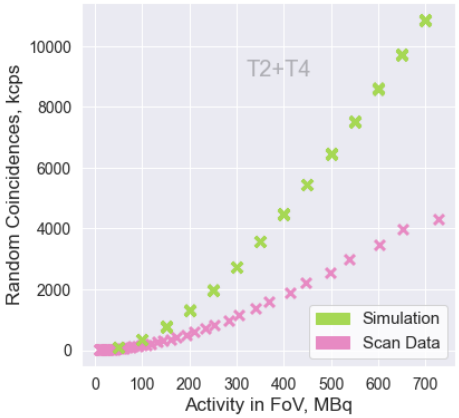
(c) *T1+T3* Scatter



(d) *T2+T4* Scatter

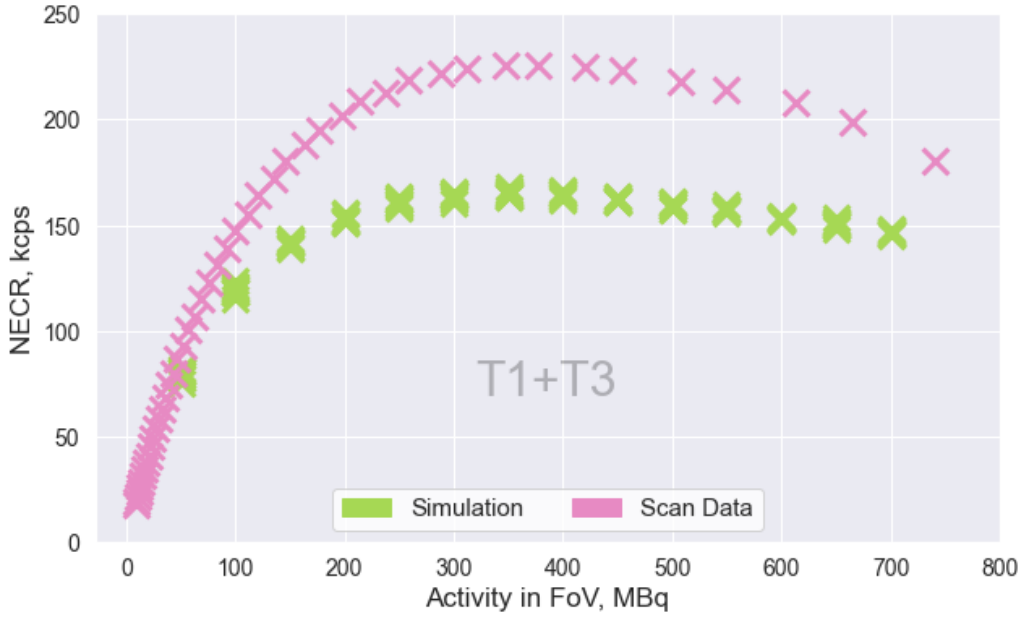


(e) *T1+T3* Randoms

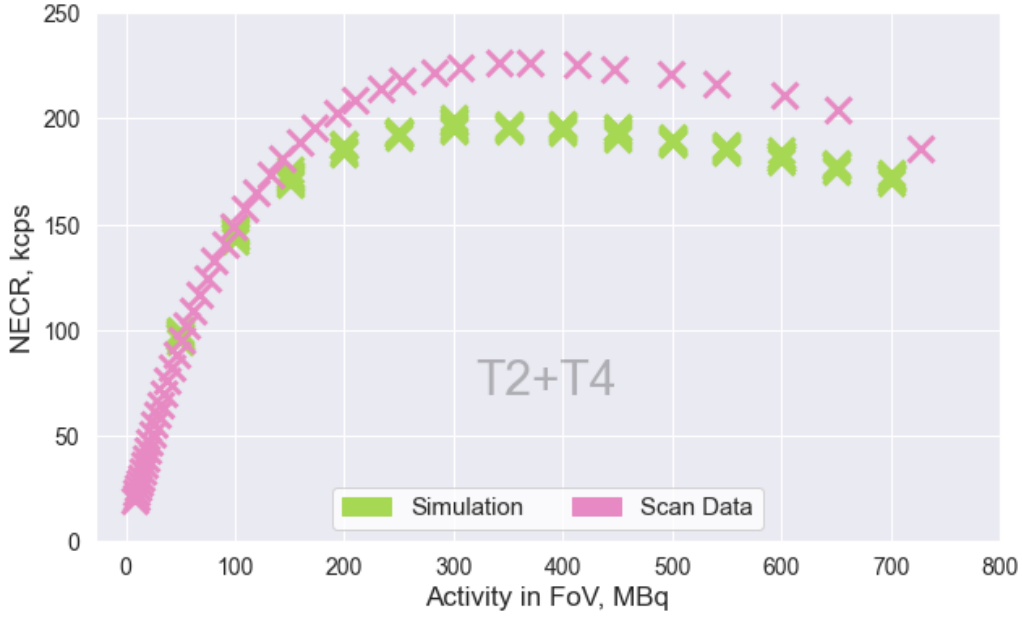


(f) *T2+T4* Randoms

Figure 7.4: Individual rates T, S and R for the simulation data (green) compared to the physical scan data (pink) for *T1+T3* (left) and *T2+T4* (right).



(a) *T1+T3*



(b) *T2+T4*

Figure 7.5: Plots showing the NECR calculated from the physical and simulation data for *T1+T3* and *T2+T4*. The NECR is calculated with $x = 1$ for simulated data and $x = 2$ for physical data.

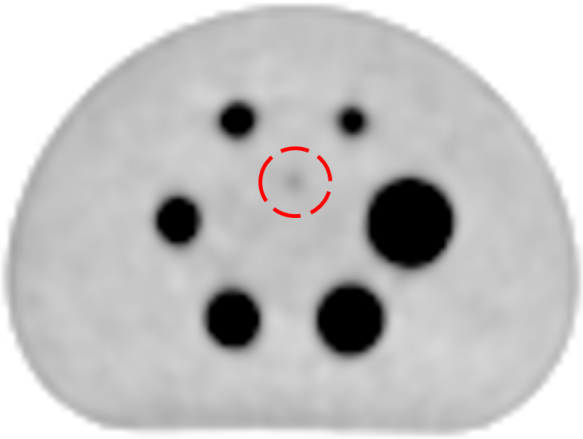


Figure 7.6: A PET image slice of the ^{68}Ga NEMA phantom. The seventh sphere, diameter 5 mm, is located by the red circle.

7.2 Validation of Image Reconstruction Software

In order for meaningful comparison between images, a full validation must be carried out between every software used in the work. As a way of demonstrating potential problems in these studies, the following subsection details an experiment performed to validate the use of Siemens e7-tools offline image reconstruction software.

A Validation Study of Siemens e7-tools

A facsimile of the NEMA Image Quality phantom was created using a ^{68}Ge-infused epoxy resin for a wider intercomparison project [119]. The standard phantom features six spheres of decreasing diameter from 37 mm to 10 mm; this version adds a seventh sphere of diameter 5 mm. For the purposes of this task, only the six largest spheres were considered. At the time of imaging, the phantom exhibited an activity of 22.9 kBq/ml in the seven spheres, and 5.7 kBq/ml in the background.

Ten single-bed position acquisitions were taken to create a core raw dataset, each of 111-seconds. This duration was chosen to be consistent with the parallel intercomparison project [119]. These frames were reconstructed using the scanner's on-line program with all permutations of:

- 1, 2 and 3 iterations of 21 subsets of;
- standard OSEM, OSEM + time-of-flight (ToF), and OSEM + ToF + point-spread function modelling for resolution recovery (UHD), with;
- no post-smoothing applied (0 mm), or 4 mm, and 6 mm width Gaussian post-smoothing filters.

Using JSRecon12, part of the package supplied by Siemens that form the e7-tools off-line framework, the raw sinogram dataset was reconstructed with the above permutations.

Parameter	**Setting**
zoom	1
MashFlag	1
matchctslice	0
AbsFlag	0
BedRemoval	1
CompressFlag	1

Table 7.1: A table of some relevant JSRecon_params.txt parameters

ROIs for the six largest spheres and the background were established using the image analysis software LIFEx v5.38, according to the NEMA NU-2 standards [56, 101]. The ROIs were drawn on a UHD, 2 iteration image with a 4 mm post-filter, and re-used for every image from both systems. LIFEx was also used to extract image metrics using its batch extraction scripting feature. For the purposes of this study, the maximum, minimum, mean and standard deviation of each region were compared.

A demonstration of the mean activity concentration voxel values are shown in Figure 7.7, plotted against the number of iterations for the largest sphere (37 mm in diameter). The colours of the distributions represent the size of the Gaussian blurring filter applied, while the left hand side and right hand side distributions

show the scanner and e7-tools reconstructions respectively. The figure shows how the distribution of the sphere's maximum values between the frames compares between the scanner and the e7-tools reconstruction. It is reasonable to claim that the e7-tools data is drawn from the same distribution as that of the scanner, yet slight discrepancies are evident. Here we see that the mean values are faithfully reproduced, with small percentage changes in the means of each distribution between the systems.

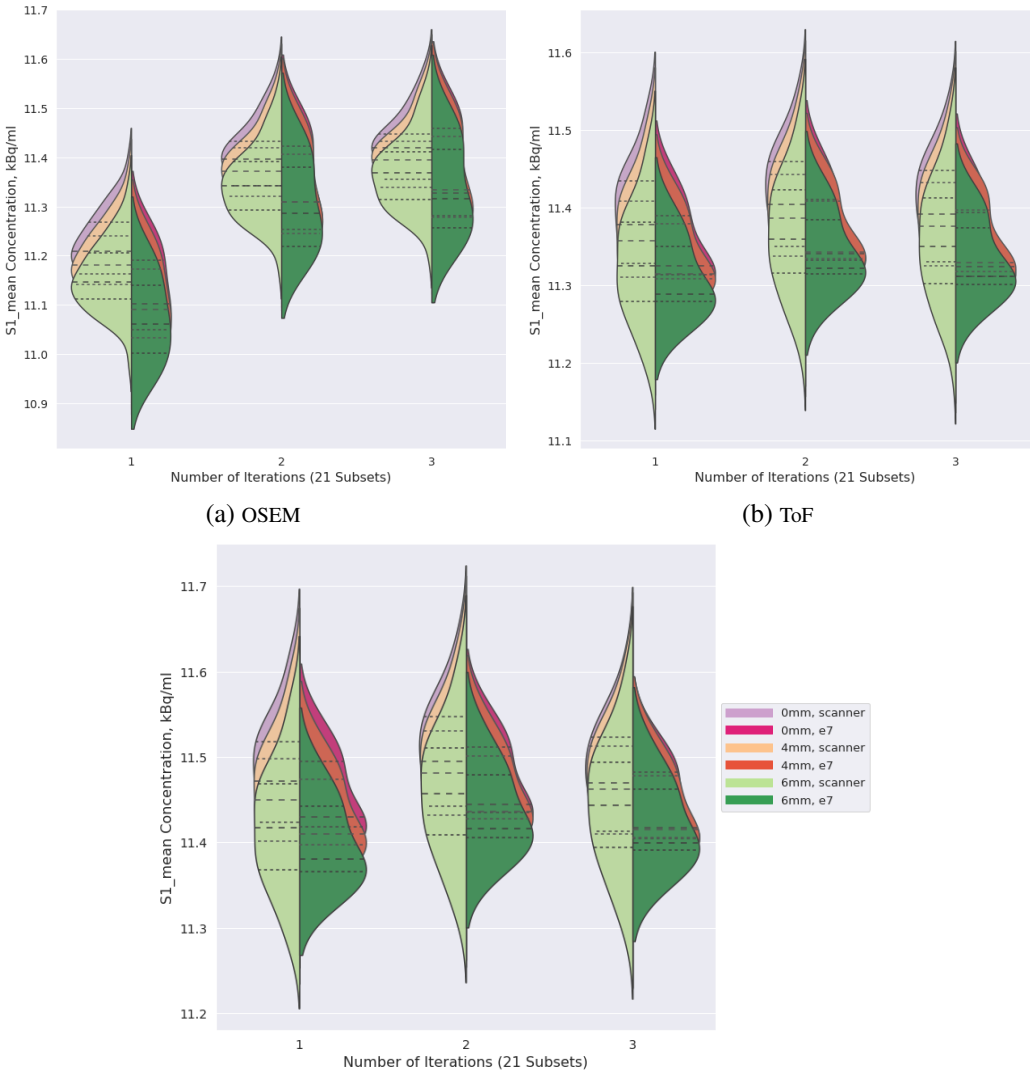
However, an interesting observation can be made regarding the difference between standard OSEM and reconstruction methods involving time-of-flight considerations. Figure 7.8 shows, in a similar manner to Figure 7.7, violin plots comparing the distribution of standard deviations in different reconstructions of the sphere. The figure demonstrates a clear distinction in the distributions of values for OSEM reconstructions that is not observed with ToF and UHD. The same effect is also seen in the equivalent plots for the maxima and minima.

Table 7.2 demonstrates this in statistical form, taking averages across all images in a dataset which includes several different sizes of post-filter blurring applied. Taking a particular - realistic - example however, using 2 iterations of 21 subsets and blurring with a 4 mm Gaussian post-filter, the trend in these results is clear. When these parameters are used in standard OSEM, the ROI standard deviation differs between the scanner and e7 reconstructions by (3.79 ± 1.00) %, compared to (0.32 ± 1.12) % for ToF and (0.27 ± 1.16) % for UHD.

Statistic	**Average Percentage Difference**		
	OSEM	**ToF**	**UHD**
Mean	(0.55 ± 0.16) %	(0.37 ± 0.14) %	(0.39 ± 0.14) %
Standard Deviation	(5.60 ± 0.31) %	(0.36 ± 0.34) %	(0.30 ± 0.39) %
Maximum	(3.95 ± 0.78) %	(0.90 ± 0.83) %	(0.93 ± 0.71) %
Minimum	(3.85 ± 1.16) %	(1.66 ± 1.39) %	(1.41 ± 1.17) %

Table 7.2: The weighted arithmetic means of the percentage differences between ROI statistics from equivalent reconstructions performed on the two software.

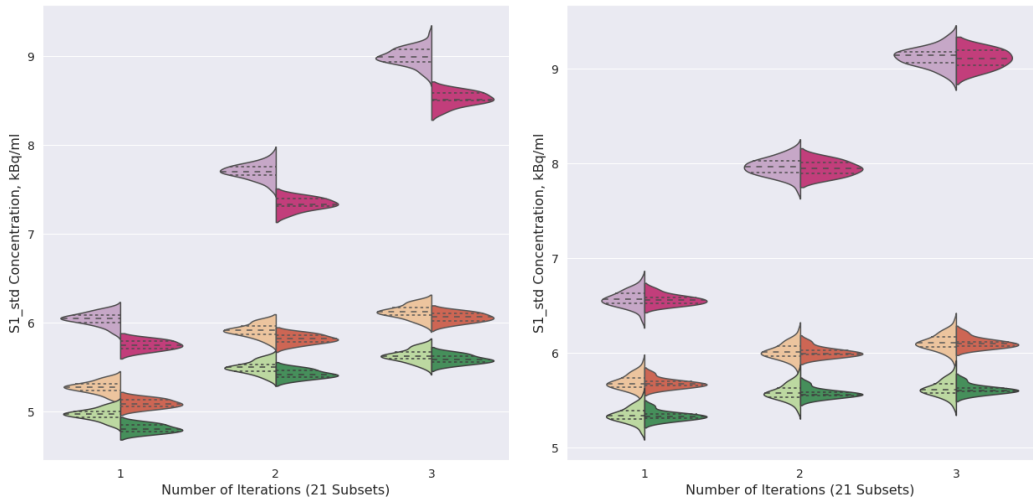
There is a good level of reproducibility of image statistics between the software for ToF and UHD reconstructions. The results show a slight discrepancy in how standard deviation, maximum and minimum values are measured in OSEM compared to ToF and UHD. This will be of an academic interest for studies de-



(a) OSEM (b) ToF

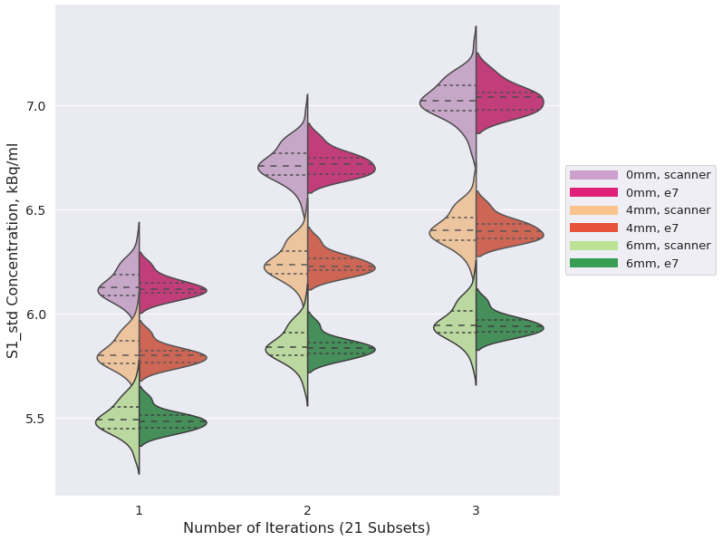
(c) UHD

Figure 7.7: Violinplots showing the distribution of mean values obtained in the ROI of S1, the largest of the NEMA spheres. Each side of the violin compares the distribution between the scanner- and e7-reconstructed images, and should ideally be symmetrical to show equivalent system performance. Number of iterations and post-filter size are also included for comparison.



(a) OSEM

(b) ToF



(c) UHD

Figure 7.8: Violinplots showing the distribution of the standard deviation of voxel values obtained in the ROI of S1. Each side of the violin compares the distribution between the scanner- and e7-reconstructed images, and should ideally be symmetrical to show equivalent system performance. Number of iterations and post-filter size are also included for comparison.

pendent on heterogeneity metrics; it is unclear at present how this discrepancy may manifest in the many higher-order metrics that rely on connectivity matrices, for example. The size of the impact is, of course, inhibited by the evermore widespread usage of ToF in PET data acquisition for these studies.

After the presentation of this data at BNMS, meetings were arranged with the algorithm developers at Siemens Healthineers. After discussions, the differences were attributed to the DICOM conversion process in e7 tools; subsequent updates are expected to have addressed potential issues in the handling of the data. This experiment nonetheless demonstrated the difficulty in producing equivalent images even on softwares issued by the same supplier.

7.3 Remarks and Conclusions

The advantages of using simulation in studies such as this are clear, as the creation of a theoretically infinite set of activity distributions would enable truly reliable assessments of feature robustness. When considering an alternative to phantom studies, the comparative low cost, high resolution and precision, and lack of the need for radiation exposure that is associated with simulation makes it an attractive avenue to pursue. In order for simulations to be used effectively, full validation must be completed. The many stages of PET data acquisition, correction and reconstruction each require separate validations, resulting in a lengthy and convoluted process. Commercially available hardware and software often progresses without active consultation of the research and open-source software communities, and this is evident by the only very recent inclusion of time-of-flight and virtual crystal capabilities into open-source reconstruction packages. The e7 tools validation study in Section 7.2 has shown that even software by the same manufacturer may produce variable results on the same data. It was evident in investigating this area that further investment is required into establishing harmonised data handling processes before truly equivalent image collection and reconstruction can be provided by simulation software.

Chapter 8

Impact of this Work & The Future of the Field

This chapter summarises and contextualises the results and conclusions of the work in this thesis. Starting with Section 8.1, the conclusions of the work are positioned as recommendations that could be argued to current researchers in the field of clinical image-based analysis and biomarker development. These recommendations are followed by a summary of the limitations and caveats of the work in Section 8.2. Subsequently, Section 8.3 reviews some key areas of cutting edge research in PET, and discusses the impact that such technologies could have on the field alongside the results of this work. The technologies discussed (Total Body PET, monolithic detectors and AI implementations) have been chosen for the magnitude of their perceived impact and relationship to this work, yet represent only a fraction of the diverse and exciting areas that PET research is exploring at the present. The final summary is given in Section 8.4, concluding the thesis.

8.1 Recommendations from this Work

This work utilised high activity phantom scans to examine the effect of changing activity on pyradiomics texture matrix-based features, using the NECR as a measure of characteristic noise. The results in Chapter 5 have shown that there is rea-

son to suspect that the 75 texture features tested are not satisfactorily robust with general radioactivity level. While the study was limited to phantom work, this could potentially impact patient PET studies where patient size or tracer choice will significantly affect the level of radioactivity in the FoV at any given bed position. There is reason to believe, however, that certain features such as GLCM IMC1 and IMC2 could be corrected to model for the quantified noise in a scan. Such a correction would take place in a feature pre-processing or normalisation stage of a classification task. In addition, there have been concerns raised over the stability of low gray level emphasis features. The volume of ROIs is an important factor in accounting for noise, as traditional global measures of data signal noise ratio (NECR) become less valid on smaller, localised observations within the image. Efforts were made to bolster this work with Monte Carlo simulation and image-based noise quantification, and can be progressed with further developments from collaborations working specifically in these respects.

Currently, the IBSI recommend that work in PET radiomics must include certain information when reporting results. These required details are listed in the IBSI guidelines. The guidelines are comprehensive within reason, requiring disclosure of all image correction and reconstruction procedures, patient preparation details including tracer information and injected activity, and basic scanner information. Patient studies are encouraged to verify obtained results with respect to previous studies that share similarities in these respects. Constructing the guidelines was a task undertaken after widespread concerns around much of the early adoption of radiomics techniques without harmonisation and standardisation [120, 19, 18].

As demonstrated in Chapter 5, this work suggests that quoting NECR information alongside other, more general scan and activity information might be a useful starting point for adopting noise correction in future work. In reality, this solution is incomplete. It is the functional form of the NECR which is of interest in the modelling, and more valuable information would be the activity in FoV at the time of the scan alongside the activity in the scan at the peak NECR. This peak information will be impossible to obtain for clinical data; for reasons pertaining to patient safety, it will be impossible to image a patient with enough activity to reach this peak NECR. Possible solutions to this could be in Monte Carlo sim-

ulation, modelling NECR by recreating the patient geometry in methods similar to those discussed in Chapter 7. This is complicated by multiple bed positions or continuous bed motion scanning, and as the NECR link is not provably solid or substantial, this extra work may not be deemed valuable for future studies to carry out.

One potential parallel in PET harmonisation is the Deauville criteria. The Deauville criteria is a 5-point scale (5-PS in literature) for categorising PET images, comparing tracer uptake in a region of interest (originally specifically in two different types of lymphoma, Hodgkin and diffuse large B-cell) to the uptake in the mediastinum and liver [26]. This reproducible classification system has proved useful in many studies for patient treatment outcome reporting, despite its deceptively simple definition [121, 122], and a similar system could be developed for categorising images which are used in radiomics-based feature analysis. One such way could be in analysing feature values in the liver as a form of baseline. In FDG PET, a cancer-free liver would be a large region of a patient with an approximately homogeneous uptake distribution which would fill a majority extent of the scanner axial FoV.

The conclusions from this work could be made more general with a wider and more diverse range of phantom activity distributions. While there is merit to using phantoms, homogeneous activity distributions even after imposed image noise will likely only occupy a narrow manifold of possible texture feature values. There is precedent for other robustness studies using phantoms with in-built ground truth heterogeneity, albeit with less anthropomorphism and a slightly different focus [94]. Such studies are vital in assessing true robustness, and future studies should ensure this variety of ground truth distributions. Ideally, with the wish to further the impact of this work, these in-built heterogeneous phantoms should undergo similar procedures of high-activity multi-frame scanning. This would enable more effective assessment of the impact of noise-based feature correction, and would enhance trust in the results of statistical tests to verify the robustness of such corrections (see the Kruskal-Wallis test in Section 5.2.6).

8.2 Caveats to this Work

The recommendations in the previous sections are provided with full knowledge of the limitations imposed by the experimental setup and environment. This section seeks to explain the impact of several key methodological factors, and how these could be addressed in future work.

The use of homogeneous regions of interest limits the impact of this work. The regions used in the study varied in shape and size; from the large and small conventional volumes of the cylinder and NEMA spheres to the non-conventional yet perhaps more clinically relevant shapes of the custom tumour inserts. The values of many texture features occupy manifestly different domains for the different ROIs (see Figure 5.22) despite the common uniformity of these activity distributions. This relates in part to the construction of the texture matrices; a difference in volume results in different restrictions that (for instance) values for Run Lengths or Size Zones may take. Suggestions for improvements here are detailed above, as part of a recommendation for how to implement the findings of this work into future studies.

The tenuousness of any direct relationship between NECR and image noise complicates the establishment of definitive NECR correlations for texture features. It is established that iterative image reconstruction imposes noise onto an image that is unaccounted for by the NECR ratio, and while this may make for clearer images, any resultant noise is more difficult to quantify from count information alone [99]. For this reason, similar studies to this have recommended that perhaps more rudimentary yet analytical reconstruction methods such as FBP should be used for data-oriented study, whereas more complex and developed iterative image reconstruction procedures, utilising techniques such as resolution recovery, should be reserved for clinical use [21]. The reconstruction used in this work used resolution recovery by point-spread function modelling, and applied a 3D Gaussian blurring filter post-reconstruction with a 5 mm width. These are useful techniques for a clinician, lessening the effect of voxelised noise on the ability to visually delineate potentially small-sized cancerous material. Nevertheless, these will significantly impact the quantification of image noise in ways that cannot be explicitly measured when combined with OSEM-based reconstruction

algorithms.

Image discretisation is one aspect of radiomics that requires the most fastidious standardisation. This work chose to implement a fixed bin number of 64 across all regions (FBN:64 to obey IBSI nomenclature). While there is currently no agreement on whether fixed bin size or fixed bin number is a more appropriate standardisation procedure, previous studies have found there to be no informational advantage to be gained from bin numbers of less than 32 [92] or greater than 64 [15, 93]. Such studies informed the choices implemented in this work, and were maintained for consistency throughout the study. Previous studies showed that most texture features show poor robustness to image discretisation settings [123, 124]. This should be unsurprising, as, similarly to changing ROI volume, this inherently impacts the potential size and shape of the resultant Haralick matrices. While discretisation settings have been stated here, future work could investigate the impact of using 32 or 64 bins to determine whether the effects of image noise are alleviated by using fewer bins. It is unclear whether the information 'gain' by using the additional bins is somehow complicit in amplifying existing noise, and more robust feature values are attainable by using fewer bins.

In addition to image discretisation, one possibly neglected aspect of the PET radiomics process is image interpolation. Possibly the biggest drawback to PET imaging is the poor spatial resolution in comparison to CT and MRI, resulting in radiomics features being calculated over fewer voxels for the same ROI. For PET radiomics information to be used alongside CT or MRI radiomics, image registration and interpolation has to be performed, requiring up- or down-sampling of the original image prior to discretisation and segmentation. There are obvious concerns to be had over potential loss of information when down-sampling, and overconfidence in misleading data when up-sampling a noisy PET image. There is evidence that some features are robust to interpolation whereas others exhibit instability [125]. Interpolation was not considered in this study, but would be an useful extension to the work in this thesis. Future work should seek to investigate this in more depth.

Further work should implement a wider range of tumour volumes in these studies. The breakdown in NECR correlations around 40 cm^3, and the similarity

of that volume to that at which there is a proposed breakdown in heterogeneity calculation [79], suggests there is merit in investigating further volumes around this value. There is merit in examining scale, printing tumours with identical shapes but varying volumes, to test robustness of texture features to this aspect. With the implementation of heterogeneous compartmental tumour inserts, similarly to Pfaehler et al (2020) [95], the extended set would enable a more comprehensive conclusion of robustness to activity for radiomics texture features.

8.3 Advancements in PET

This section will describe some of the exciting research currently underway across the cutting edge of PET research. The first advancement mentioned is that of Total Body PET. The phrase 'total body' refers to the initial proposal of a 2 m axial FoV PET scanner, enabling the acquisition of PET data from the entire patient simultaneously [108]. The advantage of the increased axial FoV is an increase in geometric sensitivity for any given emission. This has consequences on potential noise quantification, and the links to the work done in this thesis are expanded upon therein.

The following subsection details how 'monolithic' detectors are being developed to increase the spatial and energy resolution of future PET systems. A monolithic detector replaces the cut-crystal block design (see Figure 2.5) with a large single crystal attached to an array of silicon photomultipliers (SiPMs). The advantage of increased spatial resolution will greatly increase the possible matrix size of a resultant PET image, and therefore should help in reducing the possible volume limit of PET radiomics. This is only one area where improved hardware can improve image resolution and quality, but is one of the most exciting and innovative.

Artificial intelligence and machine learning have an indelible link to the work in this thesis. However the areas of patient diagnosis, treatment planning, prognosis, and many others that have been mentioned to this point, are not the only areas where AI has the potential to impact nuclear medicine research. Many new algorithms utilise machine learning models to reconstruct images from raw data, and there are hopes that in years to come these will replace the contemporary iterative

reconstruction algorithms by providing clearer images with better quantification and, potentially, voxelised uncertainty mapping [126].

8.3.1 Total Body PET

The current standard PET ring has an axial FoV of around 20 cm, with a between-detector diameter of around 70 cm. Using the formula for the solid angle viewed from the centre of an open-ended cylinder Ω_{cyl},

$$\Omega_{cyl} = \frac{4L}{\sqrt{L^2 + 4R^2}}\pi, \tag{8.1}$$

where L and R represent axial length and radius respectively, this gives $\Omega_{cyl} \approx \pi$ (and more specifically 0.99π in the case of the Biograph mCT). There is little that can be done with the current system architecture to account for the loss of counts due to photons that exit from the ends of this cylinder. Further problems come from activity outside the FoV; if one photon from outside the FoV hits the detector, this increases the likelihood of random noise.

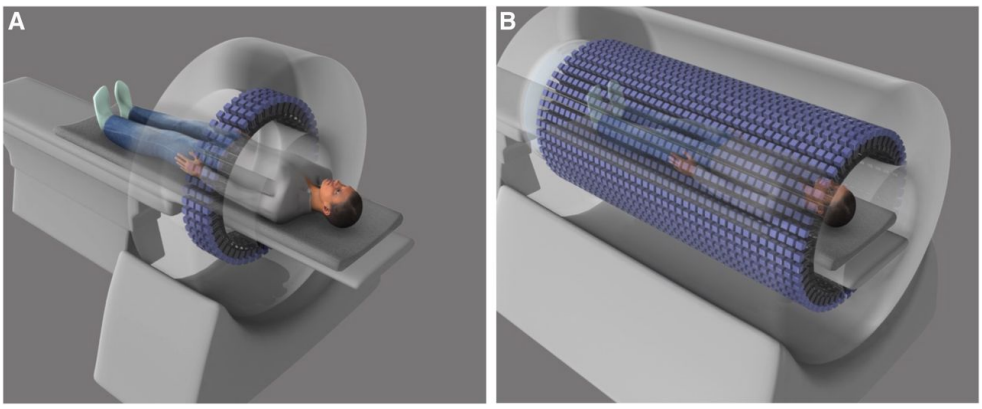


Figure 8.1: An illustration of the difference in axial FoV between conventional PET (a) and Total Body PET (b) taken from [127].

A proposed method for accounting for this is Total Body PET; constructing larger axial FoV scanners to ideally image an entire patient in a single bed position. Several such scanners have been proposed, each of varying extents. The uExplorer system features a 1.94 m axial FoV, while Siemens' Biograph Vision Quadra system consists of four back-to-back PET rings from the Biograph Vision

design, giving an axial FoV of 1.04 m [128, 129]. For the uExplorer, the solid angle covered from the centre is $\sim 3.71\pi$. These two detector designs represent the two sides to current Total Body PET research. Siemens' design is indicative of where Total Body PET might fit into clinical practice, with the smaller FoV bringing a more competitive cost at the expense of the extra gain in sensitivity given by the research-focussed uExplorer machine.

If more emissions per second are detected from an activity distribution, the same data can be acquired using either a smaller acquisition time frame, Δt, or less total activity, A. The equation

$$\mathrm{SNR}_{\mathrm{image}} \propto \sqrt{S \times A \times \Delta t} \tag{8.2}$$

can be used to describe this effect, where S denotes the effective sensitivity of the system [108]. This equation is particularly pertinent to results obtained in this work. The cylinder data demonstrated that on larger time acquisition datasets, noise in image texture statistics for large objects could be more closely modelled by the NECR, and thus become more simple to correct. Using Equation 8.2, the resultant effect of the 5-fold increase in Δt required to produce this is comparable to the expected four-fold increase in S expected when imaging a 20 cm axial length object in a Total Body PET scanner.

At the time of writing there are only around 20 Total Body PET systems installed worldwide. The concept is not new, but only in recent years has the clinical case been strong enough, detector technology been precise enough, and construction price reduced to the point where production could be possible [130]. At the time of writing, a standard PET system could be expected to cost around €3 million, while a Total Body system could cost €18 million. This increased cost is not only due to the extra materials needed to create the detector, but also the increased computing power required alongside such a scanner in order for it to run. While this cost is likely to still be prohibitive to most PET centres, it is anticipated that this increased cost could be balanced out by clinical advantages to the sensitivity gain. Using Equation 8.2, we can deduce that to produce an image of the same quality as a standard PET scan, either the activity or the scan duration can be reduced. This means that the patient dose can be reduced, which is advan-

tageous for patient safety and cost incurred at the centre. A reduced scan time per patient would also increase the throughput of any given PET centre, enabling more patients to be seen per day and enhancing the centre's efficiency. Furthermore, the potential of low-dose imaging could result in the expansion of PET into paediatrics, providing better care for a larger proportion of the population [131].

8.3.2 Monolithic PET Systems

One of the major limiting factors in PET images is spatial resolution. There are fundamental limits set by known physics phenomena, such as acollinearity, yet there are limits imposed by the detector hardware used [132]. Better timing resolution provides one way of improving spatial resolution in the resultant image - improving ToF accuracy will impose stricter limits on possible emission locations. However, the use of pixelated detector elements will always restrict the spatial resolution to a 'thick' line of response, the width and breadth of said crystal pixel size. The resultant voxel sizes in PET mean that ROIs that delineate small tumours, for instance, may only consist of a handful of voxels, and invalidate many useful aspects of deriving texture features in the first place.

Monolithic PET systems replace the pixelated crystal block with a single uniform block - the eponymous monolith. This monolithic block is then connected to a grid of semiconductor photomultipliers. The central tenet is that the physics properties of the incoming photons can be more accurately measured and, using well-trained classification techniques, the line of response of the incoming photon can be restricted to a much smaller volume. There is an analogy here to the transition from 2D to 3D PET. The constant improvement in computing power and new possibilities in backend technology enable the loosening of some restrictions on what is possible for a PET acquisition. In addition to improved spatial resolution, the lack of internal reflections caused by detector block pixelation could potentially lead to a better timing resolution in monolithic block detectors [133]. The improved spatial and timing resolution would benefit texture analysis, reducing the volume limit for the effectiveness of Haralick matrices, while potentially reducing the effect of noise between neighbouring voxels.

8.3.3 AI-Based Image Reconstruction

This work to date has concentrated on improving the extraction of data from images, largely for the betterment of future artificial intelligence processes for computer-aided diagnosis and treatment planning. Statistical characterisation is, however, only one way that artificial intelligence is being used in PET imaging. Deep learning specifically is predicted to have a significant impact in the near future on the methods by which images are produced from PET raw data. This section will list three ways in which the impact of machine learning is likely to be seen in the field in the near future; firstly in image denoising, secondly in full reconstruction, and finally in establishing image uncertainty maps.

Image denoising techniques have been developed as a post-reconstruction processing step to remove noise from a PET image. Companies such as the US-based Subtle Medical, Inc. have developed algorithms that have been proven to recover image quality and preserve image quantification for images reconstructed from data with a halved acquisition frame duration [134, 135]. These algorithms are new to the market and will likely only improve as they are able to be trained on more data, and with Subtle Medical obtaining FDA approval for their algorithm it is very likely that techniques similar to this could be seen in the clinic in the coming years. The clinical advantages will be seen in the ability to scan patients for less time or with less activity, enabling more efficient and safer clinical practice. From the research perspective, the results are more relevant for their impact on current protocols. If degradation can be recovered from half-duration images, there is a clear promise of better quantification for future 'full'-duration scans. Images with more reliable quantification and less visible noise effects would, by extension, enable more robust texture feature calculation, with the impact of noise on ground truth texture potentially extricable.

While denoising seeks to correct images post-reconstruction, other algorithms are in development to replace the reconstruction process altogether. Deep learning is so-named because of the layers to the algorithm's structure, and there have been attempts to use the layers of this structure to replicate the iterations of an iterative reconstruction process [126]. These algorithms can become very complex, such as the Learned Primal Dual reconstruction process, with two deep

learning networks teaching each other to improve the 'guesses' at each stage of an iterative process [136]. GE have developed the TrueFidelity algorithm, which is expected to become a clinical fixture in CT in the near future [137]. Clinical communities generally reserve doubts about the validity of reconstructing images with algorithms trained on a restricted dataset, yet a great deal of modern algorithm development includes a degree of in-build physics knowledge [138].

Deep-learned reconstruction is likely to make a significant impact in the wider field with some immediacy, yet there is some conflict over whether such algorithms will aid or hinder quantification of images. At present, however, OSEM reconstructions are only able to present a 'best guess' after a certain number of iterations, given very little by way of prior information. AI-based reconstruction can provide more informed priors for the reconstruction process [126]. Possibly the most exciting ways, not to mention most relevant to this work, that AI reconstruction is different is in the possible generation of companion epistemic uncertainty maps alongside the image [126, 138]. This has been successfully attempted using probabilistic Bayesian methods alongside deep learning reconstruction for brain cone beam CT [139]. If these methods could be enveloped into PET, voxelised uncertainty maps would enable thorough development and propagation of epistemic uncertainties for all manner of texture features. In the author's opinion, this is the innovation that could provide the missing piece for wider integration of radiomics into clinical practice.

8.4 The Final Word

This work sought to provide an example methodology for assessing the robustness to activity level of texture features used in radiomics for oncology PET. Using high-activity FDG phantom scans, it was determined that many texture features appear to lack this robustness. The NECR was used to provide a correction method for texture features, enabling an estimation of the values that texture features could take at clinical levels. NECR is, however, weakly linked to image SNR, and models were developed to encapsulate image-based noise into a metric to enable more informed correction. A simulation framework was developed in an attempt to model phantoms with innovative in-built heterogeneity and thus

improve the impact of this work. Progress was made in both of these regards, the outputs of which provide solid foundations for future work. While it appears, from the work in this thesis and similar published studies, that radiomics texture features are currently significantly unstable in PET images, there is great cause for optimism. Progress in PET software and hardware is not slowing down and, alongside persistent and continuous efforts improving harmonisation and standardisation across research and clinical fields, there can be little doubt that this can enable radiomics to become the burgeoning success that its proponents have always believed it could be.

Bibliography

[1] W.H. Sweet. The uses of nuclear disintegration in the diagnosis and treatment of brain tumor. *New England Journal of Medicine*, 245(23):875–878, 1951.

[2] R. Nutt. The history of positron emission tomography. *Molecular Imaging & Biology*, 4:11–26, 2002.

[3] G.F. Knoll. *Radiation Detection and Measurement*. John Wiley & Sons, 2010.

[4] D.W. Townsend. Combined positron emission tomography–computed tomography: the historical perspective. *Seminars in Ultrasound, CT and MRI*, 29(4):232 – 235, 2008.

[5] T. Turkington. Introduction to PET instrumentation. *Journal of Nuclear Medicine*, 30(2), 2002.

[6] NHS England Performance Analysis Team. Diagnostic imaging dataset 2021-22. Available at `https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostic-imaging-dataset/`, accessed 17th June 2022, published 2022.

[7] A. Scarsbrook and S. Barrington. Evidence-based Indications for the use of PET-CT in the United Kingdom. Available at `https://www.rcr.ac.uk/publication/evidence-based-indications-use-pet-ct-united-kingdom-2016`, accessed 27th June 2022, published 2016.

[8] (ARSAC) Administration of Radioactive Substances Advisory Committee. Notes for Guidance on the Clinical Administration of Radiopharmaceuticals and Use of Sealed Radioactive Sources, January 2022. Available at `https://www.gov.uk/government/publications/arsac-notes-for-guidance`, accessed 18th June 2022, published 2022.

[9] R. Boellaard, R. Delgado-Bolton, W. J.G. Oyen, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *European Journal of Nuclear Medicine and Molecular Imaging*, 42(2):328–354, February 2015.

[10] A. R. Pantel, V. Viswanath, M. Muzi, et al. Principles of Tracer Kinetic Analysis in Oncology, Part I: Principles and Overview of Methodology. *Journal of Nuclear Medicine*, 63(3):342–352, March 2022.

[11] A. R. Pantel, V. Viswanath, M. Muzi, et al. Principles of Tracer Kinetic Analysis in Oncology, Part II: Examples and Future Directions. *Journal of Nuclear Medicine*, 63(4):514–521, April 2022.

[12] G. Lucignani, G. Paganelli, and E. Bombardieri. The use of standardized uptake values for assessing FDG uptake with PET in oncology: A clinical perspective. *Nuclear Medicine Communications*, 25(7):651–656, 2004.

[13] I. El Naqa, P. W. Grigsby, A. Apte, E. Kidd, E. Donnelly, D. Khullar, S. Chaudhari, D. Yang, M. Schmitt, Richard Laforest, W. L. Thorstad, and J. O. Deasy. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognition*, 42(6):1162–1171, June 2009.

[14] R. J. Gillies, A. R. Anderson, R. A. Gatenby, and D. L. Morse. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clinical Radiology*, 65(7):517–521, July 2010.

[15] S.S.F. Yip and H.J.W.L. Aerts. Applications and limitations of radiomics. *Physics in Medicine and Biology*, 61(13):R150–R166, 2016.

[16] E. Farina, J. J. Nabhen, M. I. Dacoregio, et al. An overview of artificial intelligence in oncology. *Future Science OA*, 8(4):FSO787, 2022.

[17] G.J.R. Cook, M. Siddique, B.P. Taylor, et al. Radiomics in PET: principles and applications. *Clinical and Translational Imaging*, 2(3):269–276, 2014.

[18] M. Sollini, L. Cozzi, L. Antunovic, et al. Pet radiomics in nsclc: State of the art and a proposal for harmonization of methodology. *Scientific Reports*, 7(1), December 2017.

[19] A. Traverso, L. Wee, A. Dekker, and R. Gillies. Repeatability and reproducibility of radiomic features: A systematic review. *International Journal of Radiation Oncology Biology Physics*, 102(4):1143–1158, November 2018.

[20] M. Decuyper, J. Maebe, R. Van Holen, and S. Vandenberghe. Artificial intelligence with deep learning in nuclear medicine and radiology. *European Journal of Nuclear Medicine and Medical Imaging Physics*, 8(1), December 2021.

[21] G. A. Prenosil, T. Weitzel, M. Fürstner, et al. Towards guidelines to harmonize textural features in PET: Haralick textural features vary with image noise, but exposure-invariant domains enable comparable PET radiomics. *PLoS ONE*, 15(3), 2020.

[22] G. J. R. Cook, G. Azad, K. Owczarczyk, et al. Challenges and promises of PET radiomics. *International Journal of Radiation Oncology*Biology*Physics*, 102(4):1083 – 1089, 2018.

[23] T. Beyer, J. Czernin, and L. S. Freudenberg. Variations in clinical PET/CT operations: Results of an international survey of active PET/CT users. *Journal of Nuclear Medicine*, 52(2):303–310, February 2011.

[24] N. Aide, C. Lasnon, P. Veit-Haibach, 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, August 2017.

[25] A. Kaalep, C. N. Burggraaff, S. Pieplenbosch, et al. Quantitative implications of the updated EARL 2019 PET–CT performance standards. *EJNMMI Physics*, 6, December 2019.

[26] M. Meignan, A. Gallamini, and C. Haioun. Report on the first international workshop on interim-PET scan in lymphoma. *Leukemia and Lymphoma*, 50(8):1257–1260, 2009.

[27] R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.

[28] J. van Griethuysen, A. Fedorov, N. Aucoin, et al. pyradiomics Documentation. Available at `https://pyradiomics.readthedocs.io/en/latest/`, accessed 30th June 2022, published 2016.

[29] C. S. Levin and E. J. Hoffman. Calculation of positron range and its effect on the fundamental limit of positron emission tomography system spatial resolution. *Phys. Med. Biol.*, 44(3):781–799, 1999.

[30] A. Sanchez-Crespo, P. Andreo, and S.A. Larsson. Positron flight in human tissues and its influence on PET image spatial resolution. *European Journal of Nuclear Medicine and Molecular Imaging*, 31(1):44–51, 2004.

[31] S. Jan, C. Comtat, D. Strul, et al. Monte Carlo simulation for the ECAT EXACT HR+ system using GATE. *Nuclear Science, IEEE Transactions on*, 52:627 – 633, 2005.

[32] D. Hanahan and R.A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57 – 70, 2000.

[33] D. Hanahan and R.A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646 – 674, 2011.

[34] O. Warburg. On the origin of cancer cells. *Science*, 123(3191):309–314, 1956.

[35] M.E. Casey. *An analysis of counting losses in positron emission tomography*. PhD thesis, University of Tennessee, 1992.

[36] J. Hsieh. *Computed Tomography: Principles, Design, Artifacts, and Recent Advances*. SPIE Press monograph. SPIE Press, 2003.

[37] B. Bendriem and D.W. Townsend. *The Theory and Practice of 3D PET*. Developments in Nuclear Medicine. Springer-Verlag, 1998.

[38] S. R. Meikle and R. D. Badawi. Quantitative techniques in PET. In Dale L Bailey, David W Townsend, Peter E Valk, and Michael N Maisey, editors, *Positron Emission Tomography*. Springer-Verlag, 2005.

[39] J.W. Müller. Dead-time problems. *Nuclear Instruments and Methods*, 112(1):47 – 57, 1973.

[40] D.L. Bailey. Transmission scanning in emission tomography. *European Journal of Nuclear Medicine*, 25(7):774–787, 1998.

[41] S.K. Yu and C. Nahmias. Single-photon transmission measurements in positron tomography using 137-Cs. *Physics in Medicine and Biology*, 40(7):1255–1266, 1995.

[42] L. Theodorakis, G. Loudos, V. Prassopoulos, et al. A review of PET normalization: striving for count rate uniformity. *Nuclear Medicine Communications*, 34(11):1033, 2013.

[43] G.B. Saha. *Basics of PET imaging: physics, chemistry, and regulations*. Springer, New York, 2nd edition, 2010.

[44] P. Kinahan, D.W. Townsend, D.L. Bailey, et al. Efficiency normalization techniques for 3D PET data. pages 1021 – 1025 vol.2, 1995.

[45] M. Defrise, D.W. Townsend, D.L. Bailey, et al. A normalization technique for 3D PET data. *Phys. Med. Biol.*, 36(7):939–952, 1991.

[46] M.W. Stazyk, V. Sossi, K.R. Buckley, and T.J. Ruth. Normalization measurement in septa-less PET scanners. In *Journal of Nuclear Medicine*, volume 35:5, pages 41–41, 1994.

[47] D.L. Bailey and T. Jones. Normalization for 3D PET with a translating line pseudo-plane source. *Journal of Nuclear Medicine*, 36(5):92–93, 1995.

[48] M. Soret, S.L. Bacharach, and I. Buvat. Partial-Volume Effect in PET Tumor Imaging. *Journal of Nuclear Medicine*, 48(6):932–945, 2007.

[49] W.W. Moses. Fundamental limits of spatial resolution in PET. *Nuclear Instruments & Methods in Physics Research. Section A, Accelerators, spectrometers, detectors and associated equipment*, 648 Supplement 1:S236–S240, 2011.

[50] O.L. Munk, L. Tolbod, S. Hansen, and T. Bogsrud. Point-spread function reconstructed PET images of sub-centimeter lesions are not quantitative. *European Journal of Nuclear Medicine and Molecular Imaging*, 4, 2017.

[51] Y. Vardi, L. A. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80(389):8–20, 1985.

[52] R.E. Carson and K.Lange. A statistical model for positron emission tomography: Comment. *Journal of the American Statistical Association*, 80(389):20–22, 1985.

[53] H.M. Hudson and R.S. Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Transactions on Medical Imaging*, 13(4):601–609, 1994.

[54] R. Boellaard. Standards for PET image acquisition and quantitative data analysis. *Journal of Nuclear Medicine*, 50(SUPPL. 1), May 2009.

[55] J. J. M. Van Griethuysen, A. Fedorov, C. Parmar, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, November 2017.

[56] C. Nioche, F. Orlhac, S. Boughdad, et al. LIFEx: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. *Cancer Research*, 78(16):4786–4789, August 2018.

[57] J. O. Deasy, A. I. Blanco, and V. H. Clark. CERR: A computational environment for radiotherapy research. *Medical Physics*, 30(5):979–985, May 2003.

[58] A. P. Apte, A. Iyer, M. Crispin-Ortuzar, et al. Technical note: Extension of CERR for computational radiomics: A comprehensive MATLAB platform for reproducible radiomics research. *Medical Physics*, 45(8):3713–3720, August 2018.

[59] C F Njeh. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *Journal of Medical Physics*, 33(4), 2008.

[60] A. J. Weisman, M. W. Kieler, S. Perlman, et al. Comparison of 11 automated PET segmentation methods in lymphoma. *Physics in Medicine and Biology*, 65(23), November 2020.

[61] M. Hatt, J. A. Lee, C. R. Schmidtlein, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group No. 211. *Medical Physics*, 44(6):e1–e42, June 2017.

[62] R. T. H. Leijenaar, G. Nalbantov, S. Carvalho, et al. The effect of SUV discretization in quantitative FDG-PET radiomics: The need for standardized methodology in tumor texture analysis. *Scientific Reports*, 5(8), 2015.

[63] R.L. Wahl, H. Jacene, Y. Kasamon, and M.A. Lodge. From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors. *Journal of Nuclear Medicine*, 50(1):122S–150S, 2009.

[64] D. Hasenclever, L. Kurch, C. Mauz-Körholz, 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(7):1301–1308, 2014.

[65] G.D. Kolinger, D. Vállez García, G.M. Kramer, et al. Repeatability of [18F]FDG PET/CT total metabolic active tumour volume and total tumour burden in NSCLC patients. *European Journal of Nuclear Medicine and Molecular Imaging Research*, 9(1):14, 2019.

[66] H. Wadell. Volume, shape, and roundness of quartz particles. *The Journal of Geology*, 43(3):250–280, 1935.

[67] F. Hofheinz, A. Lougovski, K. Zöphel, et al. Increased evidence for the prognostic value of primary tumor asphericity in pretherapeutic FDG PET for risk stratification in patients with head and neck cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, 42(3):429–437, 2015.

[68] M. Hatt, B. Laurent, H. Fayad, et al. Tumour functional sphericity from PET images: prognostic value in NSCLC and impact of delineation method. *European Journal of Nuclear Medicine and Molecular Imaging*, 45(4):630–641, 2018.

[69] S. Chicklore, V. Goh, M. Siddique, et al. Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis. *European Journal of Nuclear Medicine and Molecular Imaging*, 40(1):133–140, 2013.

[70] E. Sala, E. Mema, Y. Himoto, et al. Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clinical Radiology*, 72(1):3 – 10, 2017.

[71] F.J. Brooks. On some misconceptions about tumor heterogeneity quantification. *European Journal of Nuclear Medicine and Molecular Imaging*, 40, 2013.

[72] M.-C. Desseroit, F. Tixier, W.A. Weber, et al. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non–small cell lung cancer tumors: A repeatability analysis in a prospective multicenter cohort. *Journal of Nuclear Medicine*, 58(3):406–411, 2017.

[73] M. Hatt, F. Tixier, C. Cheze Le Rest, et al. Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *European Journal of Nuclear Medicine and Molecular Imaging*, 40(11):1662–1671, 2013.

[74] A. Forgács, H.P. Jonsson, M. Dahlbom, et al. A study on the basic criteria for selecting heterogeneity parameters of F18-FDG PET images. *PLOS ONE*, 11:e0164113, 2016.

[75] F. Tixier, M. Hatt, C. C. Le Rest, et al. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in ^{18}F-FDG PET. *Journal of Nuclear Medicine*, 53(5):693–700, 2012.

[76] F.J. Brooks and P.W. Grigsby. FDG uptake heterogeneity in FIGO IIb cervical carcinoma does not predict pelvic lymph node involvement. *Radiation Oncology (London, England)*, 8:294, 2013.

[77] M. Hatt, D. Visvikis, N.M. Albarghach, et al. Prognostic value of 18F-FDG PET image-based parameters in oesophageal cancer and impact of tumour delineation methodology. *European Journal of Nuclear Medicine and Molecular Imaging*, 38(7):1191–1202, 2011.

[78] F. O'Sullivan, C. Vernon, J. Eary, et al. Incorporation of tumor shape into an assessment of spatial heterogeneity for human sarcomas imaged with FDG-PET. *Biostatistics*, 6(2):293–301, 2005.

[79] F.J. Brooks and P.W. Grigsby. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. *Journal of Nuclear Medicine*, 55, 2013.

[80] F. Tixier, C.C. Le Rest, M. Hatt, et al. Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *Journal of Nuclear Medicine*, 52(3):369–378, 2011.

[81] F.J. Brooks and P.W. Grigsby. Quantification of heterogeneity observed in medical images. *BMC Medical Imaging*, 13(1):7, 2013.

[82] E.C. de Heer, A.H. Brouwers, R. Boellaard, et al. Mapping heterogeneity in glucose uptake in metastatic melanoma using quantitative 18F-FDG PET/CT analysis. *European Journal of Nuclear Medicine and Molecular Imaging Research*, 8(1):101, 2018.

[83] G. Doumou, M. Siddique, C. Tsoumpas, et al. The precision of textural analysis in 18F-FDG-PET scans of oesophageal cancer. *European Radiology*, 25(9):2805–2812, 2015.

[84] F.H.P. van Velden, P. Cheebsumon, M. Yaqub, et al. Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies. *European Journal of Nuclear Medicine and Molecular Imaging*, 38(9):1636–1647, 2011.

[85] F.H.P. van Velden, I. Nissen, F. Jongsma, et al. Test-retest variability of a cumulative SUV-volume histogram method for quantification of FDG uptake heterogeneity. *Journal of Nuclear Medicine*, 53(1):2232, 2012.

[86] G. Thibault, B. Fertil, C. Navarro, et al. Texture indexes and gray level size zone matrix application to cell nuclei classification. *10th International Conference on Pattern Recognition and Information Processing*, 2009.

[87] R.M. Haralick, K.S. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Trans. Systems, Man, and Cybernetics*, 3:610–621, 1973.

[88] M. Amadasun and R. King. Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(5):1264–1274, 1989.

[89] J.J. Foy, K.R. Robinson, H. Li, et al. Variation in algorithm implementation across radiomics software. *Journal of Medical Imaging*, 5(4):1 – 10 – 10, 2018.

[90] I. Fornacon-Wood, H. Mistry, C. J. Ackermann, et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *European Radiology*, 30(11):6241–6250, November 2020.

[91] The Image Biomarker Standardisation Initiative. The image biomarker standardisation initiative - IBSI 0.0.1dev documentation. Available

at https://ibsi.readthedocs.io/en/latest/index.html, accessed 30th June 2022, published 2021.

[92] F. Orlhac, M. Soussan, J. A. Maisonobe, et al. Tumor texture analysis in ^{18}F-FDG PET: Relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. *Journal of Nuclear Medicine*, 55(3):414–422, March 2014.

[93] M. Hatt, M. Majdoub, M. Vallières, et al. ^{18}F-FDG PET uptake characterization through texture analysis: Investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. *Journal of Nuclear Medicine*, 56(1):38–44, January 2015.

[94] M. Carles, T. Fechter, L. Martí-Bonmatí, et al. Experimental phantom evaluation to identify robust positron emission tomography (pet) radiomic features. *European Journal of Nuclear Medicine and Medical Imaging Physics*, 8(1), December 2021.

[95] E. Pfaehler, J. Van Sluis, B. B. J. Merema, et al. Experimental multicenter and multivendor evaluation of the performance of pet radiomic features using 3-dimensionally printed phantom inserts. *Journal of Nuclear Medicine*, 61(3):469–476, March 2020.

[96] S. C. Strother, M. E. Casey, and E. J. Hoffman. Measuring PET scanner sensitivity: relating countrates to image signal-to-noise ratios using noise equivalents counts. *IEEE Transactions on Nuclear Science*, 37(2):783–788, April 1990.

[97] C. C. Watson, M. E. Casey, B. Bendriem, et al. Optimizing injected dose in clinical PET by accurately modeling the counting-rate response functions specific to individual patient scans. *Journal of Nuclear Medicine*, 46:1825–1834, 2005.

[98] M. D. Walker, J. C Matthews, M.-C. Asselin, et al. Patient Specific Noise-Equivalent-Counts from Repeated, Dose Varying ^{15}O-H_2O PET Scans. In *IEEE Nuclear Science Symposium Conference*, 2007.

[99] T. Chang, G. Chang, J. W. Clark, et al. Reliability of predicting image signal-to-noise ratio using noise equivalent count rate in PET imaging. *Medical Physics*, 39(10):5891–5900, October 2012.

[100] Siemens. *Biograph mCT TrueV Datasheet*, published 2018.

[101] National Electrical Manufacturers Association. Performance measurements of positron emission tomographs (PET). *NEMA Standards Publication NU 2-2018*, 2018.

[102] M. E. Daube-Witherspoon, J. S. Karp, M. E. Casey, et al. PET Performance Measurements Using the NEMA NU 2-2001 Standard. *Journal of Nuclear Medicine*, 43:1398–1409, 2002.

[103] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[104] Simplify3D. Materials guide. Available at `https://www.simplify3d.com/support/materials-guide/`, accessed 27th June 2022, published 2019.

[105] R. Gadd, M. Baker, K. S. Nijran, et al. *Measurement Good Practice Guide No. 93: Protocol for Establishing and Maintaining the Calibration of Medical Radionuclide Calibrators and their Quality Control*, May 2006.

[106] Bernard Gibaud. The DICOM standard: A brief overview. In Yves Lemoigne and Alessandra Caner, editors, *Molecular Imaging: Computer Reconstruction and Practice*, pages 229–238, Dordrecht, 2008. Springer Netherlands.

[107] DICOM Library. DICOM Library. Available at `https://www.dicomlibrary.com/dicom/dicom-tags/`, accessed 26th June 2022, published 2022.

[108] S. R. Cherry, T. Jones, J. S. Karp, et al. Total-body PET: Maximizing sensitivity to create new opportunities for clinical research and patient care. *Journal of Nuclear Medicine*, 59(1):3–12, January 2018.

[109] Siemens Medical Solutions Inc. Biograph mCT Flow: PET Technical and Clinical Advances with FlowMotion Technology. Technical white paper, Siemens AG, October 2013.

[110] P. E. Shrout and J. L .Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86:420–428, 1979.

[111] P. E. Kinahan and J. W. Fletcher. Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy. *Seminars in Ultrasound, CT and MRI*, 31(6):496–505, December 2010.

[112] S. Bonnini, L. Corain, M. Marozzi, and L. Salmaso. *Nonparametric hypothesis testing : rank and permutation methods with applications in R*. John Wiley & Sons, Chichester, West Sussex, 2014.

[113] Alan Agresti. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Inc., March 2007.

[114] S. Jan, G. Santin, D. Strul, et al. GATE: a simulation toolkit for PET and SPECT. *Physics in Medicine and Biology*, 49:4543–61, 2004.

[115] R. Brun, F. Rademakers, P. Canal, et al. ROOT - An Object Oriented Data Analysis Framework v6-18-02. Available at `https://doi.org/10.5281/zenodo.3895860`, accessed 27th June 2022, published 2019.

[116] K. Thielemans, C. Tsoumpas, S. Mustafovic, et al. STIR: Software for tomographic image reconstruction release 2. *Physics in Medicine and Biology*, 57(4):867–883, 2012.

[117] T. Merlin, S. Stute, D. Benoit, et al. CASToR: A generic data organization and processing code framework for multi-modal and multi-dimensional tomographic reconstruction. *Physics in Medicine and Biology*, 63, 2018.

[118] R. Twyman, L. Brusaferri, E. Emond, et al. STIR-GATE-Connection. Available at `https://github.com/UCL/STIR-GATE-Connection`, accessed 29th June 2022, published 2021.

[119] B. Sanghera. A Phantom Study in the Pandemic. *SCOPE*, 32(1):28–31, 2022.

[120] S. Sanduleanu, H. C. Woodruff, E. E. C. de Jong, et al. Tracking tumor biology with radiomics: A systematic review utilizing a radiomics quality score. *Radiotherapy and Oncology*, 127(3):349–360, June 2018.

[121] S. F. Barrington, N. G. Mikhaeel, L. Kostakoglu, et al. Role of imaging in the staging and response assessment of lymphoma: Consensus of the international conference on malignant lymphomas imaging working group. *Journal of Clinical Oncology*, 32(27):3048–3058, September 2014.

[122] E. Lopci and M. Meignan. Deauville score: the phoenix rising from ashes. *European Journal of Nuclear Medicine and Molecular Imaging*, 46(5):1043–1045, May 2019.

[123] L. Lu, W. Lv, J. Jiang, et al. Robustness of Radiomic Features in [11C]Choline and [18F]FDG PET/CT Imaging of Nasopharyngeal Carcinoma: Impact of Segmentation and Discretization. *Molecular Imaging and Biology*, 18(6):935–945, December 2016.

[124] V. Liberini, B. De Santi, O. Rampado, et al. Impact of segmentation and discretization on radiomic features in 68ga-dota-toc PET/ct images of neuroendocrine tumor. *European Journal of Nuclear Medicine and Medical Imaging Physics*, 8(1), December 2021.

[125] P. Whybra, C. Parkinson, K. Foley, et al. Assessing radiomic feature robustness to interpolation in ^{18}F-FDG PET imaging. *Scientific Reports*, 9(1), December 2019.

[126] A. J. Reader, G. Corda, A. Mehranian, et al. Deep learning for PET image reconstruction. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(1):1–25, August 2020.

[127] S. R. Cherry, R. D. Badawi, J. S. Karp, et al. Total-body imaging: Transforming the role of positron emission tomography. *Science Translational Medicine*, 9(381), March 2017.

[128] B. A. Spencer, E. Berg, J. P. Schmall, et al. Performance Evaluation of the uEXPLORER Total-Body PET/CT Scanner Based on NEMA NU 2-2018 with Additional Tests to Characterize PET Scanners with a Long Axial Field of View. *Journal of Nuclear Medicine*, 62(6):861–870, June 2021.

[129] G. A. Prenosil, H. Sari, M. Fürstner, et al. Performance characteristics of the Biograph Vision Quadra PET/CT system with a long axial field of view using the NEMA NU 2-2018 standard. *Journal of Nuclear Medicine*, 63(3):476–484, 2022.

[130] S. Vandenberghe, P. Moskal, and J. S. Karp. State of the art in total body PET. *European Journal of Nuclear Medicine and Medical Imaging Physics*, 7(1), December 2020.

[131] L. Nardo, J. P. Schmall, T. J. Werner, et al. Potential roles of total-body PET/computed tomography in pediatric imaging. *PET Clinics*, 15(3):271–279, July 2020.

[132] M. Stockhoff, M. Decuyper, R. Van Holen, and S. Vandenberghe. High-resolution monolithic LYSO detector with 6-layer depth-ofinteraction for clinical PET. *Physics in Medicine and Biology*, 66(15), August 2021.

[133] E. Lamprou, A. J. Gonzalez, F. Sanchez, and J. M. Benlloch. Exploring tof capabilities of PET detector blocks based on large monolithic crystals and analog SiPMs. *Physica Medica*, 70:10–18, February 2020.

[134] USA Subtle Medical, Inc. SubtlePET™: Improved quality and efficiency on your existing machines. Available at `https://subtlemedical.com/subtlepet/`, accessed 22nd June 2022, published 2018.

[135] K. Weyts, C. Lasnon, R. Ciappuccini, et al. Artificial intelligence-based PET denoising could allow a two-fold reduction in [18F]FDG PET acquisition time in digital PET/CT. *European Journal of Nuclear Medicine and Molecular Imaging*, (5), 2022.

[136] J. Adler and O. Öktem. Learned primal-dual reconstruction. *IEEE Transactions on Medical Imaging*, 37(6):1322–1332, June 2018.

[137] J. Hsieh, E. Liu, B. Nett, et al. A new era of image reconstruction: Truefidelity ™, technical white paper on deep learning image reconstruction. Available at https://www.gehealthcare.co.uk/-/media/files/truefidelity/truefidelity-white-paper-jb68676xx-doc2287426.pdf?rev=-1, accessed 27th June 2022, published 2019.

[138] A. J. Reader and G. Schramm. Artificial intelligence for PET image reconstruction. *Journal of Nuclear Medicine*, 62(10):1330–1333, October 2021.

[139] P. Wu, A. Sisniega, A. Uneri, et al. Using Uncertainty in Deep Learning Reconstruction for Cone-Beam CT of the Brain. In *Proceedings of the 16th Virtual International Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine*, pages 366–371, October 2021.

Appendix I

Experimental Discretisation Method

The discretisation method chosen for use in the experiments detailed in this thesis was FBN:64. The evidence considered for this selection is detailed further in Chapter 2. To test this in practice, a series of discretisation protocols were applied to the cylinder dataset as listed in Table I.1.

Bin Widths	Bin Numbers
FBS:0.005	FBN:8
FBS:0.01	FBN:16
FBS:0.05	FBN:24
FBS:0.1	FBN:32
FBS:0.2	FBN:40
FBS:0.25	FBN:48
FBS:0.5	FBN:56
FBS:1.0	FBN:64
FBS:2.0	FBN:128
	FBN:256
	FBN:512

Table I.1: A table containing the discretisation protocols used on the cylinder image dataset.

The purpose of this analysis was to demonstrate that there would be little to no informational advantage gained by selecting another discretisation protocol for analysing the phantom data. Figure I.1 demonstrates this for FBN discretisations. The figure shows an example image feature, the GLCM IMC2, for the cylinder images collected as detailed in Chapters 3 and 5 subject to the FBN discretisation protocols. It was observed that for all features, the functional form of feature

values was consistent regardless of the number of bins used, and despite any expected offset. This lead to the conclusion that any perceived correlation of feature values to NECR could be obtained regardless of the number of bins used, and that the number of bins used should be chosen in order to be consistent with the field and previous work.

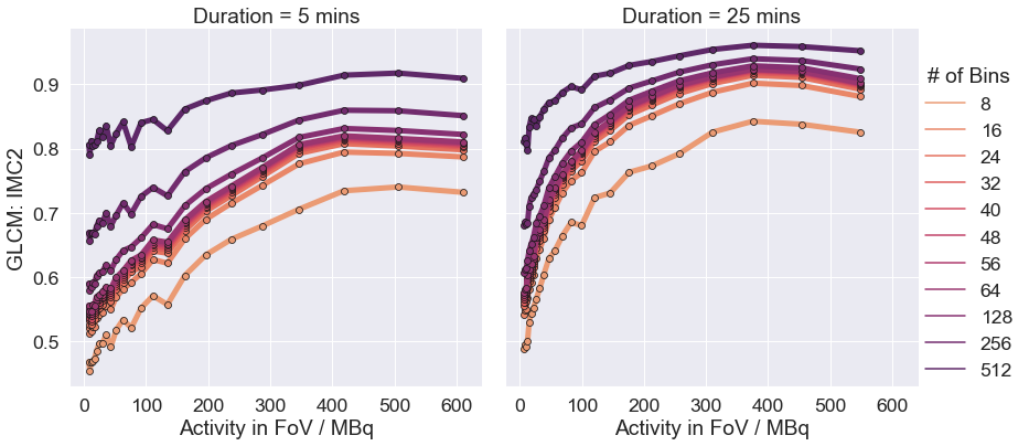


Figure I.1: The Informational Measure of Correlation 2 (IMC2) from GLCM for the cylinder datasets, comparing all FBN discretisation protocols.

While this study was not considered completely appropriate for FBS discretisation, a series of different bin widths were examined for completeness. The effect of changing FBS discretisation used on the same image set can be seen in Figure I.2, which observes the same feature seen in Figure I.1.

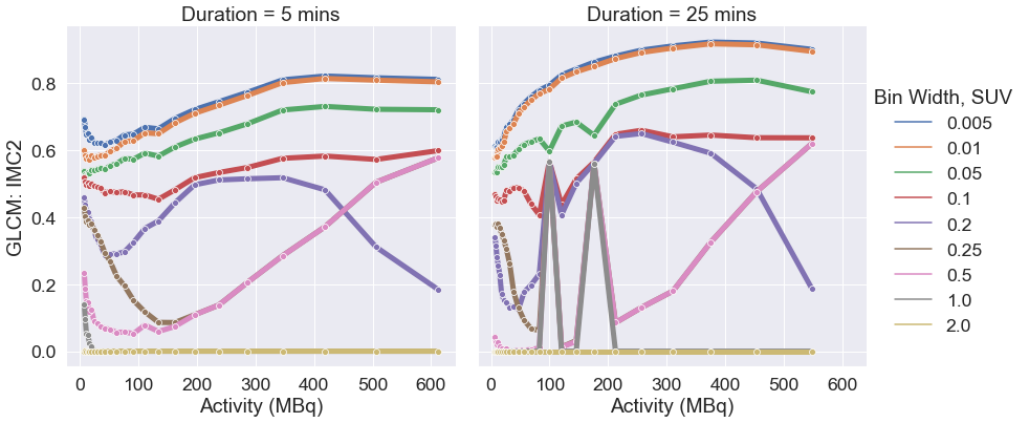


Figure I.2: The IMC2 from GLCM for the cylinder datasets, comparing all FBS discretisation protocols.

The width of the bins must be compared to the range of values in the image across the set. The voxel value ranges are shown in Figure I.3.

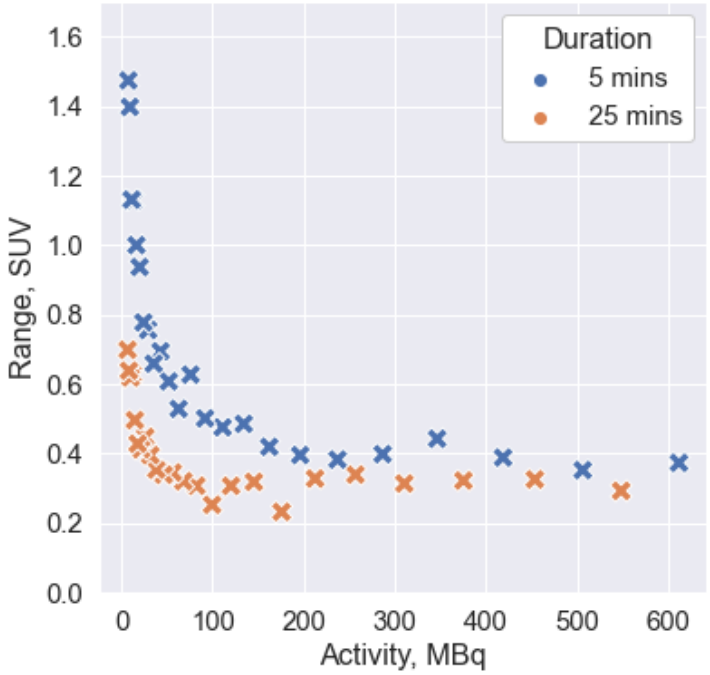


Figure I.3: The range of values, in SUV, for all images in the cylinder dataset.

The IBSI recommend a fixed bin width that guarantees between 30 and 130 bins to allow for reproducibility of image feature values [80]. From the range of values seen in Figure I.3, this determines that a value of the order of 0.01 SUV would enable this. Figure I.2 shows that there is likely to be no informational advantage to bins smaller than that. There is an adverse effect on the functional form of the feature across the dataset by increasing the bin width; the feature response appears unpredictable with increasing bin size. This can be attributed to loss of information by discretising into large bins. Should FBS discretisation be used for future work, a full evaluation of bin widths around the magnitude of 0.01 SUV should be done in order to determine the most appropriate value. It was decided that, for this work, the context did not require finding this value, and the experiment was to continue with FBN:64.